Alan Turing at 100: Artificial Intelligence and an Epistemic Stance

George F. Luger (luger@cs.unm.edu) 505-277-3204
Professor of Computer Science, Psychology, and Linguistics
University of New Mexico
Albuquerque NM USA 87131

Abstract

It has been 100 years since the birth of Alan Turing and more than sixty years since he published in *Mind* his seminal paper, *Computing Machinery and Intelligence*. In this paper Turing asked a number of questions, including whether computers could ever be said to have the power of "thinking". Turing also set up a number of criteria – including his *imitation game* – under which a human could "judge" whether software on a computer could be said to be "intelligent". Turing's paper, as well as his important mathematical and computational insights of the 1930s and 1940s led to his popular acclaim as the "Farther of Artificial Intelligence". In the sixty years since his paper was published, no computational system has fully satisfied Turing's challenge.

In this paper we focus on a different question, ignored in, but inspired by, Turing's paper: What is the nature of intelligence and how might it actually be implemented on a computational device? After sixty years of both directly and indirectly addressing this question, the Artificial Intelligence community has produced few answers. Nonetheless, the AI Community has constructed a large number of important artifacts, as well as several philosophical stances able to shed light on the nature of intelligence. This paper addresses the issue of how formal representational schemes illuminate the nature of intelligence, and further, how an intelligent agent can understand the nature of its own intelligence; this is often called the problem of *epistemological access*. The result of this access is an *epistemic stance*.

I propose to consider the question, "Can computers think?"…

Alan Turing, Computing Machinery and Intelligence, *Mind*, 1950.

Theories are like nets: he who casts, captures…

Wittgenstein, *Philosophical Investigations*, 1953

Keywords: Computational Intelligence, Epistemjc Stance, Stochastic Methods

1. Introduction: The Imitation Game

Turing proposed to answer his "Can computers think" question by introducing a gedanken experiment called the *imitation game*. In the imitation game a human, the "interrogator", asks questions of two different entities, one a human and the other a computer. The interrogator is isolated from the two respondents so that he/she does not know whether the human or computer is answering. Turing, in the language of the 1940s, comments that "the ideal arrangement is to have a teleprinter communicating between the two rooms", ensuring the anonymity of the responses. The task of the interrogator is to determine whether he/she is communicating with the computer or the human at any time during the question answering session. If the interrogator is unable to determine who is responding to questions, Turing contends that the computer has to be seen as "thinking", or, more directly, to possess intelligence.

The historical timing of Turing's paper is very instructive. It appeared before computers were challenged to understand natural human languages, play chess, recognize visual scenes, or control robots in deep space. Turing and others (Turing 1936, Church 1941, Post 1943) had already formally specified what it *meant to compute*, and had by that time hypothesized limits on *what was computable*. This *sufficient* model for any computation is often called the Church/Turing hypothesis (Turing 1936). However, the radio-tube-based-behemoths of Turing's time were used mainly to work out the trajectories of ordinance and to break complex ciphers. It is important to realize then – given the very limited nature of actual tasks addressed at that time by computers - that the most important result of Turing's imitation game was to challenge humans to consider whether or not thinking and intelligence are uniquely human skills. The task of Turing's imitation game was an important attempt to separate the attributed skills of "thinking" and "intelligence" from their human embodiment.

Of course, no informed critic would contend that electronic computers, at least as presently configured, are *universally* intelligent – they simply do a large number of specific but complex tasks - delivering medical recommendations, guiding surgeries, playing chess or backgammon, learning relationships in large quantities of data, and so on - as well as, and often much better than, their human counterparts performing these same tasks. In these limited situations, computers have passed Turing's test.

It is interesting to note, also, that many in the research community are still trying to play/win this challenge of building a general purpose intelligence that can pass the Turing test in any area that a human might challenge it. This can be seen as useful, of course, for it requires the computer and program designer to address the more complete and complex notion of building a general-purpose intelligence. Perhaps the program closest to achieving this goal is IBM's Watson, the winner of the Jeopardy television challenge of February 2011 http://en.wikipedia.org/wiki/Watson_(computer), www 9/9/12. Commercially available programs addressing the quest for general intelligence include web bots, such as Apple's Siri. The Turing test challenge remains an annual event, and the interested reader may visit the url http://en.wikipedia.org/wiki/Turing_test#Loebner_Prize, www 9/9/12, for details.

In fact, the AI community still uses forms of the imitation game to test whether their programs are ready for actual use. When the computer scientists and medical faculty at Stanford were ready to deploy their MYCIN program they tested it against a set of outside medical experts skilled in the diagnosis of meningitis infections (Buchanan and Shortliffe 1984). The results of this analysis were very interesting, not just because, in the double-blind evaluation, the MYCIN program out performed the human experts, but also because of the lack of a general consensus – only about 70% agreement - on how the human experts themselves would treat these patients! Besides evaluating many deployed expert systems, a form of Turing's test is often used for testing AI-based video games, chess and backgammon programs, computers that understand human languages, and various forms of web agents.

The failure, however, of computers to succeed at the task of creating a general-purpose thinking machine begins to shed some understanding on the "failures" of the imitation game itself. Specifically, the imitation game offers no hint of a definition of intelligent activity nor does it offer specifications for building intelligent artifacts. Deeper issues remain that Turing did not address. What IS intelligence? What IS grounding (how may a human's or computer's statements be said to have "meaning")? Finally, can humans understand their own intelligence in a manner sufficient to formalize or replicate aspects of it?

This paper considers these issues, especially the responses to the challenge of building intelligent artifacts that the artificial intelligence community has taken since Turing. In the next section we introduce deductive, inductive, and abductive reasoning (Peirce 1958) as general methodological constructs for describing intelligent activity. We also discuss epistemology, the study of our understanding of intelligence itself, and finally, the question

of epistemological access – whether, in fact, an agent can understand its own interactions with both the self and non-self worlds.

In the third section we give a brief overview of several AI programs built over the past sixty years as intelligent problem solvers. We see, often apart from the practical stance of the program's original designers, many of the earliest approaches to AI as being components of an empiricist or rationalist approach to understanding an external world. In the fourth section we present a constructivist rapprochement that addresses many of the dualist assumptions of early AI work. We also present and justify a Bayesian approach to the problem of abduction. Finally, we offer some preliminary conjectures about how a Bayesian model might be epistemologically plausible.

2. Human Reasoning and Epistemological Access

Many philosophers and mathematicians break down descriptions of human reasoning into three modes: the deductive, the inductive, and the abductive. Aristotle first described forms for *deductive* reasoning, when he pointed out in his Rhetoric that many statements must be seen as true simply because of the *form* in which they are presented. Aristotle proposed these forms as a methodology for his reader to develop convincing arguments. It required almost two more millennia, however, before mathematicians, including Boole, Frege, and Tarski, formalized a mathematical system for problem representation and the use of specific algorithms for deductive reasoning: the propositional and predicate calculi. We present one form for deductive reasoning, *modus ponens*, in Figure 1. Deductive inference plays a large role in rationalist approaches to AI, as we see in the following section.

a.		b.	
	p → q		P(X) → Q(X)
and	p	and	P(george)
implies	q	implies	Q(george)

Figure 1a. The deductive form for *modus ponens* using the propositional calculus. **Figure 1b** is modus ponens with the predicate calculus and instantiation, {george/X).

Inductive reasoning is also a formal method for building arguments about properties of specific systems. Induction is applied across potentially infinite structures whose elements can be indexed by the counting numbers, that is, by systems reflecting Peano's axioms. For an extended discussion of these axioms see http://en.wikipedia.org/wiki/Peano_axioms 9/9/12). Informally, given a base case and for each case a well formed procedure for

creating a next case, properties can be asserted for all possible countably infinite cases. Figure 2 offers an example of induction, inferring a general formula for calculating a sequence of sums. Inductive proof procedures have proven valuable for mathematicians and computer scientists, especially for proof of properties related to recursive procedures.

The most general, and seemingly open-ended practice of problem solving is referred to as *abductive*, or sometimes *empirical induction*. One way of understanding this is that in limited micro-areas of reasoning the deductive and inductive forms of inference may be appropriate but in most all aspects of human interaction these forms are not sufficient to capture what is going on. When agents interact with the stimuli afforded by the surrounding world their task is to determine *the best explanation* for the data they are experiencing. In fact, Peirce (1958), the twentieth century philosopher who has made an extensive analysis of abductive reasoning, describes it this way: Abduction is *reasoning to the best explanation, given a set of perceptual cues*.

Term number: 1 2 3 ... n

For all integers n>0: $1^2 + 2^2 + 3^2 + \ldots + n^2 = (n(n + 1)(2n + 1)) / 6$

Figure 2. An inductive argument: for P(1): $1 = (1(1 + 1)(2 + 1))/6$, so the base case is true. P(n + 1): $(1^2 + 2^2 + 3^2 + \ldots + n^2) + (n + 1)^2 = ((n + 1)(n + 1 + 1)(2(n + 1) + 1) / 6$ $= (n(n + 1)(2n + 1))/6 + (n + 1)^2 = ((n + 1)(n + 2)(2n + 3))/6$. Since the base case P(1) is true and the n + 1 case, P(n+ 1) can be built from the P(n) case, this relationship of sums is true for all n > 0.

So what might abduction include? Certainly it includes speech communication between agents. When, for example, an utterance (sound) is made by an agent in some context, the perceiving agent does not have access to the content of the speaker's head. Nor does he/she have perfect access to how the speaker is interpreting his/her environment. As a listener we are constantly working, often with comments or questions, towards the best explanation of what meaning might be afforded, given both the speaker's sounds and the context of interpretation. There are myriad similar examples of abductive inference available, for instance in medical diagnosis, where the doctor does not have perfect information of the patient, but actively struggles, often with measures and testing of the patient's symptoms, to come to the best explanation, usually by the attribution of some disease state. And in fact, even the patient's symptoms are not directly "perceived", but seen as a function of, or best explanation for, the readings from some instrument, a temperature or blood pressure measure, for example, or the "results" of some blood test.

Aristotle, in his Rhetoric, first described the abductive phenomena, commenting how the observation of a woman nursing is an indicator of a probable earlier pregnancy. Peirce introduced the topic more formally (1958), and Eco (1976) and Sebeok (1985) describe abduction, often in a literary context, as a form of semiotic analysis. Discussion of abductive phenomena by modern psychologists includes that of Schvaneveldt et al. (2010).

We present in Figure 3 a description of the abductive argument form, noting how it relates to the modus ponens model of deductive inference shown in Figure 1. Further, Figure 3 demonstrates how multiple factors, the **p, r, s**, could *each* explain the perceived set of evidence, **q**. Note that abduction is an *unsound* reasoning rule in that the argument is not guaranteed (mathematically) to be correct – it is just proposed as a possible explanation of the data. Various devices, including probability measures and certainty factors (see, for example, Section 4) may be employed in computational models to prioritize best explanations, given sets of possible explanations and supporting data. We present computational models for several examples of abductive inference in Section 6.

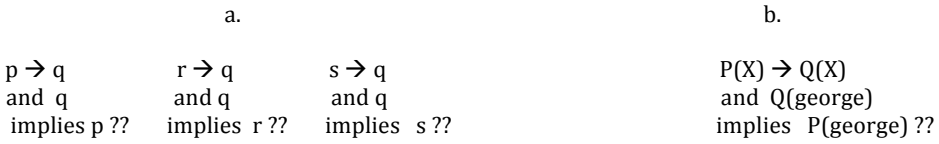


Figure 3a. The abductive form of reasoning with the propositional calculus where p, r, and s all imply q. We wish to determine which of p or s or r offers the best explanation for the fact of q. Similarly, **Figure 1b** is abduction with the predicate calculus and instantiation, {george/X).

Any characterization of the abductive processes of an active agent's problem solving – perceptions triggering possible explanations - immediately requires us to consider deeper issues of human information acquisition and insight. To ask how an agent perceives, interprets, and explains impinging patterns of "stuff" (experienced as the philosopher's *qualia*?) is fundamental to understanding how that agent copes with its world. Furthermore, the problem of whether or not the agent can come to understand and characterize its own interactions with this "stuff" introduces the issue of *epistemological access*.

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 a topic for scientific study. The empiricist and rationalist traditions have offered their specific answers to this question and artificial intelligence researchers have have made these approaches concrete with their programs, as we see in the following section. We then propose, in

Section 4, a constructivist, model-refinement approach to epistemological issues and propose a Bayesian characterization of this agent/world interaction. We present in Section 5 several Bayesian-based models of abductive reasoning and point out epistemological aspects of this approach. We conclude with some discussion of possible cognitive correlates of this class of computational model.

3. Artificial Intelligence as Adventures in Rationalism and Empiricism

Over the past sixty years, most of the research efforts in artificial intelligence can be characterized as an ongoing dialectic between the empiricist and rationalist traditions in philosophy, psychology, 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 would intersect with philosophy and psychology, and in particular, with epistemology. We describe this intersection of disciplines in due course, but first we look at these philosophical traditions themselves.

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: the reality of his own physical existence could be reestablished 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 its comprehension was enabled through a 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 Kepler's laws of planetary motion 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 our own modern life, literature, and religion (e.g., *the spirit is willing but the flesh is weak*).

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 come 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 forms of the disembodied life: learning is remembering. In his cave experience, in book seven 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. Rationalism's embodiment entering the AI-age can be found in the early twentieth century analytic philosophers, the symbol-based AI practitioner Herb Simon (1981), and especially in the works of the linguist Noam Chomsky (1957). It provides a natural starting point for work in AI as we see subsequently.

Aristotle was one of the first proponents of the empiricist tradition, although his philosophy also contained the ideas of "form" and the ability to "abstract" from a purely material existence. However, Aristotle rejected Plato's doctrine of transcendent forms, noting that the act of abstraction did not entail an independent existence. For Aristotle the most important aspect of nature is 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 relations and algorithms describe how these entities can share common characteristics, as well as be systematically altered. 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.

In the world of the enlightenment, the empiricist tradition, espoused by Locke, the early 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 view the rationalist's perfect sphere, or absolute truth, simply do not exist. Locke suggests that the human at birth is *tabula rasa*, a *blank slate*, where all language and human "meaning" is captured as conditioning across time and experience. What the human agent "internalizes" are the human-perceptible aspects 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/skeptic position: it is not that a God doesn't/can't exist, it is rather that the human agent can't know or prove that he/she *does* exist.

The empiricist tradition was especially strong in the first half of the twentieth century leading into the AI movement, where its supporters included A. J. Ayer and Rudolph Carnap, proponents of *logical empiricism*, who tried to fuse empiricism with a logic-based rationalism, as well as the behaviorist psychologist B. F. Skinner.

Many modern artificial intelligence practitioners have implicitly adopted either empiricist or rationalist views of the world. To offer several examples: From the rationalist perspective came the expert system technology where knowledge was seen as a set of clear and distinct relationships (expressed in *if/then* or *condition/action* rules) encoded within a production system architecture that could then be used to compute decisions in particular situations. Figure 4 offers a simplified example of this approach, where a rule set – the

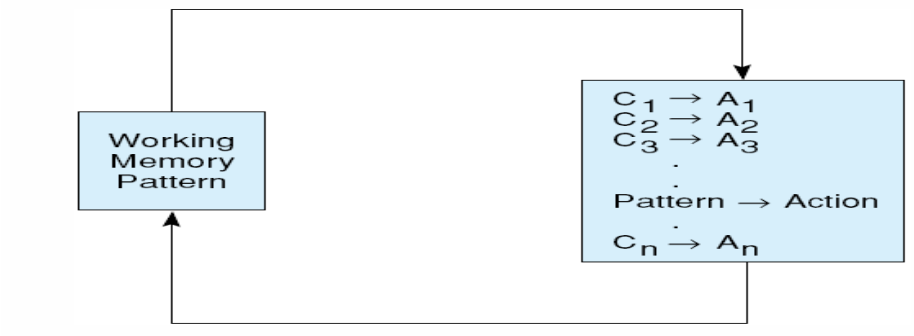


Figure 4. A production system; in the traditional data-driven mode, a pattern in the Working Memory matches the condition of a rule in the Production Memory. When this happens the conclusion of that rule is asserted (*modus ponens*) as a new pattern in Working Memory and the system continues to iterate towards a solution.

content of the production memory – is interpreted by the production system. When the *if* component of the rule is matched by the data in the working memory, the rule is said to "fire" and its conclusion then changes the content of the working memory preparing it for the next iteration of the system. The reader can observe that when the system is run in this "data-driven" mode it is equivalent to a modus ponens interpreter, as seen in Figure 1.

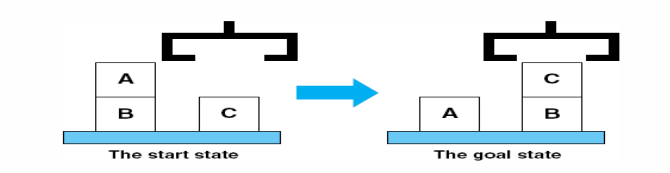
Interestingly enough, when the same production system is run in goal-driven mode it can be seen as an abductive interpreter. In this situation the goals we wish to solve – the explanations we want to prove "best" – are contained in the working memory and the production system takes these goals and matches them to the conclusions, the *action* or *then* components of the rules. When a conclusion is matched the rule again "fires" and the system puts the *if* pattern of the rule into the working memory to serve as a subgoal for the next iteration of the system, matching the conclusions of new rules. In this abductive mode (Figure 3) the system searches back through a sequence of subgoals to see if it can make the case for the original goal/explanation to be true. As noted earlier, abduction is an unsound form of reasoning, so the abductive interpreter can be seen as generating possible explanations for the data. In many cases, some probabilistic or certainty factor measure is included with each rule supporting the interpreter's likelihood of producing the "best" explanation (Luger 2009, Chapter 9).

In the work of Newell and Simon (1972) and Simon (1981) this production system interpreter was taken a further step towards cognitive plausibility. On the Newell and Simon view, the production memory of the production system was a characterization of the human long-term memory and the if/then rules were seen to encode specific components of human knowledge. On this approach, human expertise for the practicing physician or the master chess player, for example, was acknowledged to be about 50,000 such rules (Newell and Simon 1976). The working memory of the production system was seen as the human's short-term memory, or as describing a "focus of attention" for what the human agent was considering at any specific time (we now refer to this as Broadmann's areas of pre-frontal cortex). Thus the production system was proposed as a *cognitive architecture* that took the current focus of the agent and used that to "fire" specific components of knowledge (rules) residing in long-term memory, which, in turn, changed the agent's focus of attention. Furthermore, production system learning (SOAR, Rosenbloom et al. 1993) was seen as a set

of procedures that could encode an agent's repeated experiences in a domain into new if/then rules in long-term (production) memory.

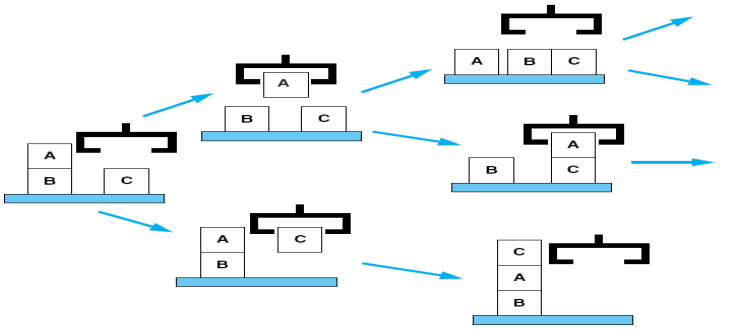
Early design of robot systems (Fikes and Nilsson 1971) can also be seen as a rationalist exercise where the world is described as a set of explicit constraints that are to be organized as "states" and searched to accomplish a particular task. "States" of the world are represented as a set of predicate calculus descriptions and then these are checked by a set of "move" rules that are used to generate new states of the world, much as the production system did in the previous example. Figure 5a presents a start state and a goal state for a configuration of blocks and a robot arm. These states are then changed by applying "move" predicates, as can be seen in the state space of Figure 5b. Problems can happen, of course, when the actual world situation is not represented precisely as the logic specifications would suggest, e.g., when one block or the robot arm accidentally moves another.

a.




```
start = [handempty, ontable(b), ontable(c), on(a,b),
          clear(c), clear(a)]
goal = [handempty, ontable(a), ontable(b), on(c,b),
          clear(a), clear(c)]
```

b.



move(pickup(X), [handempty, clear(X), on(X,Y)], [del(handempty), del(clear(X)), del(on(X,Y)), add(clear(Y)), add(holding(X))]).

Figure 5a. The start and goal states of a blocks world problem, and the set of predicate descriptions for each state. Figure 5b presents part of the state space search representing the movement of the blocks to attain a goal state. The move procedure (stated as preconditions, add, and delete constraints on predicates) is one of many possible predicates for changing the current state of the world.

Several later approaches to the design of control systems take a similar approach. When NASA wanted to design a planning system for controlling the combustion systems for deep space vehicles, it expressed the constraints of the propulsion system as sets of propositional calculus formulae. When the control system for the space vehicle detected any anomaly it searched these constraints to determine what to do next. This system, NASA's Livingstone, proved very successful for guiding the space flight in deep-space situations (Williams and Nayak 1996, 1997; Luger 2009, Sections 8.3 – 8.4).

There are many other examples of this rationalist bias in AI problem solvers. For example, case-based reasoning uses a data base of collected and clearly specified problem solving situations, much as a lawyer might look for earlier legal precedents, cases that can be modified and reused to address new and related problems (Kolodner 1993).

A final example of the rationalist perspective, is the suggestion that various forms of logic, both in representation and inference, can be sufficient for capturing intelligent behavior has long been the position of a small group of researchers in the AI community (McCarthy 1968, McCarthy and Hayes 1969). Many interesting and powerful representations have come from this work including non-monotonic logics, truth-maintenance systems, and assumptions of minimal models or circumscription (McCarthy 1980, 1986; Luger 2009, Chapter 9.1).

From the empiricist view of AI there is the creation of semantic networks, conceptual dependencies, and related association-based representations. These structures, deliberately formed to capture the concept and knowledge associations of the human agent, were then applied to the tasks of understanding human language and interpreting meaning in specific contexts. The original semantic networks were, in fact, taken from the results of psychologists' (Collins and Quillian 1969) reaction-time experiments. The goal was to design associative networks for computer-based problem solvers that captured the associative components of actual human memory. In the reaction time experiments of Figure 6, the longer the human subject took to respond to a query, for example, Does a bird have skin?, the further "apart" these concepts were assumed to be in the human memory system. Closely associated concepts would support more immediate responses.

A number of early AI programs sought to capture this associative representation, first Quillian (1967) himself, with the creation and use of semantic networks. Wilks, (1972) basing his research on earlier work by Masterman (1961) who defined around 100

primitive concept types, also creates semantic representations for the computer based understanding of human language. Schank (Schank and Colby 1975) with their *conceptual dependency* representation, created a set of primitive association-based primitives intended to support language-based human meaning as it might be used for computer understanding on translation.

From the empiricist perspective, neural networks were also designed to capture associations in collected sets of data and then, once trained, to interpret new related patterns in the world. For example, the back-propagation algorithm in training phase takes a number of related situations, perhaps surface patterns for an automated welder or phone patterns of human speech, and conditions a network until it achieves a high percentage of successful pattern recognition. Figure 7a presents a typical neuron from a back-propagation system. The input values for each neuron are multiplied by the (conditioned) weights for that value and then summed to determine whether the threshold for that neuron is reached. If the threshold is reached the neuron fires, usually generating an input signal for other neurons. The back-propagation algorithm, Figure 7b, differentially "punishes" those weights responsible for incorrect decisions. Over the time of training the appropriately configured and conditioned network comes to "learn" the perceptual cues that solve a task. And then the trained network can be used to solve new related tasks.

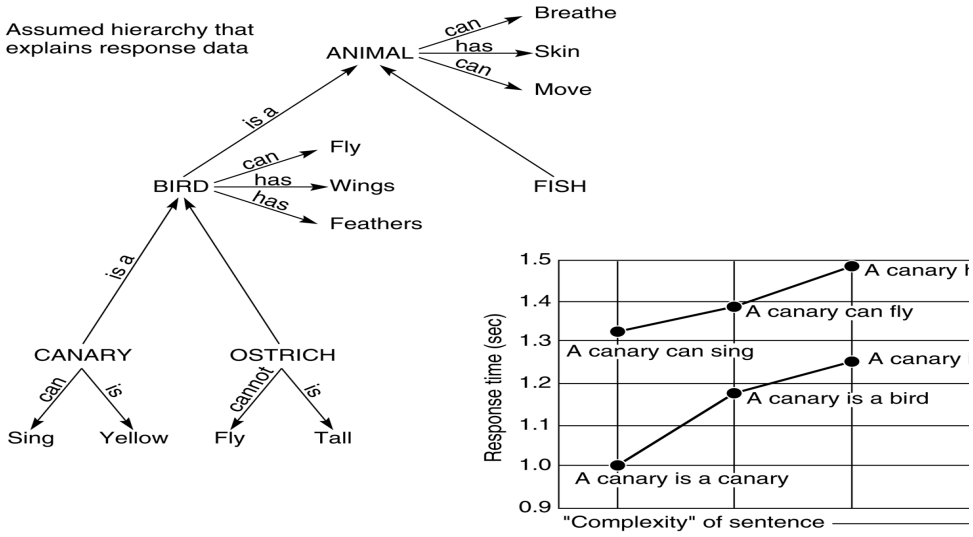


Figure 6. A semantic net "bird" hierarchy (left) that is created from the reaction time data (right) of human subjects (Collins and Quillian 1969). This figure is adapted from Harmon and King (1985).

Back-propagation networks are an example of *supervised* learning, where appropriate rewards and/or punishments are used in the process of training a network. Other network

learning can be *unsupervised* where algorithms classify input data into "clusters" either represented by a prototype pattern or by some "closeness" measure. New input patterns then enter into the basins of attraction offered by the currently clustered patterns. In fact, many successful families of networks have been created over the years. (See Luger 2009, Chapter 12 for an overview of several of these.) There have also been many obvious – and scientifically useless - claims that neural connectivity (networks) ARE the way humans performed these tasks, *and therefore* appropriate representations for use in computer-based pattern recognition.

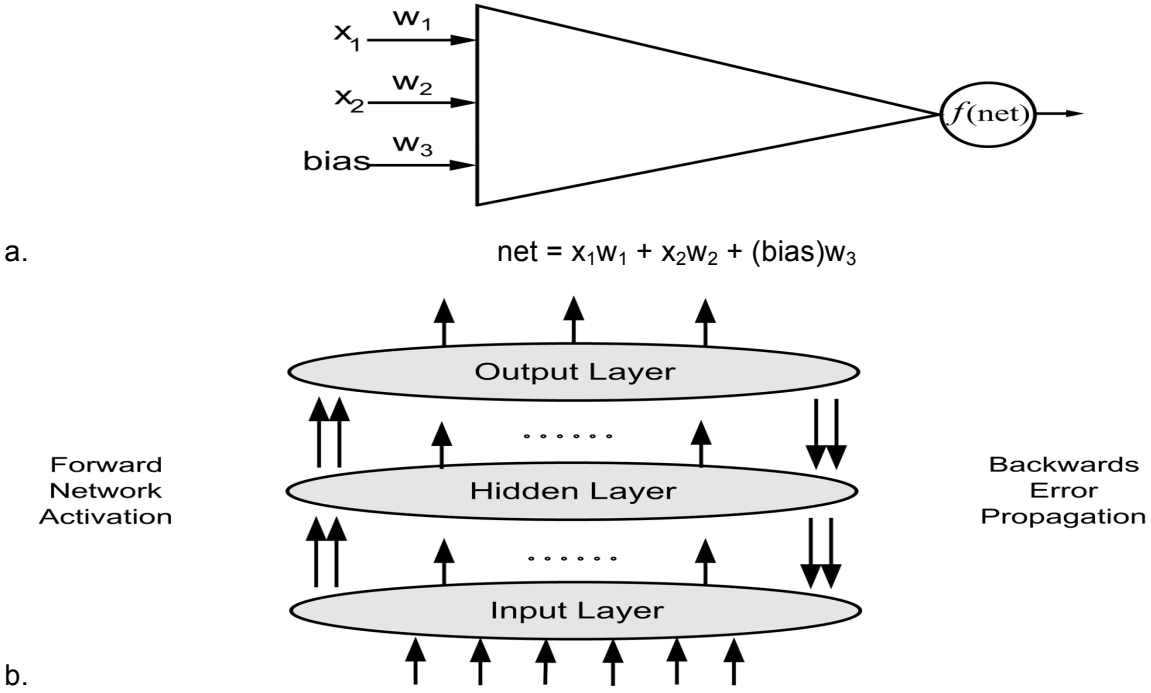


Figure 7a presents a single artificial neuron whose input values, multiplied by (trained) weights, produce a value, net. Usually using some sigmoid function, f(net), produces an output value that may, in turn, be an input for other neurons. Figure 7b is a simple backpropagation network where input values move forward through the nodes of the network. During training the networks weights are differentially "punished" for incorrect responses to the input.

In an interesting response to the earlier rationalist planners for robotics described above, Brooks at the MIT AI laboratory created what he called the "subsumption" architecture (Brooks 1989, 1991). The subsumption architecture was a layered collection of finite state machines where each level of the solver was constrained by layers below it. For example, a "wander" directive at one level of the robot's controller would be constrained by a lower level that prevented the agent from "running into" other objects during wandering.

The subsumption architecture is the ultimate knowledge-free system in that there are no memory traces ever created that could reflect situations that the robot had already learned through pattern association. Obviously such a system would not be able to find its way around a complex environment, for example, the roadways and alleys of a large city. Brooks even seemed to acknowledge this fact of a memory free solver, in entitling his 1991 paper "Intelligence without Representation" (Brooks 1991).

As final examples of representations with an empiricist bias, artificial life and genetic algorithms have been proposed by various groups within the AI community as examples of evolution-based problem-solvers. These architectures may be seen as knowledge-free association-and-reward based solvers that are intended to capture survival of the fittest. Their advocates often saw these approaches as plausible models incorporating evolutionary pressures to produce emergent phenomena, including the intelligent behavior of an agent.

It is not surprising that many of the products of the empiricist and/or rationalist approaches to problem solving have met with only 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, even to generalize and interpret their various results, they were not successful, or in the context of this paper, could not pass Turing's test. The success of the AI practitioner as the designer and builder of new and important software languages and artifacts is beyond question; the notion that this effort builds out the full set of cognitive skills of the human agent is simply naive.

The problem is both epistemological and practical. How does the human 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 overarching epistemological integration of the agent and its ever-changing environment? And how does (or even *can*) the human agent understand this integration?

This paper 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 and model refinement a plausible integration of empiricist and rationalist views. Section 5, using the insights of Bayes theorem, offers a computational model of abductive reasoning able to integrate prior expectations of a situation with

posterior perceptions presented by the phenomenal world. In Section 6, going back to the artificial intelligence tradition, we demonstrate this integration in several examples of abductive (diagnostic) reasoning.

4. A Constructivist Rapprochement

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 *schema* to describe the *a priori* structure used to mediate the experience of the external world. The term schema 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, rarely fits precisely into our preconceived and *a priori* schemata. From this tension, the schema-based biases a subject uses to organize experience are either strengthened, 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 model 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 active agent, of a contingent being in space and time.

There is a blending in constructivism 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 we 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 or transparent 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? To what extent can an agent within an environment understand its own understanding of that situation? We 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 model refinement (Luger 2011).

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 refine this approach to a form of *naïve Bayes*, which we call, as it is computed across time, the *greatest likelihood* measure. This uses continuous data acquisition to do real-time diagnosis through model refinement. Section 6 presents several examples of Bayesian-based AI research supporting a model refinement approach.

5. 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 an earlier 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 (published posthumously, in 1763, in the *Transactions of the Royal Society*) with a mathematical demonstration of how an agent's prior expectations can 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 this section we develop Bayes' insight and in the final section of this paper we conjecture how Bayes' approach might support a computational model and epistemological access.

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 visualize this situation with Figure 8, 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 that has a bad headache also has meningitis.

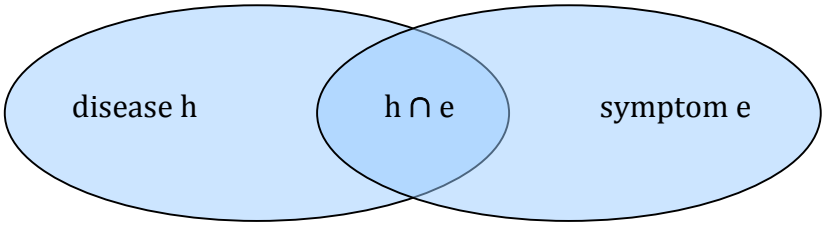


Figure 8. A representation of the numbers of people having a symptom, **e**, and a disease, **h**. Note that what we want to measure is the probability of a person having the disease, given that they suffer the symptom, denoted: **p(h|e)**.

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 considered, we then represent each number as a probability. We represent the probability of the symptom **e** given the disease **h** as **p(e|h)**:

p(e|h) = | e ∩ h | / | h | = p(e ∩ h) / p(h),

where "|" surrounding symbols, e.g., **| e ∩ h |,** indicates "the number of people in that set". The value of **p(e ∩ h)** can now be determined by multiplying by **p(h)**:

p(e ∩ h) = p(e|h) p(h)

We wish to determine the **p(e ∩ h)** value and to do so we have other information from Figure 8, including the number of people (again) that have both the symptom and the disease, **e ∩ h,** as well as the total number of people that have the symptom, **e**. So we determine the value for **p(e ∩ h)** with this information: The probability of the disease **h**, given the evidence **e**, **p(h|e)**:

p(h|e) = p(e ∩ 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:

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, $\mathbf{h_i}$, 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 union of each $\mathbf{h_n}$ with its piece of the evidence set forms a partition of the full set of evidence, **E**, as seen in Figure 9.

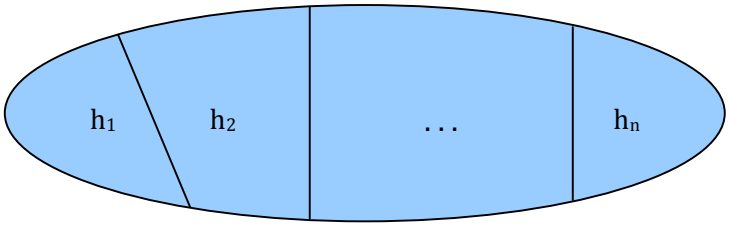


Figure 9. The set of evidence, **E**, is partitioned by the set of all possible currently known hypotheses, $\mathbf{h_n}$.

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

p(e ∩ h) = p(e|h) p(h)

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

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

$p(h_i|E)$ is the probability that a particular hypothesis, $\mathbf{h_i}$, 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.

With the general form of Bayes' theorem we have a functional (computational) description (model) for a particular situation happening given a set of perceptual evidence clues. Epistemologically, we have created on the right hand size of the equation 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 (numerical) values for all the hypotheses and evidence that can fit a 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 situations the sets of evidence are NOT independent, given the set of hypotheses. This makes the mathematical version of full Bayes just presented unjustified. When this independence assumption is simply ignored, as we see shortly, the result is called *naïve Bayes*. More often, however, the rationalization of the probability of the occurrence of evidence across all hypotheses is seen as simply a normalizing factor, supporting *the calculation of* a realistic 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, and thus, as in most natural language processing applications, is usually ignored.

A final objection asserts that diagnostic reasoning is not about the calculation of probabilities; it is about the determination of the determining the *most likely explanation*, given the accumulation of pieces of evidence. Humans are not doing real-time complex mathematical processing; 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 as well as the associated assumption of evidence independence. Naïve Bayes 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 many diagnostic situations we are 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 $\mathbf{p(E|h_i)\ p(h_i)}$:

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

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**:

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

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.

Thus, we view the evolving greatest likelihood relationship as a continuing tension between a set of 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 can either support or falsify 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, which is an important component of current research in machine learning (Luger 2009, Chapter 13). In the following section we present several computational examples from AI research of utilizing this greatest likelihood model for creating dynamic equilibration.

6. Computational Examples of Model Refinement and Equilibration

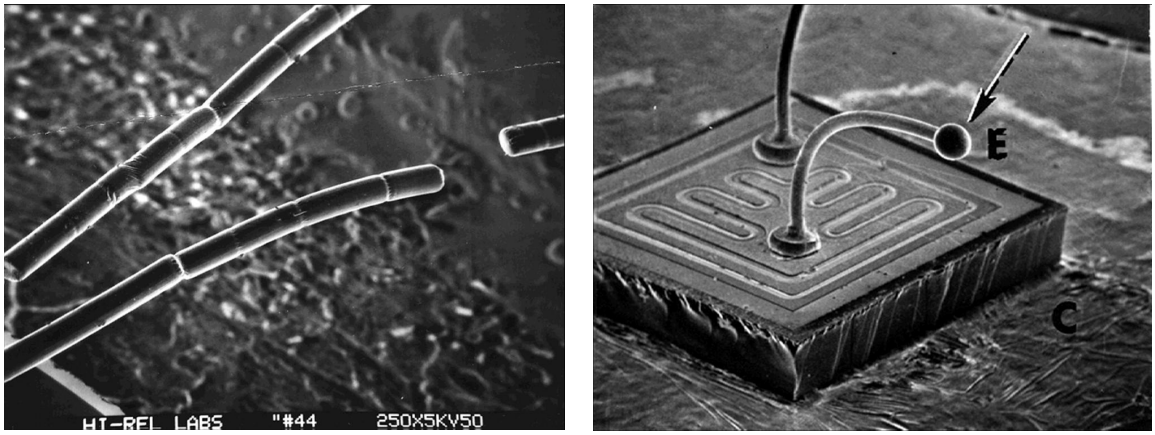
Probabilistic modeling tools have supported significant components of AI research since the 1950s when researchers at Bell Labs built a speech system that could recognize any of the ten digits spoken by a single speaker with accuracy in the high ninety percent range (Davis et al. 1952). Shortly after this Bledsoe and Browning (1959) built a Bayesian based letter recognition system that used a large dictionary that served as the corpus for recognizing hand written characters, given the likelihood of character sequences and of particular characters. Later research addressed authorship attribution by looking at the word patterns in anonymous literature and comparing them to similar patterns of known authors (Mosteller and Wallace 1964).

By the early 1990s, much of computational-based language understanding and production was stochastic, including parsing, part-of-speech tagging, reference resolution, and discourse processing, usually using tools like the *greatest likelihood* measures of the previous section (Jurafsky and Martin 2009). Many other areas of artificial intelligence, especially machine learning, became more Bayesian-based. In many ways these uses of

stochastic technology for pattern recognition were another instantiation of the behaviorist tradition, as collected sets of patterns were used to condition recognition of new patterns.

Judea Pearl's (1988) proposal for use of Bayesian belief nets (BBNs) and his assumption of their links reflecting "causal" relationships (Pearl 2000) brought the use of Bayesian technology to an entirely new importance. First, the assumption of these networks being directed graphs – reflecting causal relationships – and disallowing cycles – no entity can cause itself – brought a radical improvement in computational costs of such systems (Luger 2009, Chapter 5). Second, these same two assumptions made the BBN representation much more transparent as a representational tool that could capture causal relations. Finally, most all the traditional powerful stochastic representations used in language work and machine learning, for example, the hidden Markov model in the form of a dynamic Bayesian network (DBN), could be readily brought to this new representational formalism.

We next illustrate the BBN approach in several application domains. In the diagnosis of failures in discrete component semiconductors (Stern et al. 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 10, presenting two failures of discrete component semiconductors.



a. b.

Figure 10. Two examples of discrete component semiconductors, each exhibiting the "open" failure.

Figure 10 shows two examples of a failure type called an "open", or the break in a wire connecting components 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 a 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 10a. This suggests a revised greatest likelihood hypothesis that explains the open as a break created by metal crystallization that was likely caused by a sequence of low-frequency high-current pulses. The greatest likely hypothesis for the open of the example of Figure 10b, 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 et al. 1997) as well as reflected in a Bayesian belief net (Chakrabarti et al. 2005). Figure 11 presents a Bayesian belief net (BBN) capturing this and other related diagnostic situations.

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' analysis of known failures. Thus, the BBN can be said to capture a priori 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 explanations within its (a priori) model, given this new information. There are many inference rules for doing this (Luger 2009, Chapter 9). We 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 a second example, (Chakrabarti et al. 2005) analyze a continuous data stream from a set of distributed sensors. In monitoring of the "health" of the transmission of Navy helicopter rotor systems, a steady stream of sensor readings is analyzed; this data consists mainly of temperatures, vibrations, and pressure from the various components of the transmission running across time. An example of this data can be seen in the top portion of Figure 12, where this continuous data is broken into discrete and partial time slices.

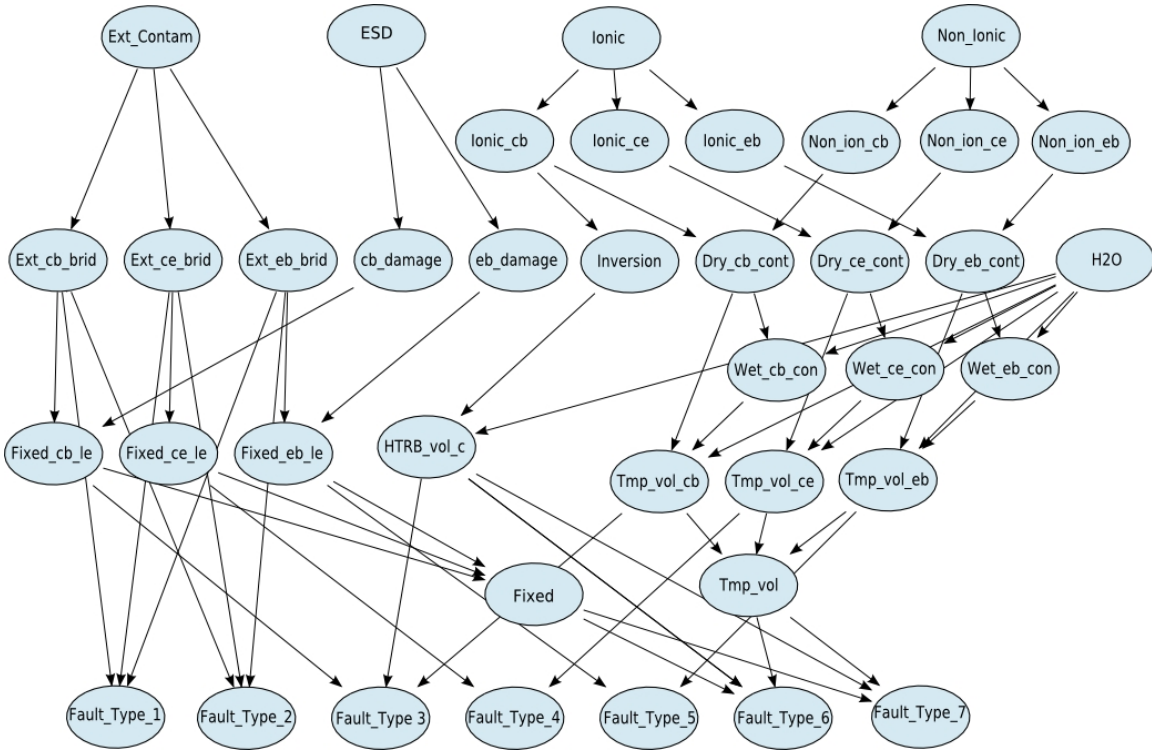


Figure 11. A Bayesian belief network representing the causal relationships and data points implicit in the discrete component semiconductor domain. As data is "discovered" the (a priori) probabilistic hypotheses change and suggest further search for data.

A Fourier transform[1] is then used to translate these signals into the frequency domain, as shown on the left side of the second row of Figure 12. 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 13. 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 states of the helicopter rotor system itself, as seen in the lower right of Figure 12.

The hidden Markov model (HMM) technology is an important stochastic technique that can be seen as a variant of a dynamic BBN. In the HMM, we attribute values to states of the network that are themselves not directly observable. For example, the HMM technique is widely used in the computer analysis of human speech, trying to determine the most likely word uttered, given a stream of acoustic signals. Training this system on streams of normal transmission data allows the system to make the correct greatest likelihood measure of failure when breakdown occurs. 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 **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 sets of data across time. Now we ask what we can be done when there is no interpretation of the model that fits the current data.

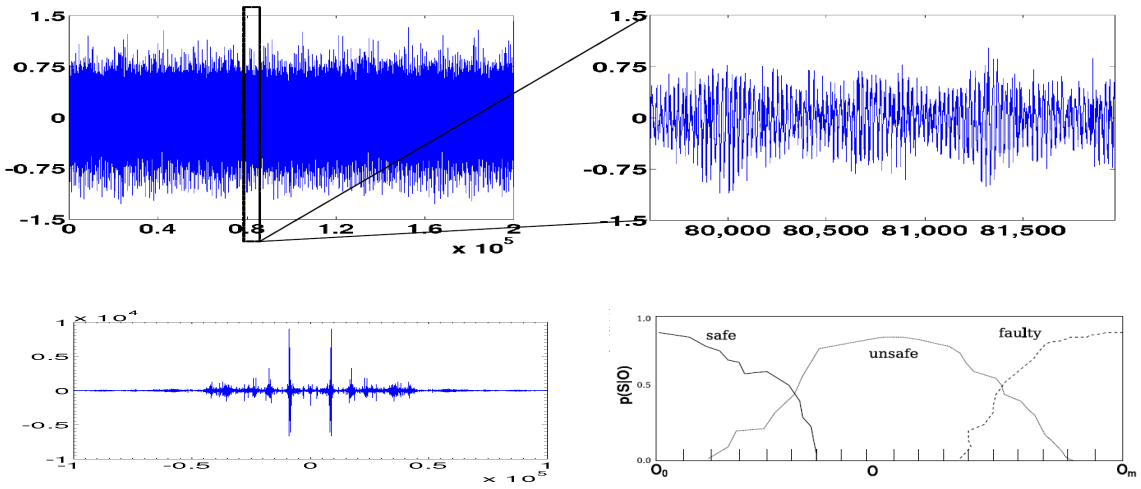


Figure 12. Real-time data from the transmission system of a helicopter's rotor. The top component of the figure presents the original data stream and an enlarged time slice. The lower left figure is the result of the Fourier transform of the time slice data (transformed) into the frequency domain. The lower right figure represents the hidden states of the helicopter rotor system.

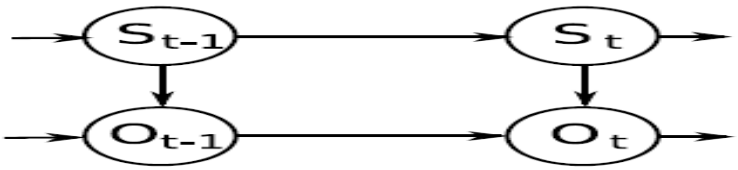


Figure 13. The data of Figure 12 is processed using an auto-regressive hidden Markov model. States $\mathbf{O_t}$ represent the observable values at time **t**. The $\mathbf{S_t}$ states represent the hidden "health" states of the rotor system, **{safe, unsafe, faulty}** at time **t**.

Figure 14 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 14 presents the COSMOS architecture (Sakanenko et al. 2009) that addresses both these tasks.

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 (Sakhanenko et al. 2009), we describe the model calibration idea in the 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.

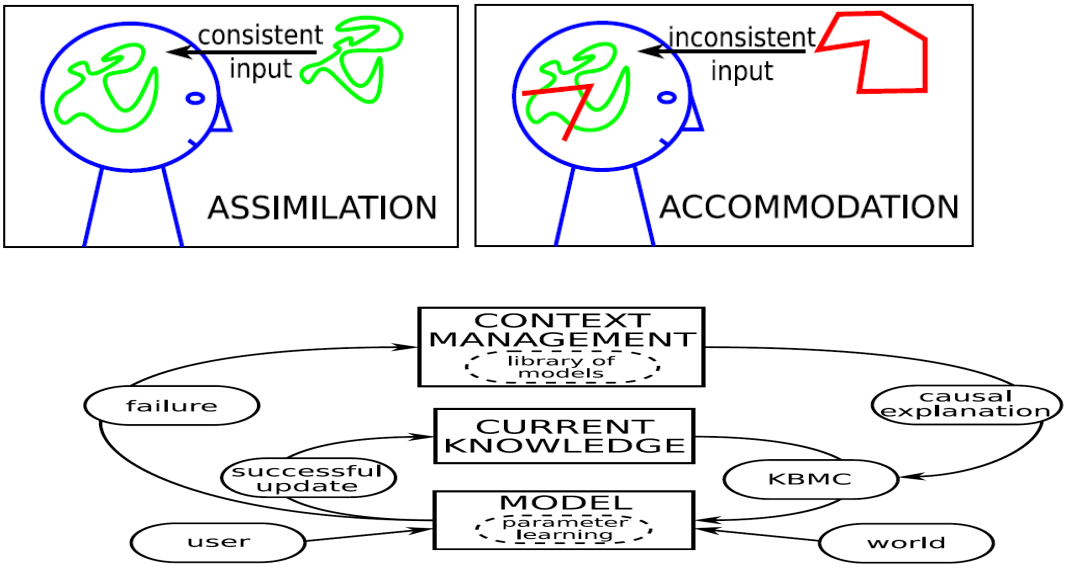


Figure 14. Cognitive model use and failure, above; a model-calibration algorithm, below, for assimilation and accommodation of new data.

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 spring monitoring (Sakhanenko et al. 2008).

Furthermore, when the alarm systems are sold in new environments it must be determined which of its library of models will best fit that new situation, or whether a new model must be discovered. If there are other important variables, such as many small earthquake tremors, this variable will probably also need to be modeled.

In this paper we considered the problem of model induction as a search for a most likely explanation, given a set of data. This, as noted, can be represented using Bayes theorem and a search for greatest likelihood parameters. There are many other examples in the literature of reasoning to the greatest likelihood for diagnostic and prognostic systems, for example, in computer based speech understanding where, with a series of n-grams, sounds and/or words, are tested against the contents of large corpora of language data. In these situations a probabilistic form of dynamic programming, often referred to as the Viterbi algorithm (Jurasky and Martin 2009), process continuous streams of data to produce greatest likelihood results, given current data. Another example is researchers attempting to determine what cortical connectivity, represented as a dynamic Bayesian network, might best explain sets of fMRI images (Burge et al 2010).

A further approach to model inductive inference was proposed by Pearl (2000) with his *Calculus of Interventions*, explored further in research with his students (Tian and Pearl 2001). In this approach, researchers force variables in a model to take on specific values, to determine causal relationships within that model. A simple example of this approach would be an infant throwing things from her highchair to see which items bounced best (or upset her father most). Although the problem of model induction in general is intractable, in many knowledge intensive situations useful new models can be created, using ideas such as calculated interventions and the greedy local search of constraints located near the points of model failure (Rammohan 2010).

In the final section we offer some conjectures about possible cognitive architectures that can support the computational calculation of the greatest likelihood schemas.

7. Conclusion: An Epistemological Stance

Turing's test for intelligence was totally agnostic both as to what the computer was composed of or the language used to make it run. It simply required the responses of the machine to be roughly equivalent to the responses of humans in the same situations. Tuting's test was not about whether the computer was made of vacuum tubes, flip-flops, protoplasm, or even tinker toys, but what it could produce. To build Turing's computational system has, however, required researchers to commit to specific data structures and search algorithms.

We have seen that researchers building out Turing's machine have also committed to various epistemological stances, most notably forms of empiricist, rationalist, or structuralist-stochastic traditions.

Although Turing is again agnostic about any sort of epistemological stance, an important subset of AI researchers since Turing have taken up such a stance, and see it as a critical component of their creating intelligent artifacts. Notable among these are Newell and Simon (1976) with their symbol system hypothesis and Sejnowski's (Meltzoff et al 2009) work in neural networks and computational neuroscience. This paper suggests that that the structuralist-stochastic tradition could have a similar role in modeling human cognition.

There now exist a number of algorithms created for real-time integration of posterior information and its propagation into (a priori) stochastic models as well as many computational inference rules for calculating the greatest likelihood in probabilistic modeling situations. An attractive choice among these is the *loopy belief propagation* algorithm (Pearl 1989) as it reflects a system constantly iterating towards equilibrium, or *equilibration*, as Piaget might describe it. A cognitive system can be in *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, often with the support for a particular 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, until 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, unobservable, or obscured data, given the state of a priori equilibrium. Sakhanenko et al. (2008) show this iterative process to be a form of expectation-maximization (EM) learning (Dempster et al. 1977). Furthermore, we suggest that a loopy belief propagation algorithm iterating to equilibrium when the original a priori belief state encounters a posteriori stimulation reflects an essential nexus between the embodied mind/brain and its environment that is compatible with current epistemological positions, including Nozick and *tracking* (1981), the *content externalism* and *knowledge as evidence* of Williamson (2000), as well as the anti-dualism of Rorty (1999).

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

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

At least as important as Turing's 1950 paper, his work from the 1930s through the early 1950s created the foundations for and suggested some of the limitations of the modern sciences of computing, artificial intelligence, as well as computational neuroscience. Turing answered Hilbert's challenge in the famous *Entscheidungsproblem* by formally specifying

what it *meant* to compute. Further, Turing, in proposing the *halting problem*, following the arguments of Godel (1931), demonstrated that there were queries within any system as rich as Peano's axioms, which a computing machine *could not* answer (Davis et al. 1976). Finally, along with Alonzo Church (1941), Alan Turing demonstrated that his Turing Machine was equivalent to other models of computation and offered a maximally powerful example of what *can be computed.* We await the answer as to whether the human processor falls within the theoretical limits of this Church-Turing thesis. Thanks to Alan Turing, however, we can now, coherently, address these questions. We end with the still cogent quote from Turing, the final sentence of his 1950s challenge first proposed in *Mind*:

We can see only a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing, Computing Machinery and Intelligence, *Mind*, 1950.

[1]The **Fourier transform** (often abbreviated **FT**) is an operation that transforms one complex-valued function of a real variable into another. In such applications as signal processing, the domain of the original function is typically time and is accordingly called the time domain. The domain of the new function is frequency, and so the Fourier transform is often called the frequency domain representation of the original function. It describes which frequencies are present in the original time function. The term Fourier transform refers both to the frequency domain representation of a function and to the process or formula that "transforms" one function into the other. **Fourier Transform**, *Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/Main_Page, 9/9/9.

Acknowledgments: The equations developed in Section 5 may be found in many texts introducing probabilistic methods. For further support for the integration of the empiricist and rationalist traditions in cognitive psychology see Piaget's *Structuralism* (1970). The hypothesis that dynamic Bayesian networks can model a structuralist epistemological stance is developed further elsewhere (Author 2012). This research was supported in part by the National Science Foundation, the Office of Naval Research, and the Air Force Research Laboratory.

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.

Bledsoe, W. W. and Browning, IO. (1959), 'Pattern Recognition and Reading by Machine', *1959 Proceedings of the Eastern Joint Computer Conference*, p 225-232, New York: Academic Press.

Brooks, R. A. (1989), 'A Robot that Walks: Emergent Behaviors from a Carefully Evolved Network', *Neural Computation* 1(2):253-262.

Brooks, R. A. (1991), 'Intelligence without Representation', *Proceedings of IJCAI-91*, pp 596-575. San Mateo CA: Morgan Kaufmann.

Buchanan, B. G. and Shortliffe, E. H. (eds), (1984), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Reading M: Addison-Wesley.

Burge, J., Lane, T., Link, H., Qiu, S., and Clark, V. P. (2007), 'Discrete Dynamic Bayesian Network Analysis of fMRI Data', *Human Brain Mapping* 30(1), pp 122–137.

Chakrabarti, C., Rammohan, R., and Luger, G. F. (2005), 'A First-Order Stochastic Modeling Language for Diagnosis' *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).

Chomsky, N. (1957), *Syntactic Structures*, The Hague and Paris: Mouton.

Church A. (1941), 'A Note on the Entscheidungsproblem', *Journal of Symbolic Logic, 1*, 40-41 and 101-102.

Collins, A. and Quillian, M. R. (1969), Retreival Time from Semantic Memory', *Journal of Verbal Learning & Verbal Behavior*, 8: 240-247.

Davis, K. H., Biddulph, R. and Balashek, S. (1952), 'Automatic Recognition of Spoken Digits', *JASA* 24(6), 637-642.

Davis, M., Matijasevic, Y., and Robinson, A. (1976), 'Hilbert's Tenth Problem: Diophantine Equations; Positive Aspects of a Negative Solution', *Proceedings of Symposia in Pure Mathematics*, 28:323-378.

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.

Eco, U. (1976), *A Theory of Semiotics*, Bloomington IN: University of Indiana Press.

Fikes, R. E. and Nilsson, N. J. (1971), 'STRIPS: A New Approach to the Application of Theorem Proving to Artificial Intelligence', *Artificial Intelligence*, 1(2), 227-232.

Godel, K. (1931), *On Formally Undecidable Propositions*, New York: Basic Books.

Harmon, P. and King, D. (1985), *Expert Systems: Artificial Intelligence in Business*. New York: Wiley.

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

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

Jurasky, D. and Martin, J. M. (2009*), Speech and Language Processing*, Upper Saddle River NJ: Pearson Education.

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

Kolodner, J. L. (1993), *Case-based Reasoning*, San Mateo CA: Morgan Kaufmann.

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

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.

Masterman, M. (1961), 'Semantic Message Detection for Machine Translation, using Interlingua', *Proceedings of the 1961 International Conference on Machine Translation*.

McCarthy, J. (1968), 'Programs with Common Sense', *Semantic Information Processing*, Minsky (ed), Cambridge MA: MIT Press.

McCarthy, J. (1980), 'Circumscription, a Form of Non-monotonic Reasoning', *Artificial Intelligence*, 13, 27-39.

McCarthy, J. (1986), 'Applications of Circumscription to Formalizing Common-Sense Knowledge', *Artificial Intelligence*, 28:89-116.

McCarthy, J. and Hayes, P. (1969), 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', *Machine Intelligence 4*, Meltzer and Michie (eds), Edinburgh UK: The University Press.

Meltzoff, A. N., Kuhl, P. K., Movellan, J., Sejnowski, T. J. (2009), 'Foundations for a New Science of Learning', *Science* 325: 284-288.

Mosteller, F. and Wallace, D. L. (1964), *Inference in Disputed Authorship: The Fereralist*, Berlin: Springer-Verlag.

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

Newell, A. and Simon, H.A. (1976), 'Computer Science as Empirical Enquiry: Symbols and Search', *Communications of the ACM*, 19(3): 113–126.

Nozick, R. (1981), *Philosophical Explanations*, Cambridge MA: Harvard University Press.

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

Pearl, J. (2000), *Causality*, Cambridge UK: Cambridge 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.

Post, E. (1943), 'Formal Reductions of the General Combinatorial Problem', *American Journal of Mathematics*, 65: 197-268.

Quillian, M. R. (1967), 'Word Concepts: A Theory and Simulation of some Basic Semantic Capabilities', Brachman R. J. and Levesque, H. J. (1975), *Readings in Knowledge Representation,* Los Altos CA: Morgan Kaufmann.

Rorty, R. (1999), *Philosophy and Social Hope*, London: Penguin Books.

Rosenbloom, P. S., Lehman, J. F., and Laird, J.E. (1993), 'Overview of SOAR as a Unified Theory of Cognition: Spring 1993', *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, Hillside NJ: Erlbaum.

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

Schank, R. C. and Colby, K. M. (1975), *Computer Models of Thought and Language*. San Francisco: Freedman.

Schvaneveldt, R. W. and Cohen, T. A. (2011), 'Abductive Reasoning and Similarity: Some Computational Tools', D. Ifenthaler, P. Pirnay-Dummer, and N. M. Seel (Eds.), *Computer Based Diagnostics and Systematic Analysis of Knowledge,* New York: Springer.

Sebeok, T. A. (1985), *Contributions to the Doctrine of Signs*, Lanham MD: The Press of America.

Simon, H. A. (1981), *The Sciences of the Artificial (2nd ed)*, Cambridge MA: MIT Press.

Stern, C.R. and Luger, G.F. (1997), 'Abduction and Abstraction in Diagnosis: A Schema-Based Account', *Android Epistemology*, Ford, K. et al., eds, Cambridge MA: MIT Press.

Tian, J. and Pearl, J. (2001), 'Causal Discovery from Changes', *Proceedings of UAI 2001*, p. 512-521. Burlington MA: Morgan Kaufmann

Turing, A. (1936), 'On Computable Numbers with an Application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, 2nd series. 42, 230-265.

Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind*, 59, 433-460.

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

Wilks, Y. (1972), *Grammar, Meaning, and the Machine Analysis of Language*, London: Routledge and Kagen Paul.

Williams, B. C. and Nayak, P. P. (1996), 'A Model-Based Approach to Reactive Self-Reconfiguring Systems', *Proceedings of the AAAI-96*, 971-978, Cambridge MA: MIT Press.

Williams, B. C. and Nayak, P. P. (1997), 'A Reactive Planner for a Model-Based Executive', *Proceedings of IJCAI-97*, Cambridge MA: MIT Press.

Williamson, T. (2000), *Knowledge and its Limits*, Oxford UK: The University Press.

Wittgenstein, L. (1953), *Philosophical Investigations*, New York: Macmillan.