Modern AI: Stochastic Models and an Epistemological Stance

George F. Luger and Chayan Chakrabarti
Department of Computer Science, University of New Mexico, Albuquerque, NM
luger | cc@cs.unm.edu

Abstract

More than sixty years ago Alan Turing 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 a computer could be said to be "intelligent". 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 work: How might the Artificial Intelligence practitioner implement "intelligence" on a computational device?

We contend that all engineering software artifacts build out the implicit epistemological stance assumed by their designers. What this means in practice is that software engineers view their problem-solving task from a particular perspective: selecting both representational devices and solution algorithms that reflect their perspective on how the problem should be perceived, understood, and solved. In the artificial intelligence research area this has led to successes in many domains of limited scope.

We consider the stochastic approaches of modern AI researchers from this epistemological perspective: how do these probabilistic programs represent their intended solution domain and are they any more likely than earlier approaches to reflect a path towards the creation of generally intelligent programs.

1. Introduction: The Question of General Intelligence

More than sixty years ago Alan Turing (1950) asked whether electronic computers could be generally intelligent. No informed critic would contend that 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) (1/4/14). Commercially available programs addressing the quest for general intelligence include web chat bots, such as Apple's Siri. The Turing challenge remains an annual event, and the interested reader may visit http://en.wikipedia.org/wiki/Turing_test#Loebner_Prize, (1/4/14), for details.

In fact, the AI community often uses forms of Turing's 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 it?

This paper considers these issues, especially the responses to the challenge of building intelligent artifacts that the modern artificial intelligence community has taken. In earlier venues (Luger 2012, Luger and Chakrabarti 2014) we presented general issues of artificial intelligence artifacts and the epistemological biases they embody. In this paper we view modern artificial intelligence, especially the commitment to stochastic models, and the epistemological stance that supports this approach. In the next section we present a constructivist rapprochement of the empiricist, rationalist, and pragmatist positions that supported early AI work and addresses many of its dualist assumptions. We also offer some preliminary conjectures about how a Bayesian model might be epistemologically plausible.

2. Modern AI: Probabilistic Models

We view a constructivist and model-revising epistemology as a rapprochement between the empiricist, rationalist, and pragmatist viewpoints. The constructivist hypothesizes that all human understanding is the result of an interaction between energy patterns in the world and mental categories imposed on the world by an intelligent agent (Piaget 1954, 1970; von Glasersfeld 1978). Using Piaget's terms, we humans *assimilate* external phenomena according to our current understanding and *accommodate* our understanding to phenomena that do 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 writing of 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. There are many current psychologists and philosophers that support and expand this pragmatic and teleological account of human developmental activity (Glymour 2001, Gopnik et al. 2010, Gopnik 2011a, 2011b, Kushnir et al. 2010).

Perceived information, Kant's *a posteriori* knowledge, rarely fits precisely into our preconceived and *a priori* schemata. From this tension to comprehend and act as an agent, the schema-based biases a subject uses to organize experience are strengthened, modified, or replaced. This *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* (Luger 2009, Ch 16).

When Piaget (1970) proposed a constructivist approach to a child's 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*.

There is a blending here of empiricist and rationalist traditions, mediated by the pragmatist 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. A Bayesian model-refinement representation offers an appropriate model for critical components of this constructivist model-revising epistemological stance (Luger et al. 2002, Luger 2012). Interestingly enough, David Hume acknowledged the epistemic foundation of all human activity (including, of course, the construction of AI artifacts) in *A Treatise on Human Nature* (1739/2000) when he stated "All the sciences have a relation, greater or less, to human nature; and ... however wide any of them may seem to run from it, they still return back by one passage or another. Even Mathematics, Natural Philosophy, and Natural Religion, are in some measure dependent on the science of MAN; since they lie under the cognizance of men, and are judged of by their powers and faculties."

Thus, we can ask why a constructivist epistemology might be useful in addressing the problem of building programs that are "intelligent". How 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 access to 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 direct access to physical phenomena and 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.

Bayes theorem (1763) offers a plausible model of this constructivist rapprochement between the philosophical traditions we have just discussed. It is also an important modeling tool for much of modern AI, including AI programs for natural language understanding, robotics, and machine learning. With a high-level discussion of Bayes' insights, we can describe the power of this approach.

Consider the general form of Bayes' relationship used to determine the probability of a particular hypothesis, $\mathbf{h_i}$, given a set of evidence **E**:

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

$\mathbf{p(h_i|E)}$ is the probability that a particular hypothesis, $\mathbf{h_i}$, is true given evidence **E.**
$\mathbf{p(h_i)}$ is the probability that $\mathbf{h_i}$ is true overall.
$\mathbf{p(E|h_i)}$ is the probability of observing evidence **E** when $\mathbf{h_i}$ is true.
n is the number of possible hypotheses.

With the general form of Bayes' theorem we have a functional (and 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 (is interpreted by) the patterns 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 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 hundreds of 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 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. Thus, 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 $\mathbf{p(E|h_i)\ p(h_i)}$:

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

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 $\mathbf{t}$:

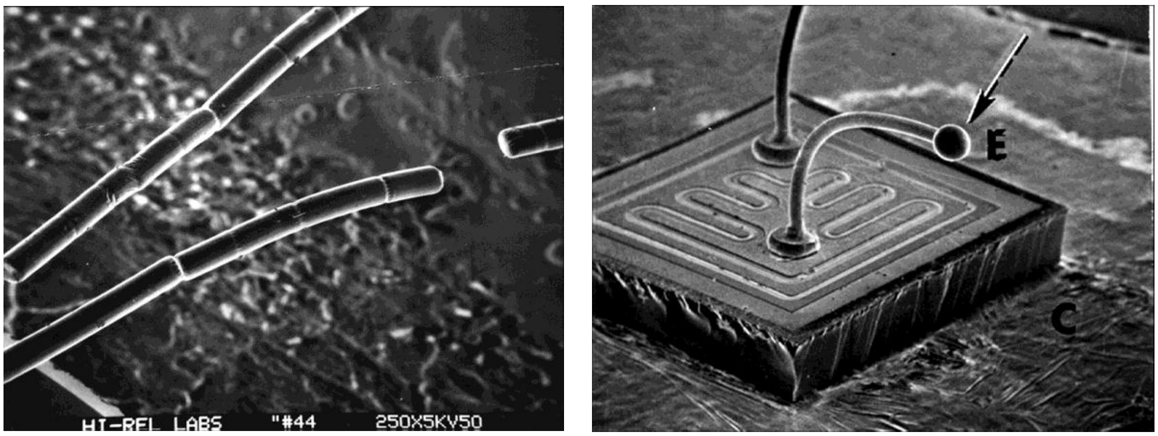
$$\mathbf{gl(h_i|E_t) = argmax(h_i)\ p(E_t|h_i)\ p(h_i)}$$

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

By the early 1990s, much of computation-based language understanding and generation was stochastic, including parsing, part-of-speech tagging, reference resolution, and discourse processing, usually using tools like *greatest likelihood* measures (Jurafsky and Martin 2009). 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 constructivist 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 to the computational costs of reasoning with BBNs (Luger 2009, Ch 9). 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 integrated into this new representational formalism.

We next illustrate the Bayesian approach in two 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 1, presenting two failures of discrete component semiconductors. The failure type is 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 (perceptual) sounds or smells when it happened? What were the resulting conditions of the components of the system?



a. b.

Figure 1. Two examples of discrete component semiconductors, each exhibiting the "open" or "connection broken" failure.

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 1a. 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 1b, 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 2 presents a Bayesian belief net (BBN) capturing these 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 explanation, 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 within the BBN.

This current example demonstrates how the most likely current hypothesis can be used to determine the best explanation, given a particular time and an hypothesis space. We next demonstrate how considering sets of hypotheses and data across time, using the most likely hypothesis at time **t**, can produce a greatest likelihood hypothesis.

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

In this model 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. If we want to expand this to the next time period, **t + 1**, we need to describe how models can evolve across time.

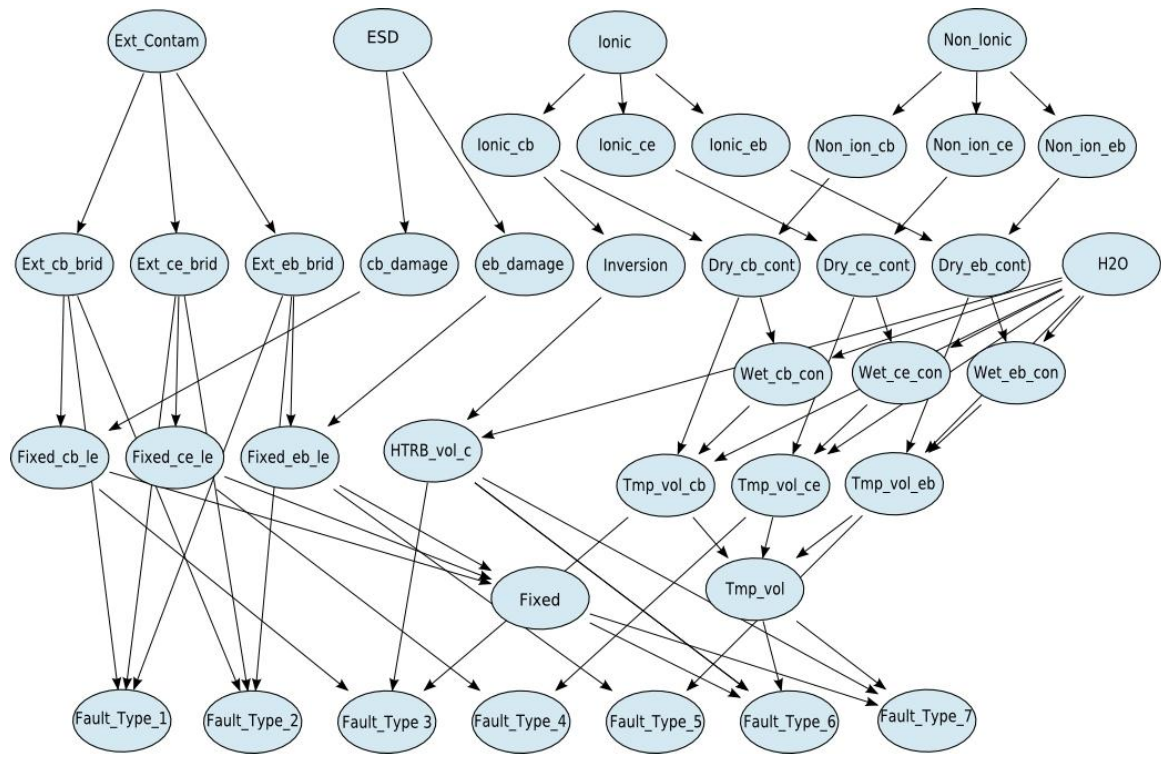


Figure 2. A Bayesian belief network representing the causal relationships and data points implicit in the discrete component semiconductor domain. As data is "discovered" the (a priori) probabilistic hypotheses change and suggest further search for data.

As an example of **argmax** processing, Chakrabarti et al. (2005, 2007) analyze a continuous data stream from a set of distributed sensors. The running "health" of the transmission of a Navy helicopter rotor system is represented by a steady stream of sensor data. This data consists of temperature, vibration, pressure, and other measurements reflecting the state of the various components of the running transmission system. An example of this data can be seen in the top portion of Figure 3, where the continuous data stream is broken into discrete and partial time slices.

A Fourier transform is then used to translate these signals into the frequency domain, as shown on the left side of the second row of Figure 3. These frequency readings were compared across time periods to diagnose the running health of the rotor system. The model used to diagnose rotor health the *auto-regressive hidden Markov model* (A-RHMM) of Figure 4. 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 3.

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 (Jurasky and Martin 2009). In our helicopter example, training this system on streams of normal transmission data allowd the system to make the correct greatest likelihood measure of failure when these signals change to indicate a possible breakdown. The US Navy supplied data to train the normal running system as well data sets for transmissions that contained seeded faults. 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 any time **t.**

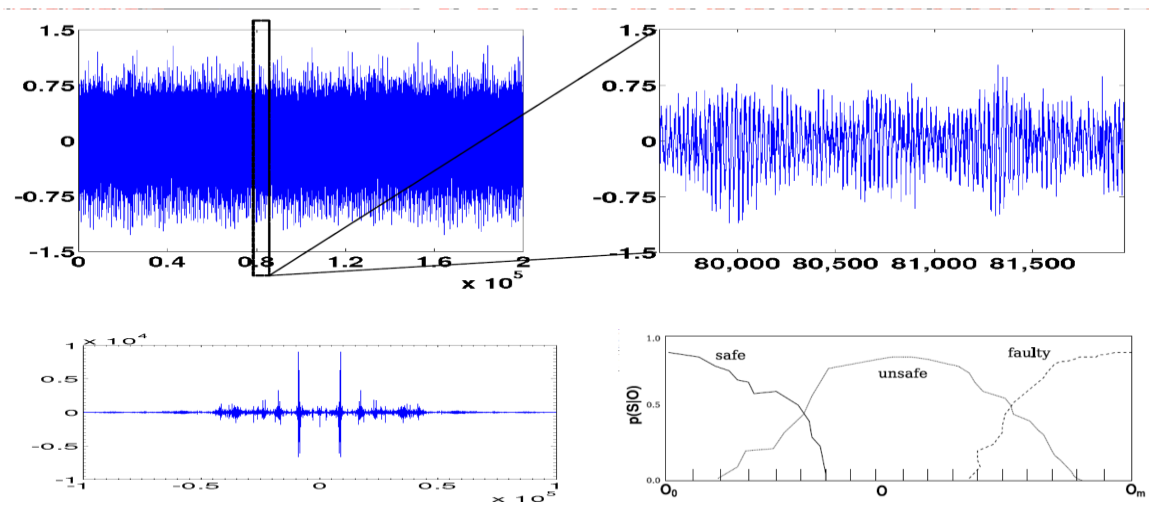


Figure 3. Real-time data from the transmission system of a helicopter's rotor. The top component of the figure presents the original data stream (left) and an enlarged time slice (right). The lower left figure is the result of the Fourier transform of the time slice data (transformed) into the frequency domain. The lower right figure represents the hidden states of the helicopter rotor system.

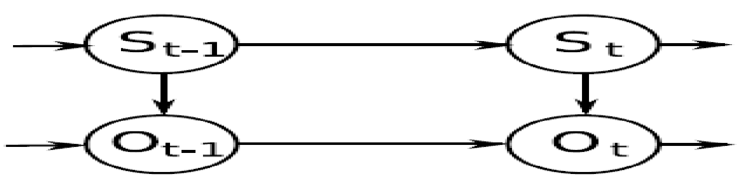


Figure 4. The data of Figure 3 is processed using an auto-regressive hidden Markov model. States $\mathbf{O_t}$ represent the observable values at time **t**. The $\mathbf{S_t}$ states represent the hidden "health" states of the rotor system, **{safe, unsafe, faulty}** at time **t**.

3. Conclusion: An Epistemological Stance

Turing's test for intelligence was agnostic both as to what a 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.

Modern AI research has proposed algorithms for the real-time integration of new (a posteriori) information into previously (a priori) learned patterns of information (Dempster 1968). Among these algorithms is *loopy belief propagation* (Pearl 1988, 2000) that captures a system of plausible beliefs 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 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, in the form of a particular greatest likelihood hypothesis.

References

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.

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).

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

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, MIT Press, 2001

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): 3-32.

Gopnik, A. (2011a). A unified account of abstract structure and conceptual change: Probabilistic models and early learning mechanisms. Commentary on Susan Carey "The Origin of Concepts." *Behavioral and Brain Sciences* 34,3:126-129.

Gopnik, A. (2011b). Probabilistic models as theories of children's minds. *Behavioral and Brain Sciences* 34,4:200-201.

Hume, D. (1739/2000). *A Treatise of Human Nature*, edited by D. F. Norton and M. J. Norton, Oxford/New York: Oxford University Press.

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.

Kushnir, T. Gopnik, A. Lucas, C. and Schulz L. (2010). Inferring hidden causal structure. *Cognitive Science* 34:148-160.

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

Luger, G. F. (2012), Epistemology, Access, and Computational Models. In *The Complex Mind*, McFarland, Stenning, and McGonigle Chalmbers, editors. London: Palgrave Macmillan.

Luger, G. F. and Chakrabarti, C. (2014), *From Alan Turing to Modern AI: An Epistemological Stance* (in submission), copies available from first author.

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.

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.

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. In Android Epistemology, Ford et al. eds, Cambridge MA: MIT Press.

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.