Smith, Robert Elliott

Article

Idealizations of Uncertainty, and Lessons from Artificial Intelligence

Economics: The Open-Access, Open-Assessment E-Journal

Provided in Cooperation with:
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Smith, Robert Elliott (2016) : Idealizations of Uncertainty, and Lessons from Artificial Intelligence, Economics: The Open-Access, Open-Assessment E-Journal, ISSN 1864-6042, Kiel Institute for the World Economy (IfW), Kiel, Vol. 10, Iss. 2016-7, pp. 1-40, https://doi.org/10.5018/economics-ejournal.ja.2016-7

This Version is available at:
https://hdl.handle.net/10419/129754

Economics
The Open-Access, Open-Assessment E-Journal

Vol. 10, 2016-7 | March 21, 2016 | http://dx.doi.org/10.5018/economics-ejournal.ja.2016-7

Idealizations of Uncertainty, and Lessons from Artificial Intelligence

Robert Elliott Smith

Abstract

At a time when economics is giving intense scrutiny to the likely impact of artificial intelligence (AI) on the global economy, this paper suggests the two disciplines face a common problem when it comes to uncertainty. It is argued that, despite the enormous achievements of AI systems, it would be a serious mistake to suppose that such systems, unaided by human intervention, are as yet any nearer to providing robust solutions to the problems posed by Keynesian uncertainty. Under the radically uncertain conditions, human decision-making (for all its problems) has proved relatively robust, while decision making relying solely on deterministic rules or probabilistic models is bound to be brittle. AI remains dependent on techniques that are seldom seen in human decision-making, including assumptions of fully enumerable spaces of future possibilities, which are rigorously computed over, and extensively searched. Discussion of alternative models of human decision making under uncertainty follows, suggesting a future research agenda in this area of common interest to AI and economics.

(Published in Special Issue Radical Uncertainty and Its Implications for Economics)

JEL B59

Keywords Uncertainty; probability; Bayesian; artificial intelligence

Authors

Robert Elliott Smith, ✉ University College London, Department of Computer Science, UK, robert.elliott.smith@gmail.com

Citation Robert Elliott Smith (2016). Idealizations of Uncertainty, and Lessons from Artificial Intelligence. *Economics: The Open-Access, Open-Assessment E-Journal*, 10 (2016-7): 1—40. http://dx.doi.org/10.5018/economics-ejournal.ja.2016-7

Received May 29, 2015 **Published as Economics Discussion Paper** July 13, 2015
Accepted March 14, 2016 **Published** March 21, 2016

1 Introduction

Economics and artificial intelligence (AI) research are both concerned with modelling human decision making. These common interests are of heightening importance as discussion is becoming more widespread as to how far human economic activity will be replaced by AI (Schwab, 2016). Interestingly, AI techniques are being used in that discussion to determine which jobs are under threat (Frey and Osborne, 2016a), ironically illustrating that AI is already becoming part of economic decision making.

Historically, AI has proceeded on both descriptive and prescriptive agendas, sometimes in silos. The former relates to efforts based on breaking down tasks to try to get machines to make decisions in the way humans do, and the latter focuses on performing tasks, whether or not the methods chosen are similar to human reasoning and behaviour. Success using decision tree approaches are an example of the former, and recent efforts to translate text from one language to another are an example of the latter.

Reviewing both lines of research, this paper argues that accurate speculation about how far AI can go in replacing humans must be informed by a clear understanding of what differs between human and mechanised decision-making. The paper will describe how functioning under uncertainty (i.e., a system's ability to robustly cope with future situations that were unforeseen by AI designers) has been the primary historical challenge for AI, and makes the point that this is likely to continue to be so.

Adopting a model of human decision-making lies at the core of both AI and economics. Economic models typically assume that human agents gather and process information about the alternative choices in any given context. Although that information may be conceived as complete or, in an important extension of the basic theory, incomplete, wider questions also exist as to how far and with what consequences rational agents can process it sufficiently and attach the appropriate meaning to it in conditions of uncertainty such as those defined by Knight (1921) or Keynes (1936).

Simon (1955, 1978) made a major contribution with his recognition that economics had been relatively silent on the questions of either the ability of human agents to process information, or how that processing is carried out. He raised crucial issues about the first of these observations in his economic theories of bounded rationality and satisficing, while his non-economic research was focused on the second issue: how human agents process information (Newell and Simon, 1972). This latter research, on modelling human decision-making processes and programming machines along the same lines, is the essence of what is referred to as descriptive AI (although, as is discussed, the ideas introduced by Simon rapidly evolved towards more prescriptive AI in practice). The means by which human agents process information to make their decisions and how to model these decisions is also at the core of economic analysis, and that is the reason that Simon's work straddled these two fields.

This paper will argue that while prescriptive AI systems have been created that are effective for many engineered domains as aids for human decision makers, successes of these uses of AI should not obscure the difficulties AI has had as a descriptive science intended to capture the robustness of human decision-making. The primary failure modes of AI have been brittleness (in deductive systems) and poor generalisation (in inductive systems). Brittleness and poor generalisation of AI (and thus of the models AI uses to characterise human decision-making) are directly related to uncertainty about the details of future decision-making situations. They are failures of pre-programmed systems to cope with circumstances unforeseen by their designers (that is to say, uncertainty about future conditions). In AI computationally intensive search and optimisation methods (benefitting from the ever larger memory and processing capabilities predicted in Moore's Law) have led to success in areas where problems can be well defined at the outset, so that uncertainty is constrained. However, computational brute force cannot overcome the problems posed by situations in which aspects of the world are continually changing, whether due to innovation (producing an unstable environment) or due to new emerging constellations of interdependent events (such as people combining and doing things in unforeseen ways). Under uncertainty of these sorts all current forms of AI prove brittle.

The unsolved problems of AI modelling just mentioned have not always been made clear, partly because in AI, as in Economics, widespread assumptions have

often been made without taking the impact of uncertainty (rather than risk) adequately into account. Recent statements endorsed by prominent scientists seem to imply the probabilistic models used in AI can model human decision-making behaviour effectively so that, so to speak, human-like AI is very near to realisation (Future of Life Institute, 2015):

> "Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents-systems that perceive and act in some environment. In this context, 'intelligence' is related to statistical and economic notions of rationality-colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilisation among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems."

This quote's comments on remarkable successes in the listed domains are certainly true. However, this paper will argue that the early challenges of AI remain unaddressed, despite advances in statistical learning and other probabilistically-based techniques. Valuable as the underlying techniques are proving to be, the modelling of probabilities (and related, statistical calculations over "big data") cannot solve the longstanding issues of brittleness and poor generalisation of AI in conditions of significant uncertainty. This paper argues that such approaches do not fundamentally overcome the persistent challenges of modelling human decision-making. In fact, the paper will argue that they only augment existing AI paradigms slightly and do not represent a fundamental break from the models of the past.

2 Overview

The remainder of the paper proceeds by reviewing the basic structure of past AI models, starting with deductive models that were based on human-programmed rule bases. Such systems reveal two challenges: brittleness of finite rule bases in the face of unforeseen circumstances, and exponential difficulty in incrementally refining rule bases to overcome that brittleness. Inductive (learning) AI systems are reviewed, along with their key challenge: selecting training data and a representation which results in robust generalisation over unforeseen data. Both these deductive and inductive AI challenges are discussed as a common failure to robustly cope with the unforeseen. Examples are also offered of another problem in both branches of AI: the tendency to wishfully label algorithmic techniques with the names of aspects of human decision-making apparatus, despite a lack of evidence of real similarity. Such labels have been called *wishful mnemonics* (McDermott, 1976), and they have been a part of cycles of hype and disappointment in AI research, and thus deserve careful consideration in evaluating the realities and impacts of computerised decision-making models

The primary method of coping with the unforeseen in modern AI systems is through probabilistic modelling and statistical learning. A discussion of the fundamentals of these techniques is provided, which shows that such systems are structurally similar to systems of the past, and, therefore, can be expected to have similar challenges. Regardless of these anticipated difficulties, there are arguments that such models are ideal representations of human decision making under uncertainty, including the those that assume Bayes' rule as a model of subjective probabilities in human decision making, and those that assume Bayesian representations are related to the fundamental nature of communications and physics (and thus evolved into humans). Later sections of this paper provide arguments to shed doubt on these assumptions, indicating that they may be as loaded with wishful mnemonics as AI systems of the past, and thus subject to the same potential problems.

The paper concludes by reviewing alternative perspectives on human decision making under uncertainty, and discussion of how such alternatives may provide a new agenda for AI research, and for the modelling of human decision making in economics, and other fields.

3 Models of Human Decision-Making

Attempts to model human decision-making have a rich history that starts with the most fundamental logic and mathematics, and continues through the development of computational mechanisms. Ultimately this history leads to the construction of modelling human reasoning in computers, through what has come to be known as artificial intelligence (AI). The history of AI has close ties to developments in economics, since economics is, at its core, about the decisions of human actors (Mirowski, 2002).

AI has a history of advances and dramatic, characteristic failures. These failures are seldom examined for their implications regarding the modelling of human reasoning in general. The uncertain circumstances and unforeseen events that human reasoning must cope with are an intimate part of the failure modes of past AI. Thus, it is unsurprising that AI has become intimately tied to probability and statistical reasoning models, as those are the dominant models of the uncertain and unforeseen future. The following section reviews some of the history of past AI models so that they can be compared to current probabilistic models of human reasoning.

3.1 Deductive Models

Consider what are called "the first AI programs": *General Problem Solver* (GPS) and *Logic Theory Machine* (LTM), created by Simon, Newell, and others (Newell and Simon, 1972). GPS and LTM work in similar manners, focusing on the idea of reasoning as a search (means-ends analysis) through a tree of alternatives. The root is an initial hypothesis or a current state of the world. Each branch was a "move" or deduction based on rules of logic. An *inference engine* employs logical rules to search through the tree towards a conclusion. The pathway along the branches that led to the goal was the concluding chain of reasoning or actions. Thus, the overall *deductive AI* methodology has of the following form:

- Model any given problem formally, creating a *knowledge representation* of the domain at hand, such that the search to satisfy goals is modelled as a formal structure and a logical search through that structure.

- Apply logical (or mathematical) search algorithms to find conclusions based on the current state of this knowledge representation.

The split here is important: knowledge is modelled for a specific domain, and search is a generalised, logical procedure, that can be applied to any domain. The generality of the logical procedures is the essential aspect that leads to the description of these techniques as AI models, rather than simply domain-specific programs, in that a general (logically deductive) theory of reasoning is assumed. For extensibility of this methodology, it must be possible to modify the knowledge representation independent of the search algorithms, if the technique is to be considered a general-purpose model of human decision-making.

However, this model leads directly to problems of limited computational power and time. Even in relatively simple-to-describe problems, the search tree can grow exponentially, making the search intractable, even at the superhuman speeds of modern computers. This leads to a key question which is relevant to both AI and economics (and had substantial bearing on Simon's contributions in both fields): how do humans deal with limited computational resources in making decisions? This question is key not only to prescriptive AI modelling (engineering), but also to any general (descriptive) theory of *in vivo* human decision-making, under the assumption of the deductive model described above.

To overcome the problem of the intractable explosion of possibilities, in both the descriptive and prescriptive arenas, Simon, Newell and others suggested the use of *heuristics*. It is interesting to examine a current common definition of this word, taken from Oxford Dictionaries Online (2015):

> **Heuristic:** (from the Greek for "find" or "discover"): 1. refers to experience-based techniques for problem solving, learning, and discovery.

This definition most likely reflects usage at the time of Simon and Newell: the word was associated with ideas like educated guesses and rules-of-thumb, as practiced in human decision-making. However, in the current definition, a second, computing-specific sub-definition exists:

> **Heuristic:** 1.1 Computing: Proceeding to a solution by trial and error or by rules that are only loosely defined.

This is a somewhat confusing definition, since rules in computers cannot be "loosely defined". This reflects how the distinction between mechanical computational procedures and human decision-making procedures can become blurred by language usage. This blurring of meaning is an effect of what some in AI have called a *wishful mnemonic* (McDermott, 1976); that is, the assigning to a computational variable or procedure the name of a natural intelligence phenomenon, with little regard to any real similarity between the procedure and the natural phenomena. The confusion in the definition reflects the danger of wishful mnemonics, in that it could be interpreted to say that loosely defined rules (like those in human heuristics) have been captured in computational heuristics. The reality in AI and computation practice is that the term heuristic has come to mean any procedure that helps where exhaustive search is impractical, *regardless* of whether that procedure has any connection to human reasoning processes. There are many parts of AI that suffer from the dangers of wishful mnemonics, and economic models of human reasoning must be carefully examined for similar problems.

The AI systems that descended from Simon and Newell's work were called *expert systems*, with implications that they embodied the way humans reasoned in a domain of their expertise. Note that this is likely a wishful mnemonic, as history has shown that capturing expert knowledge in this framework is a fraught process, as is discussed later in this section.

As an illustration of the operations of an early expert system, consider *MYCIN* (Buchanan, 1984). MYCIN was applied to the task of diagnosing blood infections. As a knowledge representation, MYCIN employed around 600 IF/THEN rules, which were programmed based on knowledge obtained from interviews with human experts. MYCIN in operation would query a human for the specific symptoms at hand, through a series of simple yes/no and textual questions, and incrementally complete the knowledge representation for the current diagnostic situation, drawing on the expert-derived rule base. An algorithm searched through the rule base for the appropriate next question, and ultimately for the concluding diagnosis and recommendation.

The answers to questions in such a process, as well as the rules themselves involves uncertainty. For MYCIN, this necessitated an addendum to a purely logical representation. To cope with uncertainty, MYCIN included *heuristic certainty factors* (in effect, numerical weights) on each rule as a part of the heuristic search

in the mechanical inference process, and as a basis for a final confidence in the result.

A typical MYCIN rule, paraphrasing from the recollections in Clancey (1997), and translating to a more human readable form, was:

> **IF** the patient has meningitis **AND** the meningitis is bacterial **AND** the patient has had surgery **AND** the surgery was neurosurgery **THEN** the patient has staphylococcus with certainty 400 **OR** streptococcus with certainty 200 **OR** the patient has e. coli with heuristic certainty factor 300

MYCIN was an impressive demonstration of AI, in that the diagnostic procedure was considered a task requiring a very sophisticated human expert. Despite early successes like MYCIN, there were extensive problems with expert systems that eventually led to dramatic events in the field. Hype, often associated with wishful mnemonics, was certainly related to this drastic setback. But one must also examine the technical difficulties that these sorts of reasoning models suffered. Two are most notable. The first is *brittleness*. It was generally observed that expert systems, when set to situations that were unexpected by their designers, broke down ungracefully, failing to provide even remotely adequate answers, even if the questions were only mildly out the system's intended purview. For instance, in the example of MYCIN, if a human medical condition included factors that effected symptoms, but were not related to bacterial infection, and more importantly not included in MYCIN's program by its designers, MYCIN could fail dramatically, and deliver the incorrect diagnosis or no diagnosis at all.

Publicly-funded for AI research suffered significant blows in the early 70's, when Sir James Lighthill (McCarthy, 1974) was asked by Parliament to evaluate the state of AI research in the UK, and reported an utter failure of the field to advance on its "grandiose objectives". In the USA, ARPA, the agency now known as DARPA, received a similar report from the American Study Group. Funding for AI research was dramatically cut. By the end of the 80s, commercial expert systems efforts also collapsed.

One might assume that to overcome brittleness, one only needed to add more rules to the knowledge base, by further knowledge extraction from human experts. This leads to a second characteristic problem in the approach: *the explosive*

difficulty of model refinement. In general it proved to be simply too costly and difficult for practicality in many real world domains to refine a model sufficiently to overcome brittleness. While human experts don't fail at reasoning tasks in a brittle fashion, those experts found it difficult to articulate their capabilities in terms that fit the idealisation of these formal reasoning systems. Formulating a sufficiently vast and "atomic" knowledge representation for the general theory expounded by Simon-framework AI has proven an on going challenge in AI.

This is not to say that expert systems have ceased to exist. In fact, today there are probably thousands of programs, at work in our day-to-day lives, with elements that fit the "domain knowledge separate from inference engine" paradigm of expert systems. However, few of these would actually be referred to as expert systems now. In fact, the wishful mnemonic "expert system" fell out of favour after the 70s, to be replaced by terms like "decision support system", a name that implies the final decision-making rests with a less brittle, human intelligence. Outside the domain of decision support, there have been human-competitive deductive AI programs, notably Deep Blue for chess playing (Campbell et al., 2002), but characteristically, such programs focus on narrow domains of application, and apply levels of brute force computation that are not reflective of human decision-making processes. Another more recent example is AlphaGo (Silver et al., 2016) (from Google's DeepMind group), which has beaten human masters in the game of Go. Similarly to Deep Blue, AlphaGo, unlike human players, employs massive look-ahead search, but unlike Deep Blue, its search is based on evaluations of game positions obtained through inductive AI techniques, like those discussed in the following section.

3.2 Inductive Models

Given the difficulty in extracting human expert knowledge to overcome brittleness in real-world decision making scenarios, one would naturally turn to the idea of learning, so that knowledge can be acquired by observation and experience. This desire has led to a broad body of research. Some of this research follows the purely logic-based model that dominated deductive AI, notably *inductive logic programming* (Mitchell, 1997), which uses a logic-based representation (like that in deductive AI systems like MYCIN) of positive and negative examples, along with "background knowledge", to derive and test hypotheses (logical rules). This

paradigm has proven highly-useful in some fields (notably bioinformatics and some natural language processing), but suffers directly from the problems of the deductive schemes discussed above.

A larger body of inductive AI techniques sought to emulate models of the mechanics of the human brain at a low level, rather than attempting to model higher-level reasoning processes like expert systems. Such models are often called *connectionist* or *neural network* models. However, quite often these models have relied more on idealisations than on the realities of the human brain. Thus, inductive AI has suffered from problems of wishful mnemonics. Notable in neural networks is the use of neuro-morphic terms like "synapse", which is often used to simply mean a parameter, adjusted by an algorithm (often labelled as a "heuristic" algorithm, but in a wishfully-mnemonic fashion), in a layered set of mathematical functions. In fact there is often very little similarity of the algorithm or the functions to any known realities of biological neurons, except in a very gross sense. In fact, most of the non-neural inductive AI techniques that exist (e.g., support vector machines, decision trees, etc. (Mitchell, 1997) (Hastie et al., 2009)), which are built for entirely prescriptive, engineering goals, are algorithmically more similar to most neural network algorithms than those connectionist algorithms are similar to demonstrable brain mechanics. While there are neural networks and other inductive AI research programmes that have focused explicitly on emulating brain mechanics with some accuracy, the majority of inductive AI (machine learning) algorithms in use are not in this category.

Inductive AI of this sort has experienced similar cycles of hype and failure to that of expert systems (Minsky and Papert, 1988)(Rosenblatt, 1957). Like expert systems, most inductive AI techniques remain limited by the representations initially conceived of by their programmer, as well as the "training data" they are exposed to as a part of learning. "Poor generalisation", or the inability to deliver desired results for unforeseen circumstances is the primary failure mode of inductive AI. This is essentially the same difficulty experienced in deductive AI: an inability to robustly deal with the unforeseen.

While this observation is seldom made, inability to cope with unforeseen circumstances is at the heart of the difficulties (brittleness and poor generalisation) of both of the largest branches of AI research. This problem is directly coupled to the computational or practical intractability of constructing a complete model,

or searching the space of all possible models, in a real-world decision-making situation that might be faced by a human being. In idealised problems, these methods have proven very effective, but their lack of generality in the face of unforeseen circumstances is the primary and persistent challenge of the AI research program to date.

This is not to say that inductive AI programs have not had considerable successes, but like deductive AI systems, they are most often deployed as decision support, and they require careful engineering and great care in selection of data, features, pre and post processing to be effective. Like deductive AI, there have been human-competitive inductive AI developments like DeepQA (Ferrucci et al., 2010) (resulting in the program Watson, which succeeded in defeating human champions in the game show *Jeopardy!*). However, like Deep Blue for chess playing, Watson exploits significant engineering for a particular task, massive computational search power, as well as huge amounts of "big data", all of which are not reflective of human decision-making processes. In another example, DeepMind's AlphaGo performs induction over huge numbers of simulated games to derive numerical evaluations of Go board positions, and then performs massive computational search over the progression of those positions to derive its human-competitive performance. Once again, it is unclear if these procedures reflect any aspects of human players learning or performing in the game of Go.

Regardless of these domain-specific successes, there has been no general solution to the challenges of brittleness and poor generalisation. In real-world applications of AI, significant human effort is expended in designing the representation of any domain-specific problem such that relatively generalised algorithms can be used to deliver decision-support solutions to human decision makers. In inductive systems, significant time is spent designing pre and post processing steps, selecting and refining appropriate training data, and selecting the architecture of the learning model, such that good generalisations are obtained for a specific application. Even with this significant human effort, inductive results often generalise poorly to situations that are truly unforeseen, outside the narrow and prescribed realms of board games, well-specified optimisation problems, and the like. In broader contexts, automatically updating knowledge representations for deductive systems and selecting the right models for inductive systems (the so-called *structural learning problem*) remain massive challenges.

4 Probability as a Model of Uncertainty

Poor coping with unforeseen contingencies is tightly coupled to the idea of uncertainty about the future: if one had a certain view of the future, and the "atoms" of its representation, one would at least have a hope of searching for an appropriate or ideal representation, to overcome brittleness and poor generalisation.

Certainty (and thus uncertainty) is a phenomenon in the mind: humans are uncertain about the world around them. From a traditional AI perspective, we must conceive of a way to computationally model this uncertainty, and how humans decide in the face of it. This has naturally turned to probability theory in both deductive and inductive AI models of human reasoning with uncertainty. In the deductive frame, probabilistic networks of inference (Pearl, 1988) of various kinds have become a dominant knowledge representation, and in the inductive frame, statistical learning has become the dominant foundation for learning algorithms (Hastie et al., 2009).

"Probability" can itself be seen as a wishful mnemonic. The word probability is defined as "*The quality or state of being probable*", and probable is defined as "*Likely to happen or be the case*." However, in common discourse and in AI research, the term probability most certainly almost always means a number, within a particular frame of algorithmic calculations. This implies the number is a representation of the human concept of probable, regardless of whether this has any similarity to the real processes of human decision-making. To consider the idea that probabilities is wishfully mnemonic, it is necessary to examine what a probability is, from a technical perspective, and how that the fundamental assumptions of probability theory underpin and effect technical models of decision making.

A probability is defined as a number between zero and one that is assigned to some "event", the occurrence of which is uncertain. Probability theory explicitly assumes that one can enumerate (or in the continuous case, formally describe with functions) all possible, mutually exclusive "atomic" events, such that the sum of their probabilities equals one. These conditions are strictly required for a number to be rightly called a probability, such that the basic laws of probability theory (discussed below) apply. Those laws, often presented via arguments that arise from Venn diagrams over the enumerated event space, give: the probability of event A not occurring (Figure 1), the probability of one event A or an alternative (not

necessarily mutually-exclusive) event B occurring (Figure 2), and (perhaps most importantly) the probability of A occurring simultaneously with B (Figure 3).

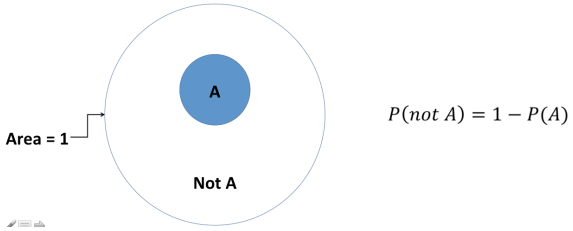


Figure 1: The Probability of event A not occuring.

In Figure 1 the white circle represents everything that could possibly happen (the enumerated space of atomic events), and thus its area (probability) is 1, by definition. The blue circle represents a subset of the atomic events that comprise event A, and the circle's area is A's probability. Thus, the probability of all events that comprise "Not A" is 1 minus the area of the blue circle.

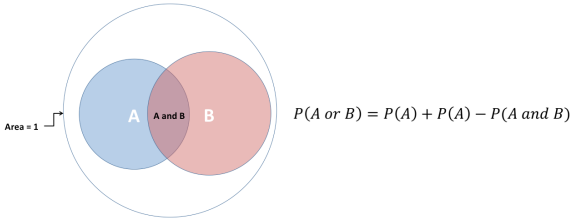
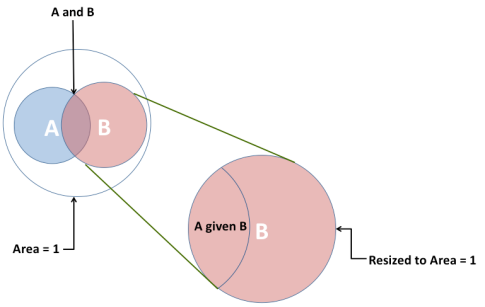


Figure 2: The joint probability of A and B

In Figure 2, the blue circle represents event A, the red circle represents event B and their overlap (the almond-shaped area) represents the probability of both A and B happening (the joint event, A and B). The probability of this joint event is given by the area of the blue circle plus the area of the red circle, minus the almond-shaped overlap, to avoid counting it twice, thus the formula.

Figure 3 shows perhaps one of the most critical relationships in probability theory, the relationship between joint probabilities and *conditional probabilities*. The interpretation of this diagram hinges upon the idea that the complete, enumer-



$$P(A\ given\ B) = \frac{P(A\ and\ B)}{P(B)}$$

Figure 3: The relationship of conditional (A given B) and joint (A and B) probabilities.

ated event space must always have a probability of one. Because of this, when we know that B has occurred (B is "given"), then the red circle is resized to area 1 (by definition), as it is then the enumerated universe of all possible events. This is represented by the re-sizing of the red circle that is shown in the figure. The scaling factor here is (the inverse of) the original size of the red circle. The almond-shaped portion (A and B) is assumed to be rescaled by the same factor. Thus, the probability of anything that was originally in the blue circle (the set of events that comprise A) happening given B has already happened (A given B) is the area of the almond-shaped area, scaled up by the inverse of the area of B.

4.1 Conditional Probabilities as Rules with Certainty Factors

The conditional probability of one event given another, $P(A \text{ given } B)$ is often written as $P(A|B)$. An important observation that connects probabilistic reasoning to non-probabilistic AI techniques of the past is that such conditional probabilities are simply rules, like those in MYCIN (recall that these rules included certainty factors). $P(A|B)$ is another way of saying

IF B **THEN** A with certainty factor $P(A|B)$

Let us assume that that the heuristic certainty factors in MYCIN are replaced by more formal probabilities. For the sake of elucidation of this connection, imagine that event B is the event that the patient has bacterial meningitis, and event A is the event that the patient has e. coli, paraphrased

from the previous MYCIN rule. This means that another way of writing $P(\text{the patient has e. coli given the patient has bacterial meningitis})$ is:

> **IF** the patient has bacterial meningitis **THEN** the patient has e. coli with certainty factor =
>
> $P(\text{the patient has e. coli given the patient has bacterial meningitis}) = P(EC|BM)$

where the final factor is a probability, adhering to the restrictions and laws of probability theory.

In much of modern deductive AI, conditional probabilities are at the core of probabilistic inference systems. An example of such a system is a Bayesian Belief Network (BBN) (Pearl, 1988). Such a network represents knowledge as an interconnected set of the conditional probabilities (which we can now see as rules with certainty factors) between many events. Deductive inference occurs when a set of (probabilistic) facts is introduced to the network (much like the answers provided by physicians to MYCIN), and an algorithm computes the probabilities of concluding events. The networks can also employ Bayes' rule to create inference in many directions through the network, reasoning from any set of known probabilities to conclusions with probabilities anywhere in the network. This addition is discussed in a later section.

BBNs employ the precisely prescribed mathematics of probability theory, resulting in conclusions that are mathematically valid within that theory. Thus, to some extent probabilistic inference systems like BBNs overcome the brittleness problem of Simon-framework AI systems by representing uncertain circumstances in "the correct way" (that is to say, in a way that is consistent with probability theory). But we must question whether this is really a superior method of representing uncertainty. Is it a heuristic, in the sense that it models a human reasoning phenomena, or merely a heuristic in the wishfully mnemonic sense: a computational convenience labelled as if it were a human reasoning process, while bearing little similarity?

5 Wishful Arguments for Probabilistic Models

5.1 Bayes Rule and Subjective Probabilities

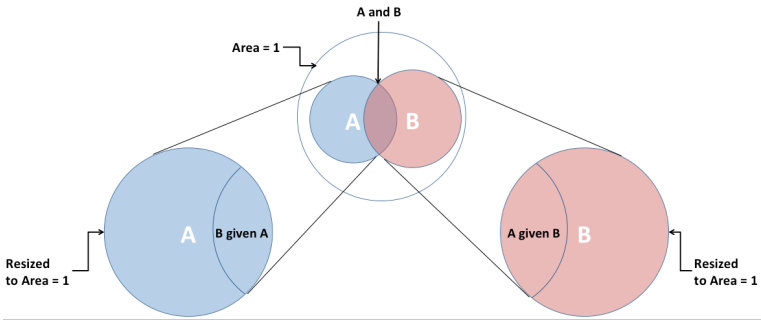
It is frequently assumed that even in cases where objective probabilities can't be explicitly derived, subjective probabilities can be assumed to be in the mind of human decision makers (Savage, 1954). This directly relates to the oft-debated frequentist and subjectivist views of probability. In the frequentist view, probabilities gain meaning in reference to future frequencies of event occurrences. In the subjectivist view, probabilities are only subjective degrees of certainty in the possible occurrence of future events: that is, they are in the mind of the decision maker.

The frequentist view is the most strongly tied to the assumptions of probability theory: the full enumeration of event spaces, the summing of their probabilities to one, and repeatable conditions that allow for the confidence built by the *law of large numbers*. This mathematical law states that if one repeats the same probabilistic experiment many times, the frequency of an event will converge towards the event's probability. Note that in assuming a probabilistic experiment, we are implicitly assuming the construction of a representation of a completely enumerated event space. The completeness of that enumeration directly implies the precise repeatability of the experiment.

It is arguable whether these careful conditions hold for many real-life situations in human decision-making experience. One instance may be well-constructed casino games, and the like. Another is problems that deal with large populations of entities, for instance in particle physics. Another population-based example is an epidemiological study, where a well-constructed set of parameters over that population (a careful representation) has been constructed, and all other differences between the subjects are explicitly to be ignored or, in effect, "averaged out". Outside these well-constructed examples, it is doubtful that a fully enumerated event space is actually ever captured by a knowledge representation, in much the same way that it has proved extremely difficult to construct non-brittle knowledge representations in AI. AI history shows that in constructing knowledge representations, finding a set of rules (which are analogous to conditional probabilities) that

cover that space of possible events has proved to be critical to robust (non-brittle) inference. Finding such a set has proven to be an intractable problem in AI.

The subjectivist view of probability has the advantage of not requiring validation through repeatability of possible event conditions into the future to confirm frequencies of occurrence: the probabilities are assumed to be subjective numbers in the mind. However, the subjectivist view has become tightly coupled to the Bayesian view of probabilities, named for Thomas Bayes, from work of his presented posthumously in 1762 (Bayes and Price, 1763). The key idea is that of Bayes' rule, which follows simply from the basic rules of probability that were previously stated, by considering both the probability of A given that B has occurred, and the probability that B given that A has occurred.



$$P(A \text{ and } B) = P(A \text{ given } B)P(B)$$
$$= P(B \text{ and } A) = P(B \text{ given } A)P(A)$$

$$P(A \text{ given } B)P(B) = P(B \text{ given } A)P(A)$$

$$P(A \text{ given } B) = \frac{P(B \text{ given } A)P(A)}{P(B)}$$

Figure 4: The derivation of Bayes' rule

Recall the "scale up" argument from Figure 3, and consider Figure 4. This shows that Bayes' rule derives from a simple re-scaling based on the ratio of sizes of enumerated event spaces and sub-spaces. The almond-shaped area in all three parts of this figure is the same set of atomic events, just in three different, fully enumerated event spaces: the white one before either A or B happens, the blue one

where A has already happened (and the blue circle is scaled up to area equals 1) and the red one where B has already happened (and the red circle is scaled up to area equals 1). Since the almond is the set of events, regardless of which scale-up was performed, one can directly derive Bayes' rule, using the ratio of the two original probabilities (scale-up factors), as shown in the figure's equations.

Note that the calculations performed draw directly on the laws of probability theory, with its requirements of fully enumerate event spaces, whose probabilities sum to one, and the implication of repeatable experiments giving the implied confidence of the law of large numbers. While Bayes' rule is closely tied to subjectivist interpretations of probability theory, its mechanical calculations (multiplying by the precise ratios of circle sizes in the Venn diagrams) derive explicitly from the frequentist foundations of probability theory. However, this may be only a detail, in that Bayes' rule has a conceptual, subjectivist interpretation that is revealing.

It is instructive to examine the terms of Bayes' rule relative to the MYCIN rule analogy. Bayes' rule states that if one has the rule 1, possibly provided by an expert:

> **IF** the patient has bacterial meningitis **THEN** the patient has e. coli with certainty factor $P(EC|BM)$

then one can infer rule 2:

> **IF** the patient has e. coli **THEN** the patient has bacterial meningitis with certainty factor =
> $P(BM|EC) = \frac{P(EC|BM)P(BM)}{P(EC)}$

In subjectivist probability, the terms in Bayes' rule have particular names that help to clarify their conceptual meaning:

- $P(EC)$ and $P(BM)$ are called *priors*, representing the pre-existing probabilities that the patient has each of those diseases, unconditionally, that is to say, conditioned on no other events. In the subjectivist sense of probabilities, priors are often called *prior beliefs*, however, this term must be carefully examined for its role as a wishful mnemonic in this context.

- $P(BM|EC)$ is called the *posterior*, representing the probability that the patient has bacterial meningitis, given the existence of rule 1 above, and the appropriate priors in the Bayesian calculation.

In terms of the sorts of inference used in rule-based AI systems, and the analogy between conditional probabilities and rules, Bayes' rule shows that that one can infer the certainty factors of a new rule that has the conditions and actions of any existing rule "reversed", through adjustment the known certainty factors of the existing rule.

Let's assume that the (unconditional) probability that a patient has bacterial meningitis is much stronger than the (unconditional) probability that the patient has e. coli, and we have an (expert or statistically) established rule 1. In such a case, the Bayes' rule states that one should increase the certainty of the rule 2 relative to that of rule 1, since the ratio of the priors is greater than one. Conceptually this is consistent with the intuitive idea that under these assumptions, most people with e. coli probably also have bacterial meningitis, because that disease is commonplace, and the certainty of rule 1 is strong.

Let's assume the opposite extreme relationship of the priors: that the (unconditional) probability that the patient has bacterial meningitis is much less strong than the (unconditional) probability that the patient has e. coli. In this case, one should lower the certainty of the reversed rule (rule 2), relative to rule 1, based on the ratio of the priors, which is less than one. This is conceptually consistent with that intuitive idea that one cannot have much certainty in a patient having meningitis just because they have the commonplace e. coli, regardless of the certainty of rule 1.

Conceptually, these extreme examples reduce Bayes' rule to an explainable, intuitive candidate as an *in vivo* human decision-making process, a heuristic in the true sense of that word, rather than a wishful mnemonic. However, consider a case where all the certainties involved, of rule 1, and the two (unconditional) priors, are of middling values. In this case, the exact certainty value derived for rule 2 (also a middling value) lends little of the intuitive confidence of the extreme cases. This is particularly true if one assumes the likelihood that the knowledge representation (set of rules and their certainty factors) may be in a model (rule set) that may be incomplete, as has been the case in the brittle AI models of the

past. This illustrates that while the conceptual, subjectivist interpretation of Bayes' rule as a metaphor for human reasoning stands up to intuitive scrutiny, the exact mechanisms are doubtful, unless one assumes the real-world can be captured in an adequate frequentist-probability-based knowledge representation. For problems of real-world scale, there may be no basis to assume this can be done any better for probabilistic models that it has been done in deterministic AI models of the past.

As noted earlier, the cases conforming to the assumptions of probability theory are limited to a few situations like casino games, population studies, and particle physics. It is unclear whether one should expect the human brain to have evolved Bayes' rule in response to these situations, as they aren't a part of the common survival experience during the majority of man's evolution. Evolution of Bayes' rule at neural or higher cognitive level may be a doubtful evolutionary proposition for this reason. In relation to this, it is important to note that while behaviour that can be described by Bayes' rule under a particular description of events may "match" behaviour of living subjects in certain experimental settings, this does not directly imply the existence of Bayes' rule in cognitive or neural processes. The intuitive "extreme cases" previously discussed could certainly be implemented by more crude calculations (heuristics).

As an analogy to logical and mathematical representations that do not include uncertainty, consider the example of catching a ball given by Gigerenzer and Brighton (2009). We are able to catch balls, which follow a parabolic motion path due to the laws of classical physics. However, just because of this human capacity, one should not expect (as Dawkins (1976) once suggested) to find the solutions to the equations of motion of the ball in the brain. Instead, one should only expect to find heuristics (where the term is used in the early, human sense of the word, rather than the more computational sense, in this instance) that allow the human to catch the ball. In fact, Gigerenzer and Brighton show such heuristics at play as real human decision-making processes.

The implication is that, even in cases where the exact calculation of a set of interdependent conditional probabilities yield optimality in the real-world, one should not expect the brain to have explicitly developed Bayes' rule calculations. Instead, the brain will have implemented heuristics that deliver the outcome that is necessary, without explicit modelling of the equations per se.

5.2 Information Theory

An important outgrowth of probability theory that has direct bearing on many AI systems, and on perceptions of uncertainty in other fields that study human decision-making, is Information Theory (Shannon, 1948)(Gleick, 2011). It is important to consider information theory in reflecting on the value of probability theory in describing human decision making under uncertainty, as the apparent relationship of this theory to fundamentals of communications and physics tends to give it added credence as a universal aspect of reasoning about uncertainty.

Shannon introduced information theory as a solution to particular problems in telephonic communications and data compression. The central mathematical expression of information theory is the following:

$$I(x) = -\log P(X) \tag{1}$$

where $P(x)$ is the probability of event x. Given Shannon's context of telephonic communications applications, the event x is typically interpreted as the arrival of some signal value at the end of a noisy communications line. $I(x)$ is often referred to as the *information content* of the symbol (or event) x. The logarithm is usually taken in base 2, giving that $I(x)$ is expressed in bits. Since information content is based on the idea of the symbol x arriving at random according to some probability, and it increases as that probability drops, it is also often called the amount of surprise. The word "surprise" here must be examined as a candidate wishful mnemonic, but even the term "information" in information theory is itself a metaphor, as Shannon himself noted:

> "Information here, although related to the everyday meaning of the word, should not be confused with it."

Certainly from this statement, "information" in "information theory" is a primary candidate to become a wishful mnemonic. This is illustrated by how, in spite of Shannon's caution on the meaning of the words he used for the symbols and calculations he employed, words with potentially different meanings in human thinking and engineering discipline become entrenched in the basics of the theory's explication:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

Clearly this core statement of Shannon's theory sounds strange if one treats the word communication in its "everyday meaning", as opposed to in the sense of engineered telephony. Communication in the human sense is tied to subtleties of meaning for which we would have exactly the same problems of explication found in trying to articulate expert systems, only more so. However, Shannon necessarily goes on to state:

> "Frequently the messages have meaning. That is they refer to or are correlated according to some system with certain physical or conceptual entities. The semantic elements of communication are irrelevant to the engineering problem."

Shannon has been very careful here to exclude meaning from what he means by information, and to point out that what he means by information is not what that word means in everyday use. However, this attempt at avoiding wishful mnemonics caused discomfort from the introduction of the theory. Commenting on one of the first presentations of information theory at the 1950 Conference on Cybernetics, Heinz Von Foerster said:

> "I wanted to call the whole of what they called information theory *signal theory*, because information was not yet there. There were 'beep beeps' but that was all, no information. The moment one transforms that set of signals into other signals our brains can make an understanding of, then information is born—it's not in the beeps."

Information theory is at its core based on a probabilistic interpretation of uncertainty, which is at its core based on the enumeration of event (or message) spaces. It is only if one accepts that interpretation as a fundamental aspect of the objects of uncertainty that these concerns over meaning dissolve, along with Shannon's concern over the name of his theory being treated as a wishful mnemonic.

5.3 Physics Analogies and Bayesian Reasoning Models

A key concept in information theory, which yields its implied relationship to physics, is the average value (or expectation) of the information content of events (symbols arriving at the end of a communication line):

$$H(x) = -\sum_{x} P(x) \log_2 P(x) \tag{2}$$

This symbol $H(x)$ is most often called information entropy, expressing an analogy to physics (specifically thermodynamics), where entropy is a measure of disorder, particularly the number of specific ways that a set of particles may be arranged in space. A simple example, provided by Feynman (2000), suffices to explain the metaphor. Consider situation A to be a volume containing a single molecule of gas, and assume that the volume is large enough that there are X equally likely locations where the particle might be. This is a uniform distribution over these X locations, and the number of bits necessary to represent all possible locations is given by

$$H_A(x) = -\sum_{x} P(x) \log_2 P(x) = -x\left(\frac{1}{x}\right)\log_2\left(\frac{1}{x}\right) = -\log_2\left(\frac{1}{x}\right) \tag{3}$$

Thus, any message indicating the location of the particle has this information content. Although the expected information content in this uniform distribution case is trivial, one can imagine this average over any distribution.

Now imagine performing work to compress the gas, and reduce the volume to half its original size (situation B). Now half the terms in the sum become zero, the probabilities in the remaining terms double, and the number of bits necessary to represent the location of the particle is given by:

$$H_B(x) = -\log_2\left(\frac{1}{2x}\right) = -\log_2\left(\frac{1}{x}\right) - 1 = H_A(X) - 1 \tag{4}$$

In effect the compression of the volume has reduced the information entropy (expected information content) of messages indicating the location of the particle by one bit.

The close analogy to physics in information theory complicates the issue of probability theory as wishfully mnemonic, in that at the lowest levels of physics,

we have uncertainties that apparently do conform to the assumptions of enumerable event spaces and repeatable experiments from (frequentist) probability theory. If the probabilistic and information theoretic interpretation of uncertainty is fundamental to reality itself, then perhaps it is a likely basis for the evolved nature of thought. Such thinking is consistent with the *Bayesian Brain* approach (Doya et al., 2007), popular in many current cognitive and neural science studies. Bayesian Brain research, like traditional AI, has manifestations in modelling both higher-level thought and neural models.

Like information theory, this research has borrowed physics concepts from physics, notable *free energy*, (Friston, 2010) which is a term used in variational Bayesian methods of modelling decision-making in (prescriptive) machine learning and (descriptive) neural models. Free energy calculations, like entropy calculations, are similar in physics and models of symbols related to communication and thought. The *Free Energy Principle* has been suggested as an explanation of embodied perception in neuroscience. Convincing results exist where technical systems derived from this theory illustrate behaviour that is similar to analogous biological systems. The success of the free energy principle in neural modelling level has led to its discussion at psychological (Carhart-Harris and Friston, 2010) and even philosophical (Hopkins, 2012) levels, as a part of higher-level thought. This analogy has historical foundations. It is unsurprising that Freud himself used the term "free energy" in elucidating concepts in his theories of psycho-dynamics, in that Freud was greatly influenced by thermodynamics in general, and Helmholtz, the physicist (who was also a psychologist and physician) who originated the term "free energy" in thermodynamics.

In light of the impact of wishful mnemonics in the history of AI, one must consider carefully tying exact algorithmic procedures to what could have only been a metaphor in Freud's time. One must carefully consider whether the exact mathematical mechanisms involved in these algorithmic representations are justified as an expected outcome of the evolution of neural structure, higher-level thought, or both, if the probabilistic view of uncertainty differs from real-world experiences of uncertainty in human (and perhaps other biological) cognition. Despite this, many feel that the inclusion of that Bayesianism is key to reforming the old AI models of human thought:

> "Bayesian Rationality argues that rationality is defined instead by the ability to reason about uncertainty. Although people are typically poor at numerical reasoning about probability, human thought is sensitive to subtle patterns of qualitative Bayesian, probabilistic reasoning.
>
> We thus argue that human rationality, and the coherence of human thought, is defined not by logic, but by probability."
>
> - Oaksford and Chater (2009)

With regard to "human rationality, and the coherence of human thought", one must examine this statement via evidence as to whether humans reason in a way that is consistent with probability theory. In fact, there are a plethora of examples (many from behavioural economics (Kahneman, 2011)) where humans decide in a way that is inconsistent with probability theory. There is a substantial literature explaining these deviations from the probabilistically-optimal solution via various biases that are intended to explain "heuristics" (where that term may or may not be wishfully mnemonic) that lead to the observed behaviour (for instance (Kahneman and Tversky, 1979)). Such theories will often introduce new variables (often weights on probabilities) to explain how the human behaviour observed is optimal (according to a probabilistic interpretation) in the new, weighted problem. The controversy amongst economists over whether such representations are realistic models of human decision-making behaviour stretches back to very early probabilistic thought experiments, where human behaviour is obviously inconsistent with probability theory (Hayden and Platt, 2009).

While introducing and tuning weights can in every instance create a probabilistic model in which a particular, isolated, observed human decision-making behaviour is optimal, it is important to note that this does not mean that the human heuristics are in fact implementing a solution procedure for the weighted problem. One might conjecture that this does not matter, because equivalence of behaviour is sufficient. However, AI history suggests that the approach of overcoming the brittleness of an idealised representation of a problem through the incremental addition of variables (rules and their weights) is fraught, and potentially computationally explosive. This situation is only potentially aggravated by multi-prior models (so-called Ambiguity Models) that attempt to characterise unknown probabilities

with probabilities of possible values of probabilities (Conte and Hey, 2013). This evolution of ideas suggests that the probabilistic approach to modelling human decision-making may suffer similar consequences to brittle models in AI, if there is a need to perpetually find and tune new bias variables to account for differences between models and observed behaviours, in order to continue adherence to idealised inference procedures.

It is reasonable to assume that at some (neural) level human thinking is dominated by the same laws of physics as any other system, and just as explicable by equations of some sort. But the assumption that this mechanism must be of a logical, probabilistic, or Bayesian formulation at all levels of human decision-making must be examined based on observation of real human decision-making. When inconsistencies with assumed formalisms are detected, incremental modification of those formalisms may be both methodologically incorrect, and ultimately practically intractable.

5.4 Statistical Learning and Big Data

In inductive AI there are recent developments that parallel those of probabilistic, deductive AI mentioned above, notably the advance of statistical learning techniques (Hastie et al., 2009). Statistical learning is in many ways similar to traditional, connectionist AI models: it uses formal algorithms to examine the statistical information in exemplars (training data), and update parameters (either probabilities, or parameters of functions that result in probabilities) of a low-level knowledge representation. Often this representation is in the form of interconnected mathematical functions that are not dissimilar to connectionist AI algorithms.

In statistical learning algorithms, the features of exemplars are inputs, and some set of their statistical properties are the corresponding "correct" outputs. The statistical learning algorithm "trains" the mathematical functions of the underlying knowledge representation such that unforeseen inputs generate outputs that have the same statistical characteristics as the exemplars, in line with probability theory.

Note that this does not overcome the fundamental problem of selecting the right representation per se: that is, the correct "atomic" features, and a sufficient set of statistical properties that represent the desired generalisation. The problem of knowing beforehand what might be the right generalisation for unforeseen cases

remains a difficulty. In fact, the structural learning problem in Bayesian networks has been shown to be amongst the most intractable of computational problems (Chickering, 1996).

However, great advances have been made in the development of computationally effective and efficient probabilistic network algorithms and statistical learning algorithms. There is the additional excitement for such approaches that comes from the enormous availability of data provided via the Internet. This so-called "big data" seems to provide a massive opportunity for large-scale inductive AI.

However, the existence of big data not only does not overcome the problem of appropriate feature and architecture selection in inductive AI, it may in fact complicate the problem, since, from a technical perspective, the space of possible models explodes with the size of the data, and the space of "spurious correlations" that can be found in the data (that is, statistical conclusions that are valid for the training exemplars, but seemingly nonsensical when examined objectively) explodes with the amount of data available (Smith and Ebrahim, 2002).

In essence, finding meaningful relationships in big data is prone to finding incorrect generalisations, without prior assumptions of a model of the data itself. As has been discussed, finding the correct model (as a problem within AI) remains a primary challenge. Models or feature sets in AI and big-data-learning algorithms are usually the work of intelligent human designers, relative to a particular purpose. Moreover, any such model leaves space for unforeseen possibilities that are yet to be uncovered. Also note that much of big data is human-generated, reflecting the thoughts and decision-making processes of humans. These are precisely the processes that have proven difficult to model in AI.

6 Alternate Representations of Uncertainty

The difficulties that have been discussed result in some measure from the fact that approaches to decision-making under uncertainty in AI, as in economics, start from sets of idealised views of decision-making (such as principles of logical deduction or probability theory) rather than from observations of how human agents actually manage in the uncertain decision-making situations they face. It is useful to begin by characterising the nature of uncertainty without reference to a

particular ideal reasoning process, like that implied in probabilistic, information theoretic, and Bayesian approaches. Lane and Maxfield (Lane and Maxfield, 2005) divide human uncertainty into three types: *truth uncertainty, semantic uncertainty*. and *ontological uncertainty*.

Consider the use of an at-home pregnancy test. If the window in the test shows a "+", one can look at the instructions, and see that this means a particular probability of being pregnant (a true positive), and another probability of not being pregnant (a false positive). Likewise, there are similar true negative and false negative probabilities in the case that the test shows a "–". These probabilities are derived from the statistics of a large population of women on whom the pregnancy test was evaluated before it was marketed. They represent *truth uncertainty*: the uncertainty of the truth of a well-founded proposition: in this case, the proposition is that the woman is pregnant. Truth uncertainty is the only type of uncertainty that is well treated by (frequentist) probabilities.

Now consider the case where the instructions for the pregnancy test are lost, and the user does not remember if the "+" in the window of the test indicates that the user is pregnant, or the opposite. In this case, the uncertainty is *semantic uncertainty*, uncertainty in the meaning of known symbols. The uncertainty in this case exists in the user's mind (possibly as subjectivist probabilities, but possibly not), and there are no meaningful statistics as to the user's confusion. Moreover, if one were to attempt to conduct experiments with similar users, it is unclear how meaningful the resulting statistics would be, given that any particular user's forgetting is the subject of a large number of features of that user's thinking, which, as we have discussed, are difficult to capture in a formal model that is not brittle. The selection of the appropriate features and architecture come into play, making the generalisation obtained somewhat pre-conceived by the designer of the test, its methods, and conclusions.

Finally, consider the case where a "*" turns up in the window of the test. This is a completely unforeseen symbol, from the user's perspective. The unforeseen possibility of symbols other than the expected "+" or "–" is a case of *ontological uncertainty*: uncertainty about the objects that exist in the universe of concern. In the case of the pregnancy test, this leaves the user in a state that requires search for a completely new explanatory model for the pregnancy test itself, and it's behaviour. It is the cause for doubt of previous assumptions and the need for deeper, more

innovative investigation. In AI, ontological uncertainty is precisely what leads to brittleness and poor generalisations in descriptive AI systems: a failure of the programmer to foresee future situations or the generalisations required for them. In effect, ontological uncertainty is precisely uncertainty about the atoms of the representation, or their relationships, and it is the precise source of the major failure mode in AI history. In the context of economics, this is analogous to the on-going possibility of emergent innovations, a topic that is not extensively treated in mainstream economic thinking, but that some have observed as a key aspect of economics itself (Beinhocker, 2007).

Examining AI as a descriptive science, it is clear that humans cope robustly with ontological uncertainty on a daily basis: the objects that exist in our universe and the connections between them are constantly unforeseen, but humans do not, in general, experience brittle failures to decide and act in the face of these novel circumstances. Even when viewed as prescriptive science, the failure modes of AI can be seen precisely as failures to cope with ontological uncertainty: failures to enumerate the proper atoms and their connections in a deductive model, or failure to create appropriate feature representations and connections for an unforeseen generalisation in an inductive model. These conditions are not made tractable by probabilistic models of uncertainty, precisely because probability theory does not treat ontological uncertainty. Clearly, the investigation of how human decision-making processes manage to cope with the full and realistic range of uncertainties is an important area of investigation for the AI project, and for economics. Some economic fields touch upon ontological uncertainties (notable Evolutionary Economics (Friedman, 1998) and the Economics of Innovation (Cecere, 2015)), however, few treatments consider the real-world psychology in the mind of the economic agent in the face of an ontologically uncertain world, instead focusing largely on the emergent complexities of interactions between simplified agents with no real psychological model.

6.1 Discussion

Ontological uncertainty, and its ubiquity, preclude logical or probabilistic ideals that suggest there is a way that humans *should* think in reality, in that there is no set of atoms for an ideal logic that can be pre-described, and no optimal

solution in probabilistic expectation over a pre-described event space. Models of human decision-making that have started with ideals of logic or probability theory (and problem domains derived from those ideals) start explicitly by eliminating the possibility of ontological certainty. This may be the reason that formal AI systems, which start with those ideals, fail with brittleness, poor generalisation, and difficulties in model scaling or the search for models. If we choose to examine human decision-making outside those ideals, we cannot begin by modelling human decision-making behaviour in an idealised abstraction, so we must examine that behaviour as it exists, without such idealised abstractions. This leads to the idea of examining human decision-making *in vivo*, as it exists, rather than through the lens of idealised representations and problems.

While much is known about *in vivo* human decision-making in the fields of psychology, sociology, anthropology, etc. surprisingly little of this work has been considered in AI, or in economic models of decision-making. In the case of AI, this may be because of the prescriptive agenda of most modern AI: it has been seen as valuable to create AI systems as tools that help overcome the computational speed of capacity limitations of human decision makers, particularly in the context of problems that are assumed to be logically or probabilistically solvable (and where a human expert is ultimately available to override this "decision support system"). However, in the descriptive framework of economics, this justification of prescriptive efficacy does not apply. Thus, particularly for the economic modelling of human decision-making, it is important to consider what can be learned from fields of research that specifically examine the realities and contexts of human decision-making. Moreover, such consideration may aid in the construction of more useful and informative AI systems.

Most fundamentally, human decision-making takes place in the physical context of the human itself. While the limitations of the brain are one physiological structure that must be examined (as was done in Simon's pivotal work, yielding the ideas of satisficing and heuristics), the human body physiology has been shown to play a critical role in the way that humans think, as well. Philosophy of the mind was long-bothered by the problem of *symbol grounding*: the problem of how words and other symbols get their meanings. Like early AI, this problem proceeds from the idea of the symbols themselves as idealisations. It is essentially the problem of how the (grounded) sensory experiences humans have are translated into the

abstract symbols humans use as representations in our reasoning. Note that this assumes that we do in fact use such symbols in reasoning. Recent cognitive science theories suggest that this may not be the case in much of human thinking: that humans actually take in their sensory experiences and largely reason about them in their modality-specific forms, and these are the most basic building blocks of thoughts. This idea is the foundation of the field of embodied cognition (Lakoff and Johnson, 1999), which proceeds from the idea that thought's atoms are sensations and metaphors to sensory experiences that are inherently tied to our bodies, rather than abstract symbols. If this insight is correct, and a significant portion of human decision-making is made through reasoning about embodied sensations, rather than abstract symbols, this should be more carefully examined in modelling human decision-making behaviour.

This leads directly to the fact that human decision-making, particularly under uncertainty, takes in a psychological context. In this area, emotions, much overlooked and disregarded in most idealisations of thought in AI and economics, must be considered not only as a reality, but as a possibly-useful and evolved feature of human decision-making processes. Consider *conviction narrative theory* (CNT) as a description of how humans cope with decision-making uncertainty (Tuckett et al., 2014)(Tuckett, 2011). CNT asserts that to cope with decision-making uncertainty, human actors construct narrative representations of what will happen in the future, as if the objects and relationships involved were certain. To continue to reason and act within these representations of the future (for instance, to continue to hold a financial asset), despite decision-making uncertainty, the human actor emotionally invests in the narrative, particularly to balance approach and avoidance emotions, such that action can be sustained. In this sense, emotion plays a role in the human decision-making process by providing a foundation upon which conviction can lead to action, despite uncertainty. Tuckett's interviews with hedge fund managers across the recent financial crisis show consistency with CNT (Tuckett, 2011).

CNT results also are consistent with the fact that human decision-making takes place in a social context. Examinations of conviction narratives in financial news (which are drawn from a large social context) are providing powerful causal results about the macro economy (Tuckett et al., 2014). These results are consistent with CNT combined with observations by Bentley, O'Brien, and Ormerod (Bentley et al., 2011), who note that in making economic decisions, humans can socially

rely on experts, or on a peer group. Thus, social sharing of ideas is an important aspect of understanding human decision-making. However, like in AI systems, it is of limited value to only consider a bounded set of atomic ideas being socially shared in the human decision-making process. We must examine how the social process innovates ideas as well. Thus, the fundamental mechanisms of how ideas are shared are of interest. One must also consider the communication mechanisms, comparison, and evaluation of differing representations that are acquired socially. Note that communication in this sense must reclaim that word from its use in information theory, where, as discussed above, it is distinctly disconnected from meaning. Considering the realities of the *in vivo* social communication of ideas may lead to understanding of the juxtaposition of representations from those shared ideas, which may provide a basis for understanding the innovation of ideas themselves. For instance, major scientific innovations can be seen as the juxtaposition of ideas that were shared socially (Koestler, 1963), and there is a small body of AI research that algorithmically and mathematically considers juxtaposition as a source of innovation (Goldberg, 2002). Evolutionary economics (Friedman, 1998) has recognised relationships to this area of AI evolutionary computation), but has yet to probe it with similar mathematical rigour regarding juxtaposition. Moreover, even work within AI has yet to develop significant theory of the development of innovations in open (social) systems. New ideas in complex systems science that directly consider the behaviour of open systems may need to be considered in this light (Kauffman, 2000). An important element of re-examining the idea of human decision-making in the face of an ontologically uncertain world may be extending these areas of work to consider the mechanisms underlying the innovation of human ideas in both the psychological and open, evolving social contexts.

7 Conclusion

This paper has reviewed basic structure of past AI models, revealing that both deductive and inductive AI has failed to robustly cope with the unforeseen, and that the limitations of AI are often masked by wishful mnemonics. More recent probabilistic models have been shown to be structurally similar to past models, and

to suffer from similar problems. This brings up fundamental questions about the future of modelling human decision making under uncertainty.

> "Essentially, all models are wrong, but some are useful." – Box and Draper (1986)

G. E. P. Box's famous quote seems particular apropos when one considers the history of modelling human decision-making in AI, and the realities of ontological uncertainty. All models are indeed wrong, and some useful, for a time. This holds not only for logical, but for probabilistic models (Box was, after all, discussing statistical models). The strength of human decision-making seems to be the ability to continually develop quick, incomplete, disposable, innovative, useful, new models, and derive the conviction to take action on their conclusions, despite the fact that they are always wrong in the face of real uncertainty. This is a fundamental strength that remains un-replicated in AI systems.

Human engineers create AI systems, and often it is their insight and experience that overcome the limitations of AI, in selecting models that are adequate for a given purpose. Such systems (cast not as "expert systems" but as "decision support systems," where the final decision rests with more robust human decision makers) serve a valuable purpose. The history of AI illustrates that the human as final decision maker is necessary to overcome the brittleness of AI systems. Current, probabilistically-based AI systems do not fundamentally overcome this necessity, because of the reality of ontological uncertainty. These factors must be carefully considered when one considers the the future of employment in light of computerisation (Frey and Osborne, 2016b) (Frey and Osborne, 2016a). It may be that many jobs will be computerised, only to result in a serious qualitative reduction in the robustness of the work generated, which may have longer-term economic impacts, as yet unaccounted for. Machine-based categorisations of jobs based on a current model (which could be brittle) might overlook the subtle involvement of human robustness in many types of employment.

Moreover, in considering descriptive modelling of human decision-making, particularly in economics, we clearly cannot overlook the limitations of AI systems. AI has shown that models of human reasoning that are based on the idealisations of mathematics or logic do not embody the real-world, robust decision-making in the

face of ontological uncertainty observed in humans. Thus, one must use caution when attempting to model the decision-making of economic actors using similar tools.

It is very important that those who descriptively model human decision-making, particularly in economics, are aware of these very fundamental limitations of AI, because they are directly applicable to their own models. As was initially noted, while economics has drawn the distinction between quantifiable uncertainties (Knightian risk) and un-quantifiable (Knightian) uncertainties (Knight, 1921), the commonplace assumption is that the later are made quantifiable within the mind, particularly as subjective probabilities. Recent enthusiasms for AI models that employ probability and statistics may seem to enforce the idea that human decision-making under uncertainty can be modelled effectively in just this fashion. However, the reality is that the modelling of human decision-making in AI is far from adequate *descriptively*.

If we are to overcome these *descriptive* limitations, we must think beyond idealisations of how humans *should* reason. An effective descriptive model must look at how humans *do* reason. The history of AI says that incremental refinements to otherwise idealised models (like the introductions of "biases") has been an ineffective strategy. It may adequately serve some *prescriptive* engineering goals, but it is an approach that seems inconsistent with observational, descriptive science. Likewise, we should not restrict our investigations of human decision-making domains that only include truth uncertainty, and thus are contrived such that there is a way that they *should* reason. We must consider how humans *do* reason under real-world uncertainty, where such idealisations do not apply. This suggests the re-inclusion of observations from sciences that have focused on *in vivo* human behaviour (psychology, sociology, anthropology, etc.). By focusing on this non-idealised decision-making, we may be able to innovate more realistic models of economic actors, as well as generate substantial advances in AI.

References

Bayes, T., and Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53: 370–418. URL http://www.jstor.org/stable/105741.

Beinhocker, E. (2007). *The origin of wealth: The radical remaking of economics and what it means for business and society*. Harvard Business Review Press.

Bentley, R., O'Brien, M., and Ormerod, P. (2011). Quality versus mere popularity: A conceptual map for understanding human behavior. *Mind and Society: Cognitive Studies in Economics and Social Sciences*, 10(2): 181–191. URL http://EconPapers.repec.org/RePEc:spr:minsoc:v:10:y:2011:i:2:p:181-191.

Box, G. E. P., and Draper, N. R. (1986). *Empirical model-building and response surface*. New York, NY, USA: John Wiley & Sons, Inc.

Buchanan, B. G. (1984). *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

Campbell, M., Hoane, A. J., Jr., and Hsu, F.-H. (2002). Deep Blue. *Artif. Intell.*, 134(1–2): 57–83. URL http://dx.doi.org/10.1016/S0004-3702(01)00129-1.

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas. *Brain*, 133: 1265–1283.

Cecere, G. (2015). The economics of innovation: A review article. *The Journal of Technology Transfer*, 40(2): 185–197. URL http://dx.doi.org/10.1007/s10961-013-9319-6.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In D. Fisher., and H-J.Lenz (Eds.), *Learning from Data: AI and Statistics*. Springer-Verlag.

Clancey, W. L. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge University Press.

Conte, A., and Hey, J. (2013). Assessing multiple prior models of behaviour under ambiguity. *Journal of Risk and Uncertainty*, 46(2): 113–132. URL http://dx.doi.org/10.1007/s11166-013-9164-x.

Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Computational neuroscience. Cambridge, Mass. MIT Press.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaefer, N., and Welty, C. A. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3): 59–79. URL http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303.

Feynman, R. P. (2000). *Lecture notes on computation*. Westview Press.

Frey, C. B., and Osborne, M. (2016a). The future of employment: How susceptible are jobs to computerisation? Discussion paper, Oxford Martin School.

Frey, C. B., and Osborne, M. (2016b). Technology at Work v2.0: The future is not what it used to be. Discussion paper, CITI GPSwReports.

Friedman, D. (1998). Evolutionary economics goes mainstream: A review of the theory of learning in games. *Journal of Evolutionary Economics*, 8(4): 423–432. ISSN 0936-9937. URL http://dx.doi.org/10.1007/s001910050071.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138. URL http://dx.doi.org/10.1038/nrn2787.

Future of Life Institute (2015). An Open Letter on AI. URL http://futureoflife.org/misc/open_letter.

Gigerenzer, G., and Brighton, H. (2009). Homo Heuristicus: Why biased minds make better inferences. *Cognitive Science*, 1: 107–143.

Gleick, J. (2011). *The information: A history, a theory, a flood*. Pantheon.

Goldberg, D. E. (2002). *The design of innovation: Lessons from and for competent genetic algorithms*. Norwell, MA, USA: Kluwer Academic Publishers.

Hastie, T., R. Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*. Springer, 2 edition. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Hayden, B. Y., and Platt, M. L. (2009). The mean, the median, and the St. Petersburg paradox. *Judgment and Decision Making*, 4(4): 256–272. URL http://EconPapers.repec.org/RePEc:jdm:journl:v:4:y:2009:i:4:p:256-272.

Hopkins, J. (2012). Psychoanalysis representation and neuroscience: The Freudian unconscious and the Bayesian brain. In A. Fotopoulu, D. Pfaff, and M. Conway (Eds.), *From the couch to the Lab: Psychoanalysis, neuroscience and cognitive psychology in dialoge*. OUP.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263–291. URL http://www.jstor.org/stable/1914185.

Kauffman, S. A. (2000). *Investigations*. Oxford University Press Oxford, New York.

Keynes, J. M. (1936). *The general theory of employment, interest and money*. Macmillan. 14th edition, 1973.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston, MA: Houghton Mifflin Co. URL http://www.econlib.org/library/Knight/knRUP.html.

Koestler, A. (1963). *The act of creation*. Hutchinson.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Lane, D. A., and Maxfield, R. (2005). Ontological uncertainty and innovation. *Journal of Evolutionary Economics*, 15(1): 3–50. DOI 10.1007/s00191-004-0227-7. URL http://dx.doi.org/10.1007/s00191-004-0227-7.

McCarthy, J. (1974). Professor Sir James Lighthill, FRS. Artificial intelligence: A general survey. *Artif. Intell.*, 5(3): 317–322. URL http://dblp.uni-trier.de/db/journals/ai/ai5.html#McCarthy74.

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *SIGART Newsletter*, 57.

Minsky, M. L., and Papert, S. (1988). *Perceptrons: An introduction to computational geometry*. Cambridge Mass.: MIT Press, expanded ed. edition.

Mirowski, P. (2002). *Machine dreams: Economics becomes a cyborg science*. Cambridge University Press. ISBN 9780521775267. URL https://books.google.it/books?id=GkrYxL0QtpcC.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc., 1 edition.

Newell, A., and Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M., and Chater, N. (2009). Precis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral And Brain Sciences*, 32: 69–120.

Oxford Dictionaries Online (2015). Definition of "Heuristic". URL http://www.oxforddictionaries.com/definition/english/heuristic.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 0-934613-73-7.

Rosenblatt, F. (1957). The Perceptron – A perceiving and recognizing automaton. Discussion paper 85-460-1, Cornell Aeronautical Laboratory.

Savage, L. J. (1954). *The foundations of statistics*. New York: Whiley.

Schwab, K. (2016). The fourth industrial revolution: What it means, how to respond. URL http://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423, 623–656.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–503. URL http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.

Simon, H. (1978). Rationality as process and as product of thought. *American Economic Review*, 68(2): 1–16. URL http://EconPapers.repec.org/RePEc:aea:aecrev:v:68:y:1978:i:2:p:1-16.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1): 99–118. URL http://dx.doi.org/10.2307/1884852.

Smith, G. D., and Ebrahim, S. (2002). Data dredging, bias, or confounding. *BMJ*, 325(7378): 1437–1438.

Tuckett, D. (2011). *Minding the markets : An emotional finance view of financial instability*. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan. URL http://www.worldcat.org/search?qt=worldcat_org_all&q=0230299857.

Tuckett, D., Smith, R. E., and Nyman, R. (2014). Tracking phantastic objects: A computer algorithmic investigation of narrative evolution in unstructured data sources. *Social Networks*, 38: 121–133. URL http://dx.doi.org/10.1016/j.socnet.2014.03.001.

http://dx.doi.org/10.5018/economics-ejournal.ja.2016-7