Artificial Intelligence: An Overview *

Vasant Honavar
Artificial Intelligence Research Laboratory
College of Information Sciences and Technology
301A IST Building
Pennsylvania State University

Last revised: August 22, 2016

Abstract

This chapter reviews common-sense definitions of intelligence; motivates the research in artificial intelligence (AI) that is aimed at design and analysis of programs and computers that model minds/brains; lays out the fundamental guiding hypothesis of AI; reviews the historical development of AI as a scientific and engineering discipline; explores the relationship of AI to other disciplines; and presents an overview of the scope, problems, and major areas of AI. Hopefully this discussion will not only provide a useful context for the technical material that follows but also convey a sense of what scientists, engineers, mathematicians, and philosophers who have been drawn to the field find exciting about AI. The views presented here are necessarily biased by the author's own research. The readers are encouraged to explore alternative views and perspectives on this subject.

1 What is Intelligence?

Try precisely defining intelligence. It is next to impossible. Despite the wide use (and misuse) of terms such as *intelligent systems*, there is no widely agreed-upon scientific definition of *intelligence*. It is therefore useful to think of intelligence in terms of an open collection of attributes. What follows is a wish-list of general characteristics of intelligence that contemporary researchers in AI and cognitive science are trying to understand and replicate. It is safe to say that no existing AI system comes anywhere close to exhibiting intelligence as characterized here except perhaps in extremely narrowly restricted domains (e.g., organic chemistry, medical diagnosis, information retrieval, network routing, military situation assessment, financial planning):

- **Perception** — manipulation, integration, and interpretation of data provided by sensors (in the context of the internal state of the system — including purposeful, goal-directed, active perception).
- **Action** — coordination, control, and use of effectors to accomplish a variety of tasks including exploration and manipulation of the environment, including design and construction of tools towards this end.
- **Reasoning** — deductive (logical) inference, inductive inference, analogical inference — including reasoning in the face of uncertainty and incomplete information, hypothetical reasoning, justification and explanation of inferences, evaluation of explanations, adapting explanations in the light of falsified assumptions or changing world states.
- **Adaptation and Learning** — adapting behaviour to better cope with changing environmental demands, discovery of regularities, explanation of observations in terms of known facts and hypotheses, construction of task-specific internal representations of the environment, discovery of procedures, learning to differentiate despite similarities and generalize despite differences, learning to describe specific

*Principles of Artificial Intelligence, Fall 2016, Pennsylvania State University ©Vasant Honavar, 1992-2016

domains in terms of abstract theories and concepts, learning to use, adapt, and extend language, learning to reason, plan, and act.

- **Communication** — with other intelligent agents including humans using signals, signs, icons, symbols, sound, pictures, touch, language and other communication media — including communication of goals, desires, beliefs, narratives of real and imaginary episodes, explanation of actions and events.
- **Planning and goal-directed problem-solving** — Formulation of plans — sequences or agenda of actions to accomplish externally or internally determined goals, evaluating and choosing among alternative plans, adapting plans in the face of unexpected changes in the environment, explaining and justifying plans, modifying old plans to fit new tasks, handling complexity by abstraction and simplification.
- **Autonomy** — Setting of goals, deciding on the appropriate course of actions to take in order to accomplish the goals or directives (without explicit instructions from another entity), executing the actions to satisfy the goals, adapting the actions and/or goals as necessary to deal with any unforeseen circumstances (to the extent permitted by the agent's physical capabilities and the environmental constraints).
- **Creativity** — exploration, modification, and extension of domains (e.g., language, mathematics, music) by manipulation of domain-specific constraints, or by other means.
- **Reflection and awareness** — of internal processes (e.g., reasoning, goals, etc.) of self as well as other agents.
- **Aesthetics** — articulation and use of aesthetic principles.
- **Organization** — into social groups based on shared objectives, development of shared conventions to facilitate orderly interaction, culture.

Most people would probably agree that the hallmark of intelligence is almost certainly not simply the ability to display some or all of the listed attributes but doing so on a broad and open-ended (not precisely pre-specified) set of domains and under a broad range of (a-priori unknown, possibly context-dependent and domain-specific) constraints (e.g., time allowed, tools available, accuracy desired). It is also clear that different systems — be they natural or artificial — can display different subsets of the attributes of intelligence to differing degrees.

2 What is Artificial Intelligence (AI)?

The attempt to understand intelligence entails building theories and models of (appropriately embodied) brains and minds, both natural as well as artificial. From the earliest writings of India and Greece, this has been a central problem in philosophy. The advent of the digital computer in the 1950's made this a central concern of computer scientists as well. The parallel development of the theory of computation (by John von Neumann, Alan Turing, Emil Post, Alonzo Church, Stephen Kleene, Markov and others) provided a new set of tools with which to approach this problem — through analysis, design, and evaluation of computers and programs that exhibit aspects of intelligent behavior — such as the ability to recognize and classify patterns; to reason from premises to logical conclusions; and to learn from experience.

The term *Artificial Intelligence* refers to the the enterprise of understanding and building intelligent systems. AI folklore credits John McCarthy (who incidentally, made several major contributions to AI and Computer Science in their infancy — by designing the programming language `LISP` and the first time-sharing operating system) with inventing the term during the workshop at Dartmouth in 1956 where the field took on its modern incarnation.

Here are some descriptions of AI:

AI is a science of intelligence. As a science, AI primarily involves falsifiable claims, i.e., testable hypotheses regarding the structures and processes that are necessary and sufficient for intelligent behavior.

AI is the study of computational models of intelligent behavior — perception, cognition, and action.

AI is the science of exploration of the space of possible and actual intelligent systems.

AI is the enterprise of design and analysis of intelligent agents.

AI is a branch of engineering which is concerned with the mechanization of tasks that are believed to require intelligence when performed by humans — e.g., proving theorems, planning trips, recognizing faces, diagnosing diseases, designing computers, composing music, discovering scientific laws, proving mathematical theorems, playing chess, writing stories, teaching physics, negotiating contracts, providing legal advice. However, all good engineering rests on a solid scientific foundation. AI is no exception.

The fundamental working hypothesis that has guided most of the research in artificial intelligence as well as the information-processing school of psychology is rather simply stated: *Cognition, or thought processes can, at some level, be modeled by computation.* The philosophical roots of this hypothesis can be traced at least as far back as the attempts of Helmholtz, Leibnitz and Boole to explain thought processes in terms of mechanical (or in modern terms, algorithmic or computational) processes. This has led to the *functional* view of intelligence which is shared explicitly or implicitly by almost all of the work in AI. Newell's *physical symbol system hypothesis*, Fodor's *language of thought*, Minsky's *society of mind*, Holland's *classifier systems* and genetic algorithms, and most artificial neural network models are all specific examples of this functional view. In this view, *intelligence* can be characterized abstractly as a functional capability independent of any commitments as to the specific physical substrates that support the functions in question.

How do we know this working hypothesis to be true? Well, we don't. In AI, like any other scientific discipline, working hypotheses are useful in formulating and testing ideas, models, and theories about the chosen domain. As long as a hypothesis is guiding us toward fruitful avenues of investigation and there is no better alternative in sight, we stick with it. But note that it is no simple task to choose among competing hypotheses. Indeed, such a choice is not possible on purely objective, rational, or scientific grounds. The interested reader is refered to the writings of philosophers of science (e.g., Popper, Kuhn, Peirce) for detailed discussions of this question. The important thing to remember is that the current working hypothesis of AI, like any other working hypothesis in science, is subject to revision or replacement if experimental evidence so warrants. Despite the fact that this hypothesis has so far led to many interesting insights into the nature of cognition as well as a number of useful technological developments, and despite — or perhaps because of — its use in some form by most researchers in AI — there is considerable ongoing philosophical debate on its meaning and its validity. Therefore, it is probably worthwhile spelling out in some detail what the hypothesis entails. This involves specifying what "computation" is.

3 Computation, Computers, and Programs

The development of the formal notion of computation — and hence, the establishment of the theoretical foundations of computer science can be traced at least as far back as Alan Turing's encounter (as a student of the mathematical logician M. H. A. Newman at Cambridge in the mid 1930s) with Hilbert's decision problem. The essence of Hilbert's decision problem is whether there exists an effective procedure for determining whether or not a certain conclusion logically follows from a given set of axioms.

The formal notion of an algorithm as we understand it today is primarily the result of Turing's attempt to formalize the intuitive notion of an effective procedure (although the informal notion can be traced at least as far back as the middle-eastern author Al Khowarizmi who wrote a textbook on mathematics around 825 A.D.). Essentially, an algorithm is a set of rules that precisely specify what to do under any given set of circumstances e.g., "2+2 = 4" or "$area = length \times width$", etc. To reflect the mechanical nature of what is involved in carrying out the steps of an algorithm, Turing invented a hypothetical computer now called the Turing machine and used it to formalize Hilbert's decision problem. Essentially, a Turing machine has a processor *head* that reads (and erases) symbols from and writes symbols to a potentially infinite memory *tape* that is divided into squares each of which is capable of storing a symbol. The behavior of a Turing machine is governed by an algorithm which is realized in terms of what we now call a *program.* A program is a finite set of instructions selected from a *general-purpose* set of simple instructions. A necessary and sufficient set of instructions is simply: 'fetch (from the tape)', 'store (onto the tape)', 'move the head one square to the left ', 'move the head one square to the right' and 'branch conditionally (if the current square contains a 0, do a, else do b)'. Turing also showed that there is a *universal Turing machine* — one that can compute anything that any other Turing machine *could possibly* compute — provided both the necessary program describing the computation and the data to be used are stored on its tape. So long as the tape is large enough to handle all the necessary data (including input, intermediate results, and the actual programs that

instruct the machine what to do), such a universal system can compute anything that can be described. The *potentially infinite tape* (which serves as the system's memory) is simply a way to ensure the system won't run out of space. Turing proved that any parallel or/and serial structure of describable processes (that is, processes that could be stated in the form of input-output functions) can be executed by such a machine.

Turing and others also showed that there could be many different versions of such a universal information-processing/computing system, including ones with 2, 3, or any number of heads and/or memories, all of which are equivalent to the universal Turing machine. About the same time, Alonzo Church and his students Stephen Kleene and Barkley Rosser at Princeton had begun to investigate several frameworks for computation including the *lambda-calculus* and *recursive functions*. Still other formulations were made, including McCulloch-Pitts neural networks, Markov algorithms, *Petri nets*, and two systems by Emil Post—an independent invention of the Turing machine and *Post productions*. In addition, all of these (given potentially infinite memory) were *proved* to be *exactly* equivalent to the Turing machine. That is, any structure of parallel and/or serial processes that could be executed by a Turing machine could in principle be executed by any of these other systems and vice versa.

There are hundreds of different general-purpose programming languages, but — just as with computers — they are all general-purpose: Anything that can be programmed (that is, that is describable) can be programmed in any language.

Since any computer can execute any structure of processes, a program can be written for any computer that translates (compiles or interprets) any programming language into that computer's machine language. Almost all computers have translators for at least one easier-to-use "higher-level language." But the more important point is that a translator from any language to any computer's machine language, or for that matter to any other general-purpose language, can always be built. And any general-purpose language or computer can do anything that any other can. However, the same computation may be millions of times faster on one computer than another, so from any practical point of view that may not be feasible. But eventually, given enough time, the results will be the same.

Turing also proved that there are limits to what can be done with computers, since there are problems where the system can never know before finding a solution (e.g., proof of a theorem in first order logic) whether there exists one and hence may never halt. Some people have used these limits to argue that computers therefore cannot possibly ever be intelligent in all the ways that human beings are. It is important to make clear that this is *not* an issue of whether computers can think, or can possibly ever be made to behave intelligently. Rather, whatever it is that a general-purpose information processor can do can *therefore* be done by writing a program in any language for any general-purpose computer.

It should be pointed out however, that recent developments in the theory of computation offer models of computation that are potentially more powerful than the Turing model and exploit quantum mechanical phenomena. The implications of *quantum computation* to computer science in general, and artificial intelligence in particular, remain to be understood. Computation using biological substrates (e.g., DNA, RNA,and proteins) offer other exciting possibilities that are only beginning to be explored.

Getting back to Turing's framework, the design of computers and programming languages often grew out of, and was strongly influenced by, one or another of the different but equivalent formal models of computation. Thus the traditional serial stored-program computers and their basic machine languages have much the flavor of a single-head Turing machine. Turing was invited to Princeton by Alonzo Church in 1936. Since von Neumann and Turing probably talked during that time, it seems likely that von Neumann's thinking on stored-program computers was influenced by Turing's work. The first electronic digital computer was built by John Vincent Atanasoff and his student Clifford Berry at Iowa State University in the late 1930's and early 1940s.

The design of the vast majority of high level programming languages that are in use today was directly or indirectly influenced by von Neumann's design of the stored program computer. The popular artificial intelligence programming language LISP (invented by John McCarthy, and later made usable as a programming language by McCarthy's student Steve Russell who wrote the first interpreter for it) is based on lambda calculus and recursive functions (The design of LISP itself was probably influenced by IPL, an earlier symbol processing language invented and used by Newell and Simon to develop their theorem-proving program, the *Logic Theorist*). COMIT, SNOBOL, and modern knowledge-based expert systems have origins in Post productions.

Neural (connectionist) networks are derivatives of McCulloch-Pitts networks which in turn were influenced

by the pioneering work of Rashevsky's group at the University of Chicago. (McCulloch worked in Chicago the 1940's and Pitts was a student of Rashevsky's.)

4 A Brief History of AI

The attempt to understand intelligence, and hence the enterprise of AI, entails building and testing models of minds and brains both actual and possible. The central questions concerning the working of minds/brains are as old as the origin of our remote ancestors that roamed the forests of Africa, plains of Asia, and the mountains of Europe. An extensive study of the intellectual history of this field involves nothing short of an exhaustive exploration of related evolutionary, historical, philosophical, scientific, and technological developments of human-kind through the ages.

Greek mythology speaks of Prometheus who earned the wrath of Gods of Olympus when he sought to steal for the benefit of the human race, not only fire, but also the gift of intelligence or *nous* (the rational mind). The notion that human efforts to gain knowledge constitutes a transgression against Gods is deeply ingrained in the Western thought as exemplified by the works of Dante, Milton, and Shakespeare. The belief that the desire for knowledge must ultimately lead to disaster has survived the Renaissance, the Age of Enlightenment, and even the scientific and technological advances of the 19th and 20th centuries.

Some of the earliest writings of the civilizations of Asia, namely Vedas and Upanishads raise the most profound questions about existence, and knowledge. They offer (largely unscientific, but nevertheless fascinating and often insightful) theories about intelligence. Despite evidence that knowledge and enlightenment were pursued by the Vedic seers at the risk of considerable physical hardship, the writings from that period also indicate a fear of the consequences of knowledge.

Mary Shelley, in her introduction to *Frankenstein* (subtitled *The Modern Prometheus*), shows us the extent to which scientific advances such as the work of Darwin and the discovery of electricity had convinced even non-scientists that the workings of nature, once regarded inscrutable divine secrets, could be understood and exploited by the human race. Frankenstein's monster is not the result of shamanistic rituals; it is assembled from manufactured components and powered by electricity. By the time Mary Shelley brought together modern science and the Prometheus myth, the work on philosophical and scientific foundations of artificial intelligence had already been underway for centuries.

Instead of dispelling the ancient fear of intelligence and knowledge, modern technology has only made these consequences appear more imminent. The legend of Prometheus has been retold many times in the language of the technological society. Thus, it is not surprising that artificial intelligence is the subject of controversy in academic, intellectual, and popular circles.

Although potentially extremely interesting and rewarding, a detailed study of the history of scientific and philosophical thought leading up to current research in artificial intelligence is beyond the scope of this handout. However, it is essential to have at least a glimpse of this history because it is impossible to appreciate where we are without some knowledge of how we got there.

What follows is a list of a few key landmarks along the path to AI (Note: the dates are approximate and refer to the period in which the work appeared):

- **Aristotle** (384 – 322 BC) distinguishes *matter* from *form* (e.g., a sculpture of Aristotle is made from the material bronze and has the form of a human), thereby laying the seeds of abstracting the medium from its representation which is at the heart of modern computer science; lays the foundations of epistemology (the science of knowing) and logic in the Western world.

- **Panini** (350 BC) develops a formal grammar for Sanskrit laying the foundations of syntactic models that led to Chomsky's theory of syntactic structures in 1956.

- Al Khowarizmi (825) introduces to the west, the eastern mathematical tradition which largely consisted of mathematical recipes i.e., *algorithms* in his text explaining the Indian system of numeration which was translated into latin under the title *Algorithmi de numero Indorum* as well as the Arabic *algebra*

- **Descartes** (1556–1650) discounts sensory experience as untrustworthy and justifies his own existence in terms of thought: *Cogito ergo sum* (I think, therefore I am) and with other contemporary thinkers, establishes the notion that the structure of ideas about the world are not necessarily the same as the

structure of their subject matter, an idea which underlies much of the methodology of AI, epistemology, psychology, mathematics, and modern literature.

- **Hobbs** (1650) proposes that thinking is a rule-based computational process analogous to arithmetic.
- **Leibnitz** (1646-1716) seeks a general method in which all truths will be reduced to a kind of calculation.
- **Boole** (1815-1864) puts forth his study of logic and probability as an investigation into the laws of thought.
- **Russell, Frege, Tarski** (1910-1950) formalize logic and reduce large portions of mathematics to logic; Russell writes an influential book *Principia Mathematica.* Tarski introduces the theory of reference for relating objects in a logic to objects in the world, laying the foundations of formal semantics.
- **Hilbert** (1862-1943) presents the decision problem . Is there an effective procedure for determining whether or not a given theorem logically follows from a given set of axioms?
- **Godel** (1906-1978) shows the existence of an effective procedure to prove any theorem in Frege's logic and proves the incompleteness theorem
- **Turing** (1912-1954) invents the Turing Machine to formalize the notion of an effective procedure
- **Turing, Church, Kleene, Post** (1930-50) Turing and Church put forth the Church-Turing thesis that Turing machines are universal computers. Kleene and Post propose other Turing-equivalent models of computation
- Several special purpose analog and digital computers are built (including the Atanasoff-Berry Computer)
- **Chomsky** (1956) develops the Chomsky hierarchy of languages formalize the common-sense notion of computation in terms of effective procedures and universal computers.
- **Rashevsky, McCulloch, Ashby, Rosenblatt** (1930-60) — work on early neural network models.
- **Von Neumann, McCulloch** (1940-1956), investigate the relationship between the brain and the computer
- **Von Neumann and Morgenstern** (1946) develop a formal framework for rational decision making under uncertainty
- **Shannon** (1948) develops information theory, laying the foundations of coding and communication
- **von Neumann (1956)** works out a detailed design for a stored-program digital computer
- **Wiener, Lyapunov (1956)** develop the science of cybernetics to study control and communication in humans and machines.
- **McCarthy, Minsky, Newell, Selfridge, Simon, Turing, Uhr, et al.** (1956) establish AI in a workshop organized at Dartmouth College at the initiative of John McCarthy.
- Several digital computers are constructed and universal languages for programming them are developed e.g., Lisp, Snobol, Fortran.
- **Wittgenstein** (1950) challenges many of the underlying assumptions of the rationalist tradition, including the foundations of language, science, and knowledge itself; and proposes that the meaning of an utterance depends on being situated in a human, cultural context (e.g., the meaning of the word "chair" to me is dependent on my having a physical body that can take on a sitting posture and the cultural convention for using chairs).
- **Dantzig and Edmunds** (1960-62) introduce reduction, a general transformation from one class of problems to another

- **Cobham and Edmunds** (1964-65) introduce polynomial and exponential complexity
- **Cook and Karp** (1971-72) develop the theory of NP-completeness which helps recognize problems that are intractable
- **Cerf** (1974) invents the Internet
- **Husserl, Heidegger** (1960-1975) articulate the view that abstractions must be *rooted* or *grounded* in the concrete *lavensvelt* or *life-world* i.e., the rationalist model of Aristotle is very much secondary to the concrete world that supported it; in the existentialist/phenomenological view, intelligence is not knowing what is *true*, but knowing how to cope with a world that is constantly changing.
- 1960-1970 — untempered optimism fueled by early success on some problems thought to be hard (e.g., theorem proving) is tempered by slow progress on many problems thought to be easy (e.g., visual pattern recognition); much of the AI work is based in the rationalist/logical tradition; the field is fragmented into sub-areas focused on problem-solving, knowledge representation and inference, vision, planning, language processing, learning, etc.
- 1970-mid 1980s — investigation of knowledge representation and reasoning leads to many practical tools such as expert systems; the difficulty task of *knowledge engineering* draws attention to the need for systems capable of learning from experience and interaction with the world
- Internet is rolled out in 1984
- Mid 1980s-1990 — some of the failures of rationalist/logical approaches to AI lead to renewed interest in biologically inspired neural network and evolutionary models which lead to modest successes on some problems (e.g., pattern recognition) prompting a few to prematurely proclaim the death of "good old fashioned AI". Some propose alternative approaches (in the Heideggerian tradition) while others discover (and rediscover) the limitations of the alternatives being proposed.
- Mid 1980s-mid 1990s — progress in algorithmic models of learning begins to offer promising and practical alternatives to knowledge engineering and AI technologies begin to be used in critical components of large software systems; The most successful approaches incorporate the elements of both the rationalist/logical/symbolic tradition and the existential/phenomenological/non-symbolic tradition; proposals for reconciling the two approaches begin to appear; maturing of the several subfields of AI such as vision, language processing, knowledge representation, planning, etc. leads to insights on the capabilities as well as limitations of the techniques that were developed and redirects attention on the problem of building intelligent agents as opposed to subsystems and this further fuels synthesis of hybrid models.
- **Berners-Lee** (1989-91) invents the world wide web
- Mid 1990s-present — AI technologies continue to find applications in web search, big data analytics, natural language processing, computer vision, customizable software, smart devices (e.g., homes, automobiles), agile manufacturing systems, autonomous vehicles, health informatics, medical informatics, etc. slow but steady progress on fundamental AI research problems continues. Synthesis of traditional logic-based systems, soft and adaptive computing technologies (e.g., neural networks, probabilistic models, etc.), crossdisciplinary work in cognitive sciences, and synthesis of software agents and multi-agent systems leads to the emergence of *nouvelle* AI which views intelligence as an emergent behavior resulting from interactions (e.g., communication, coordination, competition) among large numbers of autonomous or sem-autonomous entities (be they neurons, computer programs, individuals) that are situated in the world, display structure and organization at multiple spatial and temporal scales, and interact with the world through sensors and effectors; a host of fundamental problems such as design of individual agents, inter-agent communication and coordination, agent organizations, become topics of active research.

5 Relation of AI to other disciplines

The invention of digital (and analog) computers in the 1940s and 1950s and the work in the theory of computation, information theory, and control that accompanied it provided the experimental tools and the theoretical underpinnings of AI research. Much related work has taken place in related fields addressing similar questions (e.g., bionics, cybernetics, neural networks, statistical pattern recognition, syntactic pattern recognition, expert systems, computer vision, robotics, computational linguistics, decision theory, cognitive psychology, artificial life, computational neuroscience, computational organization theory, etc.). AI, broadly interpreted, is closely intertwined with, and often subsumes, much of the work in most of these fields.

AI is often regarded as a branch of computer science. AI's special relationship with computer science is due to the fact that the language of computation is to the study of mind what calculus was to the study of physics. Calculus provided the mathematical tools for formulation of questions and search for answers in classical physics (It should come as no surprise that Newton and Leibnitz were among the inventors of calculus). But physics is more than calculus; it developed its own armamentarium of experimental methods to probe its domain — the physical universe. AI (and more recently, cognitive science) continue to develop their own experimental tools and theoretical frameworks. In the process, AI has contributed over the years, a wide variety of concepts and tools to Computer Science — `LISP` — one of the earliest high-level programming languages, the first multi-tasking operating system, logic programming, constraint programming, heuristic search, object-oriented programming, neural networks, computational learning theory, temporal logic, deductive databases, high-dimensional grammars, evolutionary programming — to name a few. AI problems have stimulated research in other areas of computer science — massively parallel architectures for vision, theoretical research in complexity of reasoning and learning, and so on. AI is occasionally viewed as a sibling of psychology. Psychology is concerned with the formulation and experimental verification of theories of behavior with human or animal subjects. AI is concerned with computational models that exhibit aspects of intelligent behavior. It is not generally committed to any particular (e.g., human-like) set of mechanisms or any particular ways of implementing the chosen mechanisms. Yet, the information processing models coming out of AI research have strongly influenced contemporary research in human and animal psychology and neuroscience.

Insofar as intelligent behavior is normally associated with living systems, AI shares some of the concerns of the field of study that has been provocatively, and misleadingly, labeled *artificial life*.

AI can also be thought of as applied epistemology (the branch of philosophy that is concerned with the nature of knowledge). AI research has brought to light entirely new questions and new ways of looking at old problems in epistemology.

AI is often treated as a branch of engineering that is concerned with the design, implementation, and evaluation of intelligent artifacts. AI research has resulted in a number of useful practical tools (programs that configure computer systems, diagnose faults in engines, software agents that scour the Internet for information on demand, etc.).

AI attacks a long-standing mix of problems from a number of more established disciplines like philosophy, psychology, linguistics, anthropology, engineering, and neuroscience. While freely borrowing from these disciplines, it brings to the study of intelligent behavior, a unique approach, and a unique set of tools and in the process, sometimes raises entirely new questions due to its use of computation as a substrate for theory-construction and experimentation. This has led to arguably one of the most important scientific developments of this century, the birth of *Cognitive Science* (which attempts to integrate insights and results from its constituent disciplines better than most (though by no means all) of the work in AI). All of this gives us a new perspective on some of the long-standing questions about the nature of mind. But it does not make the questions themselves necessarily any easier!

Every discipline has a domain of enquiry. For AI, it is the entire range of human and non-human intellectual enterprise spanning the entire space of actual and possible intelligent adaptive systems. As a result, AI gets deeply involved in the conceptual and methodological questions in any area in which it is applied: The use of AI in synthesis of artistic objects (e.g., drawings and paintings) necessarily has to involve an understanding of the specification of ways of representing the knowledge used by an artist as well as theories about creativity in the domain of art; the use of AI tools to model the process of scientific exploration in some area (say molecular biology) necessarily entails an understanding of the scientific method and is likely to yield new insights on hypothesis formation, experimental design, and theory selection in that

area. As a consequence, AI is one of the most interdisciplinary fields of study currently taught in our universities.

6 Goals of AI

The primary goal of AI research is to increase our understanding of perceptual, reasoning, learning, linguistic and creative processes. This understanding is no doubt helpful in the design and construction of useful new tools in science, industry, and culture; Just as the invention of the internal combustion engine and the development of machines like airplanes resulted in unprecedented enhancement of the mobility of our species, the tools resulting from AI research are already beginning to extend human intellectual and creative capabilities in ways that our predecessors could only dream about. Sophisticated understanding of the underlying mechanisms and the potential and limits of human as well as other forms of intelligence is also likely to shed new lights on the social, environmental, and cultural problems of our time and aid the search for solutions.

7 AI and Society

As many recent studies, including one conducted by the World Economic Forum on the Future of Jobs note, the Fourth Industrial Revolution unleashed in part, by the advances in artificial intelligence (including in particular, machine learning and robotics), can be expected to bring about widespread disruption not only to business models as well as labor markets. The Economic Report of the President published earlier this year predicts that there is an 83% chance that workers who earn $20 an hour or less could have their jobs replaced by robots in the future. For example, the prospects of automated driving, a capability that is expected to become a reality within the next few years, underscores the moral imperatives of artificial intelligence. On the one hand, the potential benefits of automated driving are extremely compelling: The number of deaths from car accidents worldwide every year is around 1.25 million. Given that over 90% of the accidents are caused by human error, we could save over a million lives and avoid countless injuries by automating driving. On the other hand, automation of driving would result in a significant job loss. In the U.S. alone, nearly 10% of all jobs involve operating a vehicle and the majority of these jobs would be lost as a result of automation of driving. Assuming that the saving of lives and prevention of injuries must take precedence, we need to develop and deploy automated driving. However, we can only do so if we are prepared to address the downside of automated driving - the loss of employment for millions (truck drivers, bus drivers, taxi operators, etc.) who currently make their living by driving a vehicle. As advances in artificial intelligence lead to automation of tasks requiring significant levels of expertise and skills in fields like healthcare, law, journalism, banking, among others, what will happen to the workforce that was once occupying those jobs? Will the same technological advances that are making many of the jobs of today obsolete, help create entirely new types of jobs, and improve the quality of life for the general population as the industrial revolution (1760-1840) did, for the first time in history? How can individuals equip themselves with the knowledge and skills throughout their life that are critical for success in a world where the jobs available as well as the skills needed both change at a rapid pace? How can we develop systems that optimally leverage the unique and complementary strengths of artificial intelligence and robots on the one hand and humans on the other? How can societies anticipate and respond to the technology driven disruptive changes that are almost surely on the horizon, so as to maximize their benefits and minimize their harm to the society at large?

8 Practical AI

AI, in short, is about the design and implementation of *intelligent agents*. This requires the use of AI tools and techniques for *search*, *knowledge representation*, and *adaptation and learning* and their application to *problem-solving, planning, analysis, design, knowledge acquisition, discovery*, etc. Some would argue that much of contemporary AI work essentially involves reducing problems requiring intelligence into search problems using appropriate ways of representing the knowledge necessary for the solution of such problems. A

side-effect of this is a wide range of practically useful tools (theorem-provers, game-players, vision programs, natural language interfaces, stock-market analysts, programmer's assistants, assembly line robots, physician's assistants, tutoring programs, architect's assistants, internet softbots, and so on).

8.1 Problem-Solving as State Space Search

The dominant paradigm for problem solving in AI is *state space search*. It can be shown that problem-solving, planning, learning, scientific discovery, mathematical theorem proving, etc. are at an abstract level, essentially search problems. A lot of work in AI has to do with the detailed reduction of such problems to (and the solution of the resulting) search problems.

States represent snap-shots of the problem at various stages of its solution. Operators enable transforming one state into another. Typically, the states are represented using structures of symbols (e.g., lists). Operators transform one symbol structure (e.g., list, or a set of logical expressions) into another. The system's task is to find a path between two specified states in the state-space (e.g., the initial state and a specified goal, the puzzle and its solution, the axioms and a theorem to be proved, etc.).

In almost any non-trivial problem, a blind exhaustive search for a path will be impossibly slow, and there will be no known algorithm or a procedure for directly computing that path without resorting to search. As a result, much early work in AI focused on the study of effective heuristic search procedures. For example, AI systems handle games like chess as follows: The initial board is established as the given, and a procedure is coded to compute whether a win-state has been reached. In addition, procedures are coded to execute legal moves and (usually) to compute heuristic assessments of the promise of each possible move, and to combine the separate heuristic assessments into an overall value that will be used to choose the next move. Finally, all these are put into a total structure that applies the appropriate heuristics, combines their results and evaluates alternative moves, and actually makes a move, then waits for and senses the opponent's moves, uses it to update the board (probably checking that it is indeed a legal move), and loops back to make its own next move. (For simplicity the look-ahead with minimax that most game-players use has been ignored, but that is essentially more of the same.) Conventional search algorithms designed for serial von Neumann computers, artificial neural network realizations of certain search algorithms, and the so called genetic algorithms — are all — despite misguided assertions by some to the contrary — examples of this general class of AI problem-solving techniques.

Search problems are hard because they tend to be computationally intractable. The size of the space being searched typically grows exponentially with the size of the problem. There are only 26^4 or 4.6×10^5 ways in which to enter letters into a 2×2 cross-word. But a typical NY times crossword may admit 26^{190} or 6.9×10^{268} possibilities. A good bit of work in AI has to do with finding ways to attack such search problems with limited computational resources available in practice. (Some have suggested — in a lighter vein — that AI is a branch of computer science which seeks to find computationally feasible solutions to NP-hard problems). This might involve the analysis of various search spaces, design and evaluation of suitable *heuristics* that support efficient search strategies, and the design of mechanisms for automating the discovery of appropriate heuristics, and so on. Even where AI systems do not explicitly make use of search algorithms, search serves as a useful metaphor for thinking about tasks requiring intelligence.

Search in general can be guided by the knowledge that is at the disposal of the problem solver. If the system is highly specialized, the necessary knowledge is usually built into the search procedure (in the form of criteria for choosing among alternative paths, heuristic functions to be used, etc.). However, general purpose problem solvers also need to be able to retrieve problem-specific and perhaps even situation-specific knowledge to be used to guide the search during problem-solving. Indeed, such retrieval might itself entail search (albeit in a different space). Efficient, and flexible representations of such knowledge as well as mechanisms for their retrieval as needed during problem solving are, (although typically overlooked because most current AI systems are designed for very specialized, narrowly defined tasks), extremely important.

8.2 Knowledge representation

Any intelligent system has to know a great deal about the environment in which it is situated. It is generally accepted in artificial intelligence and cognitive science that knowledge has to be *represented* in some form in order for it to be used. Much effort in AI is devoted to finding ways to acquire and encode such knowledge

in a form that can be used by the machine. This is free of any commitment as to how a particular piece of knowledge is internally represented. However, implicit in this view is a commitment to use *some* language (e.g., first order logic, production rules, lambda calculus or LISP) to express and manipulate knowledge. Expressions in any such language can be syntactically transformed into any other sufficiently expressive language — this follows from the universality of the Turing framework. This is tantamount to saying that systems that use knowledge are simultaneously describable at multiple levels of description. And systems (such as living brains or robots) that exist in the physical world would have physical descriptions — just as the behavior of a computer can be described at an abstract level in terms of data structures and programs, or in terms of machine language operations that are carried out (thereby making the function of the hardware more transparent) or in terms of the laws of physics that describe the behavior of the physical medium which is used to construct the hardware.

8.2.1 Nature of Knowledge Representation

Given the central role played by knowledge representation in functional approaches to understanding and engineering intelligence, the *nature of representation* is among one the most fundamental questions in artificial intelligence and cognitive science. Some insight into this question can be obtained by considering a concrete example. A common way to represent knowledge is with logic. It is important to emphasize that logic is not the knowledge itself; it is simply a way of representing knowledge. (However, logic can be viewed as a form of meta-level knowledge about how to represent and reason with knowledge.) What logic enables us to do is represent the knowledge possessed by an agent using a finite set of logical expressions plus a process (namely, the inference rules of logic) for generating a (potentially unlimited) set of other logical expressions that are part of the agent's knowledge. Thus if we represented an agent's knowledge in the form of expressions a and b, and if $a \wedge b \models c$, the agent has (implicit) knowledge of c even though c was not part of the (explicit) representation. In fact, first order logic is universal in that it is powerful enough to represent essentially any knowledge that can be captured by a formal system. However, for certain types of knowledge to be used for certain purposes (e.g., knowledge of the sort that is captured by maps of some geographical region or a city), first order logic representation may be awkward, indirect, or overly verbose.

If on the other hand, we were to choose a different way of representing knowledge of an agent, one which did not permit any logical deduction, then the agent's knowledge could be limited to those expressions that were explicitly included in the representation. Such a representation is in essence, simply a lookup table for the expressions in question. Thus, (for lack of a better term), the *knowledge content* of a representation may be limited by restricting either the inferences allowed, the form of the expressions that may be included (that is, limiting the expressive power), or both. Indeed, it is often necessary to impose such limits on the power of representation in order to make their use computationally feasible (perhaps at the expense of logical soundness, completeness, or both).

In order for any system to serve the role of a representation (as used in most artificial intelligence and cognitive science theories) it must include: an encoding process that maps the physical state of the external environment into an internal state; processes that map aspects of the physical state(s) of the external environment into appropriate (internal) transformations of the internal state; a decoding process that maps an internal state into a physical state of the external environment — all subject to the constraint that the result of decoding the result of application of internal transformations of an internal state (obtained from encoding a physical state of the environment) is the *same* as the result of directly transforming the physical state of the external environment. This is perhaps a stronger requirement than is necessary — most likely influenced by the emphasis on logic. It is easy to see several ways of relaxing this constraint — by allowing the correspondence to be only *approximate* instead of exact, or attainable only with a certain probability. Representation, and the associated inference mechanisms, in the broadest sense of the terms, need not be complete or precise or explicit. In short, representations are caricatures of selected aspects of an agent's environment that are operationally useful to the agent. Thus, certain mental operations on the representation can be used to predict the consequences of performing corresponding physical actions on the environment in which the agent operates.

Note that the internal transformations may be performed using LISP programs, or production systems, a suitably structured artificial neural network, or a collection of finite state automata (among other possibilities).

An additional requirement for representations thought to be essential by many (e.g., Newell) is that the application of encoding (sensing), internal transformations, and decoding (acting) must be executable on demand to the extent required to serve the purposes of the organism (which could be viewed essentially as the sensed internal environment of needs, drives, and emotions).

8.2.2 Where Do The Representations Come From?

Representations may be discovered by organisms (or evolution) by identifying the right medium of encoders (transducers) and decoders (effectors) and the right dynamics for the transformations for specific tasks. This would lead to a large number of task-specific *analogical* representations. Indeed, strong evidence for such analogical representations can be found in living brains: the retinotopic maps in the visual cortex and the somatotopic maps of the sensory-motor cortex provide good examples of analogical representations.

Alternatively, or in addition, a set of encoders and decoders may be used in conjunction with the ability to compose whatever sequence of transformations that may be necessary to form a representation. Many AI systems take this route to the design of representations — by using a sufficiently general language (e.g., LISP, or primitive functions computed by simple processors of an artificial neural network) that allows the composition of whatever functions that may be necessary to satisfy the task-specific desiderata of representations.

Irrespective of the approach chosen, the discovery of adequately powerful, efficient, and robust representations for any non-trivial set of tasks is still a largely unsolved problem. In early AI systems, the representations used by the system were hard-wired by the designers of the system. But it is unrealistic to expect that the designer of the system can provide adequate representations for anything but precisely defined, narrowly restricted caricatures of environments of the sort that a robot or a child has to contend with. So then where do such representations come from if they are not hard-wired by design? This question has brought into focus learning and evolutionary processes as possible candidates that enable AI systems to select, transduce, encode, abstract, transform, and operationalize signals, knowledge and skills necessary to interact successfully with the environment.

Learning and evolution must build the representations that perception and cognition utilize. One of the most informative characterizations of learning to date is in terms of storage of results of inference in a form that is suitable for use in the future. Learning can clearly provide an avenue for the discovery of the necessary compositions of transformations which is a major aspect of representation. This makes the initial representation or encoding extremely important. If it is not properly chosen (by the designer of the system or by evolution), it places additional (and perhaps insurmountable) burdens on the learning mechanisms (e.g., if the initial representation failed to capture spatial or temporal relations at a level of detail that is necessary for dealing with the problem domain). The past decade has seen much progress in machine learning although many problems remain to be solved. Learning algorithms, once mere laboratory curiosities, have now become essential components in the repertoire of tools used to build flexible, adaptive, information systems for a wide variety of practical applications (e.g., investing in the stock market, detection of credit card fraud, medical diagnosis, autonomous robots, computer operating systems, and adaptive intelligent agents).

8.2.3 Semantic Grounding of Representations

From the discussion above, it must be clear that many AI systems presuppose the existence of some representation before they can discover other useful representations. Therefore it appears that representations cannot come into existence without the existence of physical transducers and effectors that connect such systems with the physical world, leading to the *grounding problem*. Many in the artificial intelligence and cognitive science research community agree on the need for *grounding* of symbolic representations through sensory (e.g., visual, auditory, tactile) transducers and motor effectors in the external environment on the one hand and the internal environment of needs, drives, and emotions of the organism (or robot) in order for such representations (which are otherwise devoid of any intrinsic meaning to the organism or robot) to become imbued with meaning or semantics. At the same time, one cannot rule out the possibility that a large part of the representational machinery in intelligent systems may not make use of discrete symbols at all. (Mark Bickhard, (among others), has advocated a continuous dynamical systems approach to modeling intelligent systems). In any event, it seems that the case for *embodied* or *environmentally situated* intelligence is rather strong. Thus autonomous robots provide an interesting testbed for artificial intelligence.

9 How do we know if an agent is intelligent?

No one who is engaged seriously in the enterprise of AI can ignore the question — How do we know when AI has succeeded? In other words, how do we know if an agent (natural or artificial) is intelligent? Turing, among others, suggested that a system or agent can be said to be intelligent when the agent's performance cannot be distinguished from that of a human performing the same task. Consider the following recipe suggested by Turing: Put an artificial agent or a computer and a human in one room and a human interrogator in another room. The interrogator can direct questions to either the human or the computer (or agent) refering to one as A, and the other as B. The interrogator is not told the identity of A or B. The only means of communication allowed is through an intermediary system (say a a terminal and a keyboard). A and B both compete to convince the interrogator that he/she/it is the human. If the artificial agent or computer wins on the average as often as the human, it is said to have passed the Turing test for intelligence.

While Turing test offers a starting point, it is far from clear that Turing had the last word on the subject. For example, is it necessary for an agent to display the sorts of weaknesses that humans are prone to (e.g., in performing long and involved arithmetic calculations quickly) in order for it to pass the test?

We can get more sophisticated with the test — there is no need to restrict the domain of questions to those that can be entered or answered by words typed on a keyboard. For example, the interrogator may ask A or B to perform certain tasks (move a chair to a certain location in the room; draw a picture; describe the taste of some item of food that is being cooked in the kitchen next door...).

Why does the definition of intelligence have to be so anthropocentric? It seems somewhat arbitrary to evaluate an intelligent entity (whether natural or artificial) without regard to the context (i.e., the environment) in which it evolved. Perhaps the most important test for intelligence is an entity's ability to survive as a species in a changing environment. Perhaps we ought to be open to forms of intelligence other than our own. Perhaps it is downright arrogant if not stupid of us not to recognize the immense richness and variety of intelligence found in the world around us – from compassionate chimpanzees that protect infants of another species to dogs that can sense an impending epileptic seizure and do everything that they can to protect their epilepsy-prone human companions to parrots that can be taught abstract concepts. Perhaps the question of whether an agent is intelligent is a shallow one. Perhaps "intelligence", like a number of other loaded terms (e.g., "life", "love", "soul") used in everyday language turns out to be too complex to assign a scientifically agreed–upon meaning. Perhaps it is more meaningful to ask what sorts of tasks an agent can perform well, what kinds of inference it can handle, what sorts of knowledge acquisition mechanisms it has, and so on. Perhaps. Most of these issues are the subject of ongoing debate in AI and cognitive science.

10 Quo Vadis?

The history of AI, like that of any other science, is one that is filled with surprises, false starts, and occasional successes. Probably the most important lesson to date from research in AI is that problems that were thought to be easy (i.e., the ones that a two-year-old can handle effortlessly — e.g., recognizing faces) are extremely difficult to mechanize (given the current state of art); Problems that were thought to be hard (i.e., the ones that require years of formal training — e.g., proving theorems in mathematics — not to be confused with deciding what theorems to prove) are embarassingly easy to mechanize. Is it really the case that cognition can be fully understood in terms of computation? Are cold logic and rationality sufficient for intelligence? Are they necessary? Do we need anything else? If so, what? AI research meshes with some of the most exciting philosophical questions. It also offers tremendous scientific and technological challenges.

The state of AI as a scientific discipline today is much like that of physics in Newton's time. The scientific problems of AI and cognitive science are embarassingly simple to state in everyday language (e.g., "How do we recognize everyday objects?" or "How do we learn and use language?" or "How could an AI professor have evolved from a primordial soup of chemicals?") much like some of the questions that Newton sought to answer (e.g., "Why does an apple released from a tree fall to the ground?"). Are there simple satisfactory answers to the simply stated questions of AI?

The rise and fall of fashions and "hot topics" in contemporary science has probably to do as much with sociology and economics of science as with scientific progress. AI has had its share of hot topics – from the good-old-fashioned-AI and expert systems of the 1960s and 1970s to artificial neural networks and genetic

algorithms of the 1980s and intelligent agents and distributed AI of the 1990s, machine learning and semantic web of the 2000s, big data and deep learning of the 2010s.

Despite, or perhaps even because of, occasional false starts, research in each of these areas has led to, and continues to provide, important insights, and real (albeit modest) successes, that could be stepping stones to a fundamental understanding of the origins and workings of intelligent systems (including ourselves, our societies, and cultures — in short, the very essence of our existence and being). That is what makes AI (broadly interpreted), one of the most important, exciting, and challenging scientific, intellectual, and technological developments of the twentieth century.

Acknowledgements

Some of the material in this chapter is adapted from:

1. Honavar, V. Symbolic Artificial Intelligence and Numeric Artificial Neural Networks: Toward a Resolution of the Dichotomy. In: *Computational Architectures Integrating Symbolic and Neural Processes*. Sun, R. & Bookman, L. (Ed.) New York, NY: Kluwer, 1994.

2. Uhr, L. & Honavar, V. Artificial Intelligence and Neural Networks: Steps Toward Principled Integration. In: *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*. Honavar, V. & Uhr, L. (Ed). New York, NY: Academic Press, 1994.

I am indebted to Professor Leonard Uhr of the University of Wisconsin-Madison, my mentor and friend, for introducing me to research in artificial intelligence and cognitive science and for countless stimulating discussions on the topic. His work in Artificial Intelligence dating back to the 1950s with its emphasis on learning and adaptation, agents and emergence of intelligent behavior, situated perception and action, and numeric/symbolic hybrid systems foreshadowed the work that became part of the mainstream many years later. This has, and will continue to, even after his untimely demise in 2000, inspire those of us who have had the good fortune of having known him.

I am grateful to members of the artificial intelligence research laboratories at the University of Wisconsin-Madison (during 1985 through 1990) and Iowa State University (from 1990 to 2013), and the Pennsylvania State University (from 2013 onwards) for many useful discussions on AI and cognitive science. I have also benefited from my interactions with faculty and students of the Center for Neural Basis of Cognition at Carnegie Mellon University (during my sabbatical there in 1998), and of the University of Wisconsin Machine Learning Research Group (during my sabbatical there in 2002), and the staff of the National Science Foundation and the larger research community (during my service at the Foundation during 2010-2013).

I am grateful to the United States National Science Foundation, the National Institutes of Health, the US Department of Agriculture, the Department of Defense, Iowa State University, the Indian Institute of Science, and the Pennsylvania State University for support of my research in AI.