L. Sweeney, *That's AI?*, Carnegie Mellon University, School of Computer Science, Technical Report, CMU-CS-03-106. Pittsburgh:January 2003.

That's AI?

A History and Critique of the Field

Latanya Sweeney
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
latanya@cs.cmu.edu

ABSTRACT

What many AI researchers do when they say they are doing AI contradicts what some AI researchers say is AI. Surveys of leading AI textbooks demonstrate a lack of a generally accepted historical record. These surveys also show AI researchers as primarily concerned with prescribing ideal mathematical behaviors into computers -- accounting for 987 of 996 (or 99%) of the AI references surveyed. The most common expectation of AI concerns constructing machines that behave like humans, yet only 27 of 996, (or 2%) of the AI references surveyed were directly consistent with this description. Both approaches have shortcomings – prescribing superior behavior into machines fails to scale to multiple tasks easily, while on the other hand, modeling human behaviors in machines can give results that are not always correct or fast. The discrepancy between the kind of work conducted in AI and the kind of work expected from AI cripples the ability to measure progress in the field.

Keywords: history of AI, philosophy of AI, human intelligence, machine intelligence, Turing Test, reasoning

INTRODUCTION

More than 50 years have passed since the origin of artificial intelligence (AI). Recently, Warner Brothers and DreamWorks released Steven Spielberg's film, "A.I.", thereby advancing the state of science fiction from "E.T." [Spielberg 1982], which is a story about an extra terrestrial who comes to earth and befriends a boy, to "A.I" [Spielberg 2001], which is a story about a human-like robot that believes he is a boy. Inspiring a movie is an accomplishment, but this success begs the question, "where is the real AI?" In 1950, Turing presented his dream of a machine that would behave like a human. He predicted it would be realized by the year 2000 [Turing 1950]. Originally, the term artificial intelligence was used exclusively in the sense of Turing's dream [Wilkes 1992], but very little progress has been made along those lines. Perhaps now is a natural time for AI to re-examine itself – reflect on its past, take stock in its present and plot directions for its future. This work provides one such examination.

The meaning of AI today of course depends on whom you ask. Perhaps for an individual researcher this is not a problem because she can guide her work by a personally adopted definition. But if a field has too many different and sometimes conflicted meanings, as is the case in AI, then problems emerge when measuring progress and building on previous work. Such problems are not merely the result of different researchers seeking to achieve a common goal but doing so by following diverse research paths. Instead, it seems no unified vision of AI has emerged or been widely adopted; and as a result, progress made under one characterization of AI is not be viewed as success by others who operate under a different perception of it. This tends to undermine confidence in the entire field itself, promote premature conclusions of what can and cannot be accomplished and limit progress and funding along research paths. In fact, Hayes-Roth [1997] reports that traditional AI funding sources themselves face serious

budget constraints. So, they now continually ask hard and pinpointed accountability questions about dollars they invest in AI research. They want to see demonstrated progress in AI. But what are the operational goals against which progress is to be measured in the absence of a common vision of AI?

BACKGROUND

In covert and overt observations and in informal surveys involving 30 computer science graduate students (15 at MIT, 5 at Harvard, 5 at CMU and 5 at Stanford), overwhelming discontent with AI and expressions of its demise were found. Granted these are anecdotal findings; but surprisingly, virtually no attention has been paid to these matters in the AI literature.

Some references heralded the success of AI with no self-criticisms of the field along the lines of an overall objective [Wegner and Doyle 1996; Reddy 1996]. A few exceptions were found. Brooks [1991] declared that AI researchers do not talk about replicating the full power of human intelligence anymore. Wilkes [1992] foretells that Turing's dream cannot be realized by digital computers. Doyle [1996] warned that unless AI redefines itself, it might face "disrespect, decay and dissolution." The absence of self-analysis with respect to an overall agenda leaves a sense that the field is no longer concerned with such matters. But with no common vision, in what direction is AI headed? This paper reports on surveys that were conducted to identify stated goals of AI and to relate those goals to the character of research promoted in AI. The results reveal mismatches between the overall expectations of AI and the work actually conducted.

METHODS AND RESULTS

In an attempt to describe AI's goals and its work, 3 surveys are presented. The first survey categorizes definitions of AI found in leading AI textbooks in an attempt to identify what is considered AI. The second survey classifies almost 1000 AI references, which are cited in a leading AI textbook, to describe the kind of work promoted as AI. The final survey examines definitions of AI found in computer science and basic references to consider what others depict AI to be. The categorizations from the first survey (based on AI textbook definitions) do not adequately support the findings in the second survey (based on AI cited work), and the results from the second survey (based on classifications of cited work in AI) and the third survey (based on definitions of AI by others) are opposed. The remainder of this section describes the methods used and the results found.

Terms like machine and hardwiring have precise meanings in this work. All pursuits of AI involve the construction of a *machine*, where a machine may be a robot, a computer, a program or a system of machines whose essence these days is assumed to be rooted in digital computer technology (though biological machines or combined biological and digital machines may be possible in the future [Knight and Sussman 1997]). The construction of a machine requires *hardwiring*, which is the knowledge, expertise or know-how that is incorporated a priori into the machine. While self-refinement within the machine is possible such as modifying internal state, adjusting parameters, updating data structures, or even modifying its own control structure, hardwiring concerns the construction of the initial machine itself. Machines are hardwired to conduct one or more tasks.

The first two surveys rely heavily on Russell and Norvig's textbook on artificial intelligence [1995]. While this decision alone generates particular biases, Russell and Norvig's textbook may be considered the leading textbook on artificial intelligence in terms of general coverage of the field. Reportedly, the book has been adopted in 561 courses at 463 schools [Russell and Norvig 1999]. Thirty of the top 40

graduate programs in computer science are noted as using the book [Russell and Norvig 1999]. Over 39,000 copies have been sold [Russell and Norvig 1999].

Survey of AI Textbooks

Russell and Norvig [1995] surveyed eight contemporary artificial intelligence textbooks [Haugeland 1985; Bellman 1978; Charniak and McDermott 1985; Winston 1992; Kurzweil 1990; Rich and Knight 1991; Schalkoff 1990; and, Luger and Stubblefield 1993] and organized the definitions of artificial intelligence found within them into the four categories shown in Figure 1 and labeled AI_{HT}, AI_{IT}, AI_{HB} and AI_{IB}. The categories in the left column (AI_{HT} and AI_{HB}) measure success in terms of human performance while those in the right column (AI_{IT} and AI_{IB}) compare the machine's performance to an ideal mathematical standard. The categories in the top row (AI_{HT} and AI_{IT}) compare the machine's performance in thinking and reasoning skills while those on the bottom row (AI_{HB} and AI_{IB}) compare the machine's performance in behavioral terms.

	Human compliance	**Ideal compliance**
Thinking & reasoning	Machines that think like humans. 2 AI_{HT}	Machines that think rationally. 2 AI_{IT}
Behavior	Machines that act like humans. 2 AI_{HB}	Machines that act rationally. 2 AI_{IB}

Figure 1. Definitions of AI divided into four categories [Russell and Norvig 1995].

Each cell in the grid shown in Figure 1 represents the definitions from two leading AI textbooks. The implication is that AI work appears somewhat evenly distributed over these categories as noted in definition 1.

Definition 1. *(artificial intelligence with respect to AI textbooks, $AI_{textbooks}$) Artificial intelligence is the study of ideas to bring into being machines that perform behavior or thinking tasks ideally or like humans.*

Definition 1 defines $AI_{textbooks}$ as encompassing the AI_{HT}, AI_{HB}, AI_{IT} and AI_{IB} definitions, but there are biases. While the 8 books that provided these definitions, are "leading" AI textbooks, the percentage of students taught with each is not known. Similarly, the total market share accounted for by these 8 textbooks is not known. A better metric might be to weigh each definition by the number of students potentially influenced by its definition. There may be more or different textbooks needed. Nevertheless, the authors of these "leading" AI textbooks are themselves noted AI researchers, so this survey gives a reasonable starting point for characterizing how those in the field define AI.

The next survey shows that work in AI is far from being evenly distributed across these definitions. Also, the definitions are not inclusive of work done by Brooks [1995] and his colleagues; a column named "animal compliance" must be added; and so, in the remainder of this paper, such is included.

Survey of AI References

In an attempt to characterize works that are promoted as notable AI research, references found in an AI textbook were classified. In particular, there are 1374 references to scientific work in Russell and Norvig's bibliography [1995]. Of these, 378 are not based on AI research, but are related references in psychology, logic, philosophy or other fields or could not be located. The remaining 996 were reviewed

and classified into the following classes: (1) human thinking; (2) human behavior; (3) ideal thinking; (4) ideal behavior; and, (5) animal behavior. Each class is inspired by a definition of AI found in Figure 1. Of the 996 references, 987 (or 99%) were classified as being compliant to ideal thinking or ideal behavior. In the next paragraphs, the criteria for classification and detailed findings are provided.

Before continuing, however, the term *work* (as a noun) needs to be clarified. A *work* refers to the thesis that is the subject of a bibliographic reference as it relates to the construction of one or more machines. Each of the 996 references related to one or more machines, so each of the 996 works was classified based on the machine(s) that were the subject of its thesis. A work can therefore fit into one or more classes.

Definition 2. ***(human thinking class, HT)*** *The human thinking class, HT, is the set of works whose machines pose or employ a model of how humans reason when humans perform the same tasks.*

Definition 2 describes the human thinking class (HT) as consisting of works that seek to perform tasks by modeling how humans might reason while doing those same tasks.

Example (human thinking class).

An example of a work in the human thinking class is General Problem Solver [Newell and Simon 1961]. This program attempts to solve problems by following the same reasoning steps as humans. Problems are solved by successive decomposition of a goal into sub-goals and by establishing new goals based on differences between current and desired states.

Of the 996 references, only 22 (or 2%) were works that modeled human thinking.

Definition 3. ***(human behavior class, HB)*** *The human behavior class, HB, is the set of works that account for human phenomena found when humans perform the same tasks.*

Definition 3 describes the human behavior class (HB) as consisting of works that attempt to behave like humans, but unlike the human thinking class, works that model human behavior do not have to necessarily model the human reasoning process. The human behavior class includes works that do not merely perform tasks which humans perform, but works which do so by imitating or duplicating the behavioral patterns, standards or paradigms observed in the human performance of those tasks.

Example (human behavior class).

The references had few examples of works in the human behavior class. Looking beyond the references for examples is the Scrub system [Sweeney 1996], which reliably locates personally identifying information in unrestricted text. Humans seem to use localized knowledge when asked to quickly recognize names, phone numbers and other personal identifiers within letters, notes and messages, and to locate these identifiers easily without fully reading the contents of the text itself. Modeling this human text scanning approach involved numerous detection algorithms competing in parallel and sharing information through a blackboard-like architecture. No claim is made that such processing happens within the minds of humans. Yet, the resulting system behaved comparable to humans, finding 99-100% of all personal identifiers while the leading techniques used at the time found only 30-60%.

Of the 996 references, only 5 (or 1%) were classified as being works on human behavior that did not also model human thinking. These included ELIZA [Weizenbaum 1966], which could engage in conversation on any topic and performed such using linguistic pattern matching.

Consider the human thinking class (HT) and the human behavior class (HB). Works that model human thinking and reasoning are a subset of works that model human behavior. After all, humans are basically black boxes that can be observed, surveyed and experimented on, but products of human thinking are evidenced only by human behavior. This is illustrated in the set notation shown in Figure 2, in which works modeled after "human thinking" are a subset of works modeled after "human behavior." That is, all works in the human thinking class are works in the human behavior class, but the reverse is not necessarily true; this can be expressed as $HT \subseteq HB$. In the classification of the 996 references, only 5 references concerned works that modeled human behavior, exclusive of modeling human thinking; that is, $|HB - HT| = 5$.

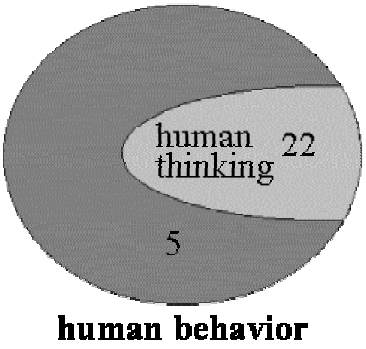

Figure 2 References compliant to human behavior.

Definition 4. ***(ideal thinking class, IT)*** *The ideal thinking class, IT, is the set of works that provide only rational thinking or "right thinking" when performing tasks humans do.*

Definition 4 describes the ideal thinking class (IT) as containing works that provide rational thought. These works seek to find the best or optimal answer of all possibilities, even in situations where humans do not typically provide the best or optimal answer.

Example (ideal thinking class).

An example of a work in the ideal thinking class is the very large knowledge base and inference engine known as CYC [Lenat 1998], which has been under development since 1984. The intention of CYC is to construct (by hardwiring) a repository of commonsense knowledge so that programs accessing CYC can be more flexible and useful because they will have access to information about common objects, events and relationships. All the knowledge in CYC is represented declaratively in the form of logical assertions. The inference engine derives new conclusions using deductive reasoning. No numerical methods such as statistical probabilities are used. Likewise, CYC does not utilize neural networks or fuzzy logic. By 1997, CYC contained over 1.5 million facts, rules-of-thumb and heuristics the developers considered necessary for reasoning about some of the objects and events of everyday life. The CYC knowledge base is still growing.

CYC is not a work in the human thinking class. It does not make any attempt to understand how the human mind works, or test any particular theory of intelligence. CYC does not claim to model human behavior either. To make a decision, CYC generates all assertions consistent with its knowledge base and then logically examines them (using deduction) in order to reach conclusions.

Of the 996 references, 769 were works on rational thinking, as shown in Figure 3. References involving logical reasoning systems dominated this class. Such systems derive mathematically sound conclusions from formal declarative knowledge and are usually defined by abstract rules of inference. Reasoning strategies utilized in these works include causal, deductive and probabilistic approaches.

Definition 5. ***(ideal behavior class, IB)*** *The ideal behavior class, IB, is the set of works that provide only rational actions or "right actions" when performing tasks humans do.*

Definition 5 describes the ideal behavior class (IB) as consisting of works that provide rational actions. Examples of works found in this class that are not also in the ideal thinking class include works based on functional components such as perception, learning and planning in robots and robotic vehicles. Other examples include game playing such as chess, checkers, backgammon and Go.

Example (ideal behavior class).

Of all games to play, chess, in particular, became a goal for some in AI because it was viewed as a masterful game of wits. In 1957, Herbert Simon predicted that AI would produce a world chess champion in 10 years [Simon 1958]. Some coined this "the Chess Test." In 1997, forty years later, the IBM computer Deep Blue defeated the world chess champion Garry Kasparov. IBM's web pages devoted to the competition at that time claimed that Deep Blue did not use AI. The accuracy of that claim depends on one's meaning of AI. Deep Blue substituted overwhelming computer power for modeling human thought.

Human chess players cannot examine all moves at every position; so instead, they focus on promising moves for exploration. In comparison, Deep Blue was capable of evaluating 100 million chess positions per second [Newborn 1996]. It used a 32 node IBM RS/6000 parallel computer, where each node had six specially designed chess processors. Human chess players have to abstractly analyze board positions and integrate learned opening and ending strategies. Deep Blue used stored histories of games and crude computer power to make up for its inefficiency of analysis. It had no learning whatsoever.

Early work in AI on chess programs did attempt to more closely model human chess playing, but these programs became extremely complicated. A common approach at the time involved arranging all possible moves in a search tree and then deploying algorithms for selecting the best move by forward pruning the tree. These approaches provided mediocre play, and were later abandoned in favor of simpler approaches relying on speed and storage improvements in computer hardware.

Of the 996 references to classify, 192 of them concerned ideal behavior, exclusive of ideal thinking; this can be expressed as $|IB - IT| = 192$. References specific to the ideal behavior class and not included in the ideal thinking class typically involve machines that exhaust or reasonably exhaust the space of all possibilities to determine which action is best. Such direct computations are usually beyond human ability.

In these discussions of the ideal thinking class (IT) and the ideal behavior class (IB), the word "ideal" has been used not to imply that such machines never make any error whatsoever. That is not correct. Machines that comply with an ideal standard always seek to provide the right or rational answer, where "right" is determined from the perspective of omniscience. However, in the face of uncertainty, the ultimate right or rational answer may not be decidable with absolute confidence. In these cases, what is right is qualified but such machines still seek to get the right answers regardless of the kinds of answers humans may provide.

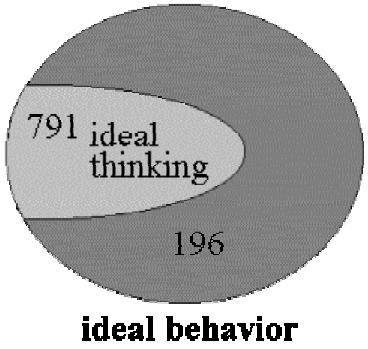

Figure 3 References compliant to ideal (or rational) behavior.

Works that model ideal (or rational) thinking and therefore are in the ideal thinking class (IT) are a subset of works that model ideal (or rational) behaviors; that is, IT ⊆ IB. This is illustrated in the set notation shown in Figure 3. Of the 996 references, 987 (or 99%) are works in the ideal behavior class, and of these, 791 (or 79%) are works in the ideal thinking class.

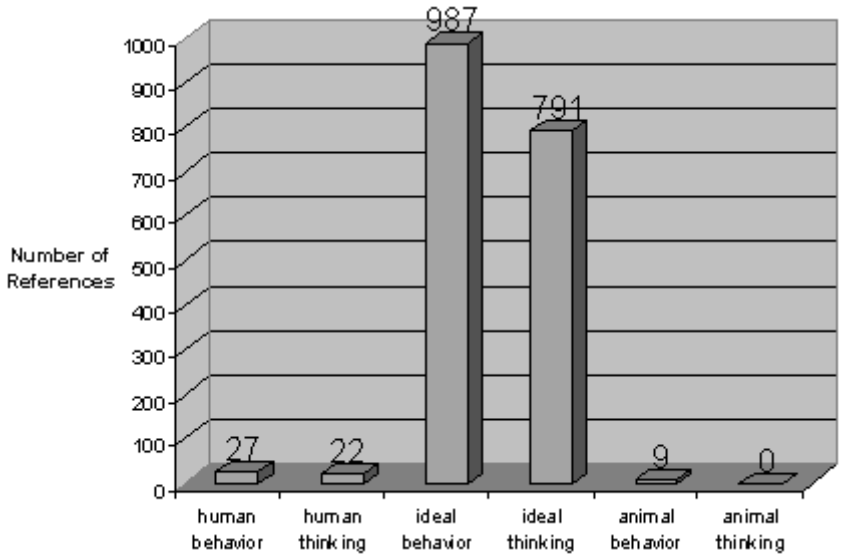

Figure 4 Number of references of AI work in an AI textbook by class.
One reference appears in both the animal behavior and human behavior classes.

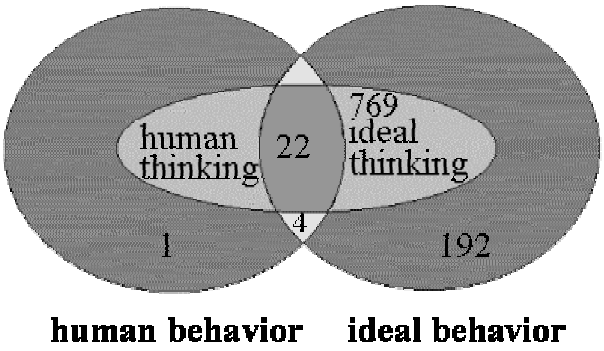

Figure 5 References compliant to human and ideal classes.

Figure 4 reports the findings for each class. Figure 5 shows the complete distribution of references realized by intersecting the findings reported in Figure 2 and Figure 3. This distribution is also itemized in Figure 6. As shown in Figure 5, sometimes humans think and act rationally. Therefore, there are intersections between these classes. Surprisingly however, the only works that appear in that intersection are those that appeared in human thinking and human behavior (excluding ELIZA, the fifth, because it is not considered rational).

	Compliant to Humans	Compliant to Ideal	Compliant to Animals
Thinking & reasoning	22 (2%)	791 (79%)	0
Behavior (excluding thinking)	5 (1%)	196 (20%)	9 (1%)

Figure 6 Classification of 996 AI references.
One reference appears in both the animal behavior and human behavior classes.

Definition 6. ***(animal behavior class, AB)*** *The animal behavior class, AB, is the set of works that account for phenomena observed when insects or animals perform the same tasks.*

Definition 6 describes the class of works on animal behavior (AB) as those that model behaviors observed in insects and animals. Examples include works in behavior-based robotics [Brooks 1995], which is an approach to robot design that advocates the construction of autonomous robots having very little (or no) internal state. Typical behaviors are wandering, avoiding obstacles, wall following, delivering an object, cleaning the floor, and following someone. These behaviors are also observed in humans, but there are animal behaviors not shared with humans. The animal behavior class includes machines that model behaviors observed in both humans and animals as well as behaviors observed only in animals.

Example (animal behavior class).

To better understand the animal behavior class, the behavior-based approach to robotics is compared to the more traditional AI approach to robot design. First, for clarity some terms are introduced. Works on robot design that were classified in the ideal behavior class are termed *ideal robots*. Similarly, works on robot design that were classified as animal behavior are termed *animal robots*.

Here then is a comparison of ideal robots and behavior-based animal robots. Ideal robots are deliberate. Behavior-based animal robots react to their environment. Behavior-based animal robots have little or no internal representation of the specific world in which they operate, where as ideal robots have a symbolic representation of their world. Behavior-based animal robots tend to be robust and fault tolerant and operate in real-time. Ideal robots tend to be slower, following the computationally intensive algorithm: (1) perceive world; (2) update internal world model; (3) make a plan; (4) take action; and then, (5) go to step 1. Behavior-based animal robots directly link sensors for perceiving the world to actuators for taking action with no substantive internal processing steps in the middle.

So far, behavior-based animal robots have demonstrated low-level intelligence, where ideal robots have been aimed at high-level intelligence. Some wall-following behavior-based animal robots have been likened to the meandering foraging trails left by animals of the late Precambrian and early Cambrian periods approximately 530-565 million years ago [Prescott and Ibbotson 1997]. On the other hand, Minerva, an ideal robot that talks and moves among people in public spaces, recently debuted in the Smithsonian's National Museum of American History. It interacted with an estimated 100,000 people, actively approaching people, offering tours and then leading visitors from exhibit to exhibit [Thrun et al. 2000].

This sharp contrast may seem somewhat definitive, but such is far from reality. A behavior-based animal wall-follower robot can operate in many different settings with no changes to its hardwiring whatsoever. Minerva, on the other hand, can only work properly in the Smithsonian during the showing of those particular exhibits. Change one exhibit or the floor plan and Minerva's hardwiring has to change. Many difficult challenges and open questions remain to be solved before robots like Minerva become robust, adaptive and fault tolerant enough to operate in many different settings with no changes to their hardwiring. On the other hand, the commercial availability of robotic pets, which interact with humans for human amusement, shows evolution in animal robots beyond the behavior-based creatures described above.

Robots enjoy a special place in AI, related in part to the human tendency to personify them. We often project human qualities onto robots when interpreting what we consider a robot to "think" or "feel." For example, consider a robot with a light sensor that is hardwired to move away from light. When a light beam from a flashlight hits its sensor, it scurries away. It has been observed that humans often describe the behavior in human terms, such as "the robot is shy."

Of the 996 references, only 9 of them are works in animal behavior (AB) and all 9 of them are works in behavior-based robotics. Eight of these 9 references are somewhat specific to animal behavior (AB), exclusive of human behavior (HB). The remaining work models the behaviors shared with humans that were described above, so the work is included in both the animal behavior class (AB) and the human behavior class (HB).

This work began with classifications established by Russell and Norvig's division of AI definitions found in leading AI textbooks [1995]. The set of classes was expanded to include animal behavior. Finally, Figure 6 contains a summary of the results from the classification of 996 references; 987 (99%) were classified as being work on ideal behavior or thinking. This is very different from the even distribution implied by the results shown in Figure 1. These results lead to the definition of AI stated in definition 7 and referred to as $AI_{references}$.

Definition 7. ***(artificial intelligence with respect to AI references, $AI_{references}$)*** *Artificial intelligence is the study of ideas to bring into being machines that perform tasks typically done by humans but that performs them ideally.*

The classes HB, HT, IB and IT are directly related to the definitions of AI provided as AI_{HB}, AI_{HT}, AI_{IB}, AI_{IT} in Figure 1. Let w_i be a work. If the work w_i is in the HB class (i.e., $w_i \in HB$), then it is said w_i "complies" with AI_{HB}. If $w_i \in HT$, then w_i complies with AI_{HT} and AI_{HB}. Similarly, if $w_i \in IB$, then w_i complies with AI_{IB}. And finally, if $w_i \in IT$, then w_i complies with AI_{IT} and AI_{IB}.

It is clear from the classification of the references reported in Figure 6, that if a work, w_j, appears as the thesis of an AI reference in Russell and Norvig's textbook, it is highly likely $w_j \in IB$. Specifically, the probability $Pr(w_j \in IB) = 99\%$. So then, it is highly likely that w_j complies with AI_{IB}. Therefore, it is not surprising that $AI_{references}$, which is the definition of AI inspired by the classification of the references, is the same as AI_{IB}.

Validating and Generalizing these Findings

Bibliographical references found in a leading textbook are a good basis for this kind of analysis because they reflect works that are presented as being representative of or important to AI, and as such, are being promoted to the next generation of researchers. Nevertheless, there are biases resulting from the way in which they are used in this study. First, the classes were determined a priori and there may exist another set of classes that better describe these references. However, the classes used were determined from surveying definitions found in leading AI textbooks so they are believed to be sufficient.

A second bias concerns the selection of references. The references that were classified in the previous experiment reflect works in AI that the authors of a leading AI textbook, Russell and Norvig, determined as important and interesting. There are other ways to survey AI research such as money or time spent or using citations in AI research journals. However, most of the research done in AI or reported in the field's noted journals is sponsored by a funding agency, thereby making it unlikely that surveying AI research dollars spent or papers appearing in AI research journals would provide drastically different results. As a point to note, of the 1374 references found in the bibliography of Russell and Norvig's textbook, 532 (or 39%) are from books, 517 (or 38%) are from journal articles, 232 (or 17%) are from conference proceedings and 93 (or 7%) are from technical reports, theses and unpublished manuscripts. No significant changes in the character of the earlier findings were found by classifying references respecting these type categories.

A possible bias concerns what may be the changing focus of AI over time. Earlier definitions of AI were rooted in human performance, consistent with Turing's dream, where as later definitions of AI involve achieving an ideal standard. Therefore, one would suspect definitions of AI described as human behavior (AI_{HB}) and human thinking (AI_{HT}) would appear in the earlier textbooks, where as the definitions of AI described as ideal behavior (AI_{IB}) and ideal thinking (AI_{IT}) would appear in the more recent textbooks. The textbooks surveyed by Russell and Norvig (see Figure 1) do not completely support this suspicion. Figure 7 displays each of the 8 textbooks surveyed, along with its publication year and the classification of the definition of AI found within. While the two oldest textbooks have definitions of AI concerned with human thinking (AI_{HT}) and the two most recent textbooks have definitions of AI concerned with ideal standards (AI_{IT} and AI_{IB}), we cannot confidently conclude from this small sample that the definition of AI, as reported in contemporary AI textbooks, shifted.

Author(s)	Publication Year	AI Definition Class
Bellman	1978	human thinking AI_{HT}
Haugland	1985	human thinking AI_{HT}
Charniak & McDermott	1985	ideal thinking AI_{IT}
Kurzweil	1990	human behavior AI_{HB}
Schalkoff	1990	ideal behavior AI_{IB}
Rich & Knight	1991	human behavior AI_{HB}
Winston	1992	ideal thinking AI_{IT}
Luger & Stubblefield	1993	ideal behavior AI_{IB}

Figure 7 AI definitions from AI textbooks

There may be bias stemming from the publication year of the textbook. Perhaps classifying references from a leading AI textbook published earlier would produce radically different findings. Figure 8 plots the references found in the bibliography of Russell and Norvig's book by publication year. Notice that most of the references were recent to its publication date (1995). Bellman's textbook, in contrast, was published in 1978. As a result, Bellman's textbook could share no more than 38% of Russell and Norvig's references. Similarly, Haugland's textbook was published in 1985, thereby possibly sharing no more than 54% of Russell and Norvig's references. As a final observation, Winston's textbook was published in 1992 and could possibly share 91% of the references found in Russell and Norvig's bibliography. These books provide a sample from each of the last three decades.

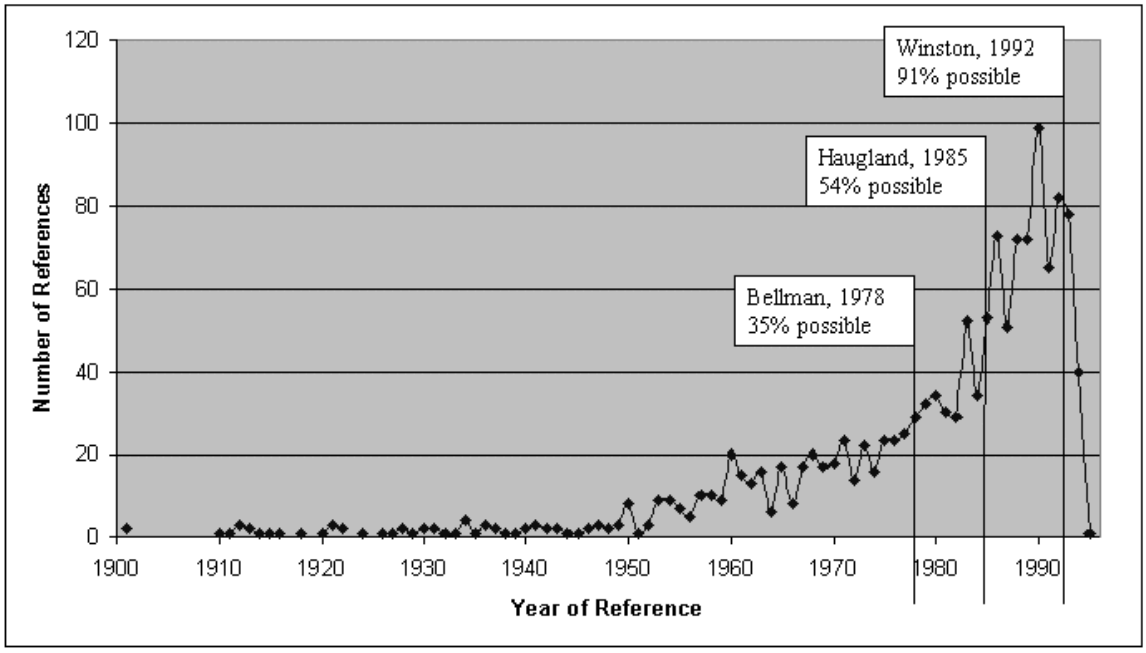

Figure 8 Bibliographic counts (by year) of references from Russell and Norvig, 1995

Despite the possibility that occurrences of the same reference could appear in different textbooks, few references were in fact found in common. Let RN_{78}, RN_{85} and RN_{92} be the sets of references from Russell and Norvig's bibliography whose year of publication is less than or equal to 1978, 1985, and 1992, respectively. Figure 9 reports the number of references found in each set. RN_{78} has 477 references. Bellman's textbook published in 1978 has 182 references (the set of which is represented as

B). Only 11 ($|RN_{78} \cap B| = 11$) references are common in both books. RN_{85} has 741 references. Haugland's textbook published in 1985 has 122 references (the set of which is represented as H). Only 33 ($|RN_{85} \cap H| = 33$) are common in both books. Finally, RN_{92} has 1255 references. The 1992 edition of Winston's textbook (represented as W) has 533 references. Only 110 ($|RN_{92} \cap W| = 110$) are common in both books.

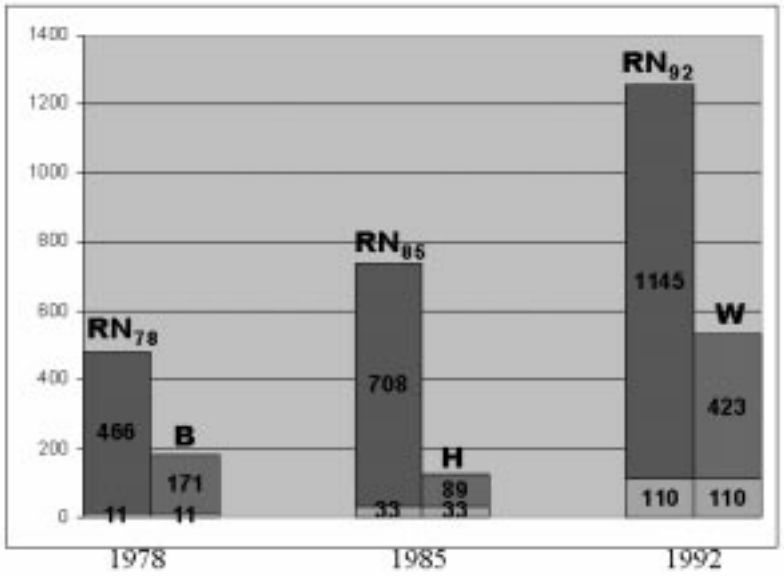

Figure 9 Bibliographic comparison of references from leading AI textbooks: Russell Norvig (RN), Bellman (B), Haugland (H) and Winston (W). The number of common references appears on the bottom of each bar.

It is difficult to report the extent to which the results reported in Figure 9 reflect a selection bias by the authors. A specific machine or work can have more than one reference and duplications based on particular machines or works were not removed. Likewise, Russell and Norvig may have included references that were extensions of earlier machines or works without reporting the earlier references because they would then be outdated. Further, selection bias by the authors based on source of publication was evident. Almost 1/3 (51 of 182 or 28%) of Bellman's references are from mathematical sources; yet, none of the references common in both B and RN_{78} are from mathematical sources. Of the 33 references found in both H and RN_{85}, most (21 or 64%) are books. But of the 110 references common in W and RN_{92}, most (42 or 38%) are from papers published in *Artificial Intelligence* (the academic journal). In order to combat some of these issues, the distinct authors found in the bibliographies of each of these textbooks were examined.

Figure 10, Figure 11, and Figure 12 show the number of common authors found in all possible 2-way, 3-way and 4-way combinations of these textbooks. In all cases, surprisingly few authors are found in common and there does not appear to be any other remarkable characteristic that stands out. In the following examples, let B_A, H_A, W_A, and RN_A be the set of authors found in the bibliographies of Bellman's, Haugland's, Winston's and Russell and Norvig's textbooks respectively. Consider the time between publication years. W_A and RN_A are published only 3 years apart and 31% of W_A's authors are found in RN_A. Likewise, B_A and RN_A are published 17 years apart and only 7% of B_A's authors are found in RN_A. These findings imply that the greater the time between publication years, the fewer the number of authors found in common. Sounds good, but this implication doesn't hold. H_A and RN_A are published 10 years apart and 27% of H_A's authors are found in RN_A, while B_A and H_A are published only

7 years apart and only 9% of H_A's authors are found in B_A notwithstanding the fact that H_A and B_A both define AI as AI_{HT} (human thinking).

The findings reported in Figure 9, Figure 10, Figure 11, and Figure 12 are disturbing. They show that a significant percentage of references and authors do not appear in multiple AI textbooks, even when those textbooks are published within a few years of each other. There are several possible reasons. Perhaps few works (or authors) are fundamentally important to the field so far. If so, many of the works being published do not seem to be universally accepted as progress from within the field itself. Some of this is expected as new ideas and directions are being promoted rather than established first principles. After all, at the time of publication, an author cannot predict whether recent efforts will yield a fruitful research path. Nevertheless, it is important to note that different textbook authors are promoting different works.

What can be said about so many cited works appearing in one leading textbook but not others published in the same time period? Are they redundant works done by different authors? If so, improvements are needed to report findings in such a way that researchers can build on previous works and not waste valuable resources replicating efforts. Perhaps the field is fragmented into too many sub-fields. References can be drawn (or not) based on the author's knowledge or promotion of a particular sub-field within AI. If so, a common vision for AI and a supporting framework is needed to make sure results are generalized and reported across sub-fields.

In order to quantify some of these concerns, the set of references from Winston's textbook (W), which was published just three years prior to Russell and Norvig's (RN) and the set of references from authors common in both W and RN, were also classified into the human behavior, human thinking, ideal behavior and ideal thinking classes. The results revealed a 10-20% change in some of the classes, but such change did not alter the essence of the findings reported earlier. $AI_{references}$ describes works in AI as being members of the ideal behavior class IB.

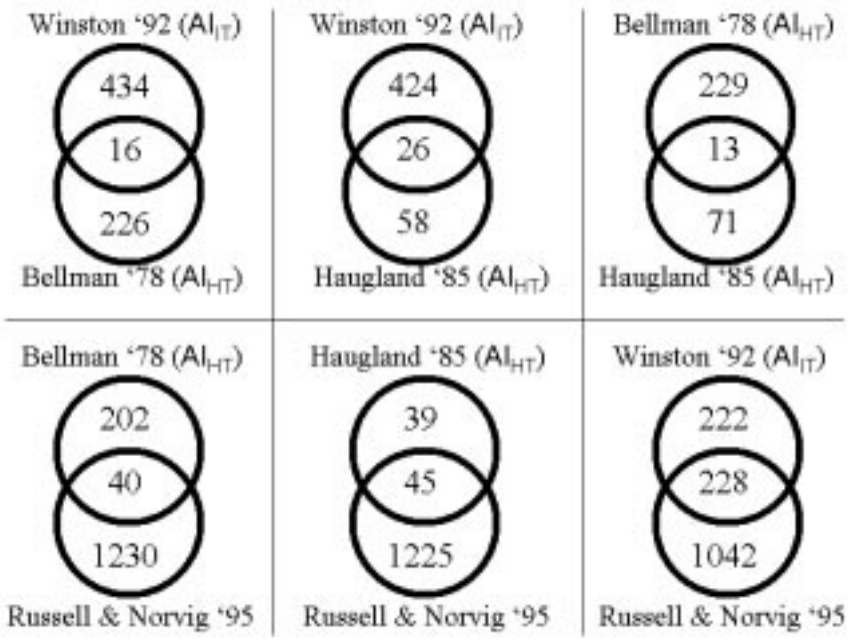

Figure 10 Two-way intersections showing the number of authors in references from AI textbooks

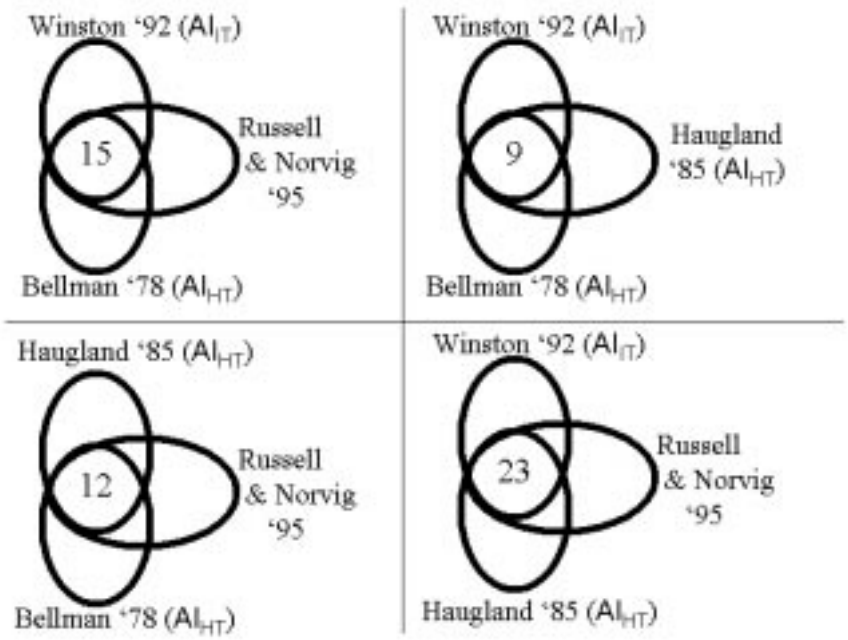

Figure 11 Intersections (3-way) showing the number of authors in references common in some AI textbooks

Author
Edward A. Feigenbaum
Marvin Minsky
Allen Newell
Zenon W. Pylyshyn
Roger C. Schank
Herbert A. Simon
Joseph Weizenbaum
Terry Winograd

Figure 12 List of authors found in all 4 of the AI textbooks surveyed

Survey of General References.

Here is a summary of what has been reported so far. First, definitions of AI found in AI textbooks were characterized to report how the AI community defines AI; see Figure 1. Then, AI references were classified to see how actual AI research compared to AI definitions; see Figure 6. In summary, the textbook definitions described AI as being more diverse and encompassing than its referenced research revealed.

Now, as a final characterization, the way in which others perceive AI is examined. Ten computer science and general references were surveyed to determine how they defined AI. A list of these definitions and sources is contained in Appendix 1. The results are shown in Figure 13. Nine (90%) of the definitions describe AI as being work that is compliant to humans.

Definitions of AI from outside the field (see definition 8) are opposite to the classification of actual AI work in the field. This implies that outside expectations of AI are counter to the work actually performed. It is not uncommon for lay definitions to be obtuse, but this problem is different in some

respects. The surveyed references are from computer science and general references, so they are not completely removed from the AI community. The difference may result from historical lag. That is, AI was first defined in the spirit of Turing's dream even though AI work has clearly followed a different model. Even still, half the AI textbooks surveyed (4 of 8 books) agreed with these outside definitions. So these findings may be more of a sign of confusion within the AI community itself.

Definition 8. ***(artificial intelligence as defined by others, $AI_{general}$)*** *Artificial intelligence is the study of ideas to bring into being machines that simulates human thinking and behaving.*

	Compliant to Humans	Compliant to Ideal	Compliant to Animals
Thinking& reasoning	6	1	0
Behavior	3	0	0

Figure 13 Survey from computer science and general references.

The results reported in Figure 13 have some bias and the biggest concern stems from problems with small number statistics. Also, the references are not qualified in terms of how much impact or exposure they have. Nevertheless, the contrast in these results and those of Figure 6 remains quite interesting.

In summary, the following probabilistic implications were found.

$$AI_{textbooks} \Rightarrow AI_{IB} \cup AI_{HB}$$
$$AI_{references} \Rightarrow AI_{IB}$$
$$AI_{general} \Rightarrow AI_{HB}$$

The surveys reported in this paper show that AI has not been consistently defined. While no commonly accepted definition of AI was found, the definitions of AI as being AI_{HB}, AI_{HT}, AI_{IT} or AI_{IB} were found with equal emphasis in AI textbooks. The work promoted in the field has been overwhelmingly limited to AI_{IB}. Yet, general references tend to consider AI as AI_{HB}.

In the next section, which is the final section, these findings on definitions of AI are related to how the field appears to be evolving.

DISCUSSION

AI began as a pursuit to construct a machine that would behave like a human (AI_{HB}). AI has become a pursuit to construct a machine that adheres to an ideal standard (AI_{IB}). These are not the same. So, questions remain. Why did the definition of AI for AI researchers go from modeling human behavior to ideal behavior and will the focus switch back? How have works in AI evolved? What development paths are works in AI following? What are the driving forces? Discussion in this section attempts to shed some light on these questions by contrasting definitions of AI, which were realized in the previous section, to related economic and technical realities.

Beforehand, clarification is needed regarding what is meant by *success*, *ideal machines* and *human-like machines*. A machine is *successful* if it is deployed and in regular use. One machine is more successful than another if its use is more widespread. The term *ideal machines* refers to those machines consistent

with AI_{IB}, and the term *human-like machines* refers to those machines consistent with AI_{HB}. Of course, there can exist ideal human-like machines, but typically, as shown earlier, ideal machines seek rational results and in so doing, ideal machines do not usually behave the same as observed human phenomena.

AI as constructing a superior human

The perception of AI as constructing machines that comply with an ideal standard (AI_{IB}) has an advantage. What good is an answer if it is not the best answer? For example, a patient in serious medical condition may not be able to tolerate much human error in his care, yet the opportunities for human mistakes are enormous. It is desirable to have machines that avoid such errors.

Consider then what may be the ultimate machine of AI_{IB}. Imagine a single machine that can run faster, jump higher, think faster and reason better than any human. Rather than just having human-like vision, it also has telescopic ability; rather than just having human-like hearing, it also has ultrasonic ability; and, the list of comparisons continue such that each human capability appears improved and better in the machine. In this view, constructing the ultimate AI_{IB} machine posits AI as a pursuit to construct a superior human.

Example (AI expert systems in medical decision-making).

The task of diagnosing human medical conditions is to map a patient's symptoms and disease manifestations to diseases believed to be the causes of such manifestations. In the 1970's and 1980's numerous works in AI sought to diagnose human medical conditions. Here are a few. MYCIN [Davis et al. 1977] was influential in initiating AI interest. It used implications as its primary representation to diagnose bacterial infections of the blood. MYCIN spawned a family of related programs – EMYCIN for anti-microbial therapy, PUFF/VM for pulmonary function, ONCOCIN for outpatient chemotherapy, and HEADMED for psychopharmacology. The CASENET System [Weiss et al. 1978] used a causal association network model of disease to diagnose Glaucoma and other diseases of the eye. INTERNIST[1] [Myers et al. 1982] used a database of 500 known disease profiles with an inference engine that scored and partitioned profiles, to ambitiously diagnose all the diseases of general internal medicine.

Each of these programs was judged comparable to human expert physicians in some kind of test. However, the performance of these systems degraded substantially in cases outside their narrowly defined domain of coverage, in cases involving complex interactions, and in cases involving multiple disorders. Perhaps the most significant limiting factor of these systems was an inability to capture the adaptive reasoning abilities of humans. Knowledge, which was considered useful by the systems' designers, was abstracted from human physicians and medical textbooks and hardwired into an inflexible framework. Despite being more than 20 years old, none of these systems are in widespread clinical use today.

Nevertheless, there is one success story. By 1996, Pathfinder, commercialized as IntelliPath [Heckerman 1991], had been approved by the American Medical Association and deployed in hundreds of hospitals worldwide. IntelliPath is a probabilistic system that diagnoses hematopathology. It uses Bayesian networks, which are graphical models of the relationships between symptoms and diseases. In comparison to MYCIN, CASENET, and earlier approaches, IntelliPath uses a more rigorous and accurate mathematical framework. The system also includes videodisc libraries of histological slides. It creates a differential diagnosis of plausible diseases

[1] INTERNIST's name was changed to CADUCEUS in 1983 for legal reasons.

based on the histological features presented. A human physician can work with the system to help distinguish among those competing diagnoses.

Despite appearances, IntelliPath has serious limitations. A paper in Human Pathology [Miller et al. 1997] reports that IntelliPath is useful, but its diagnostic accuracy is not acceptable. The authors recommend its use be limited to training physicians and not be used as an on-the-job diagnostic tool. There are currently no published uses of IntelliPath as an on-the-job diagnostic tool. Also, the 105,000 conditional probabilities on which IntelliPath bases its diagnostic reasoning have never been disclosed and reportedly result from one expert physician's subjective assessments and are not the result of statistical techniques applied to a foundation of empirical data.

As the example above discussed, none of the systems that have attempted to diagnose human medical conditions have demonstrated sufficient competence and reliability to make decisions regarding human life.

This underscores a concern with the notion of AI as constructing ideal machines or a superior human. So far, resulting machines typically perform limited tasks in abstracted and disjoint environments and no single machine performs a broad array of tasks in complex and robust environments like humans. This is not simply a matter of scaling because the computational demand to achieve rationality as the environment becomes more complicated usually grows exponentially. Conversely, this may imply a limit on the size and nature of tasks that can be pursued by rational machines, or at least, by traditional approaches to constructing ideal machines.

Of course, the role advances in technology can play in providing machines with exponentially more computational power must not be forgotten. The evolution in chess playing machines, as discussed earlier, provides an example. Also, the ability for IntelliPath to include videodisc libraries gives it advantages over its predecessors. As technical advances in computer hardware continue to provide machines that process instructions faster, utilize greater storage capacities, and have improved human-computer interactions, ideal machines will often work in increasingly more complex environments as a result of hardware improvements alone. In these cases, it may be electrical, mechanical, chemical and computer engineers, as much as AI researchers, who will be responsible for advancing machines consistent to AI_{IB}.

AI as constructing smarter machines

Often AI researchers qualify tasks of interest to AI as being those that make humans seem intelligent [Winston 1977]. But there are many problems in identifying which tasks those might be. In any case, there is tremendous breadth in the tasks done by humans and some tasks are not very interesting, whether they are considered intelligent or not. For example, some consider chess playing an interesting task while others consider the taking of dictation (i.e., transcribing spoken words to text) an uninteresting task. Often mundane or uninteresting tasks are important tasks and are the kinds of tasks society wants machines to do. So by extension, it is not surprising to consider AI as expanding the usefulness of today's machines by constructing smarter computers. This perspective is summarized in definition 9.

Definition 9. ***(artificial intelligence as smarter computers)*** *Artificial intelligence is the study of ideas to construct smarter computers.*

Definition 9 frames AI research as a vehicle for feeding new machines and technology to computer science and other fields as well as to society. Evolution and fitness under this model are based on market supply and demand. Funding often determines the kind of machines that get produced, but in this

pursuit, such funding may be based on the potential economic impact of the machine expected within a short time period.

There is no doubt that many AI success stories fall into this category. This is not surprising for two reasons. First, AI researchers have to be paid. And secondly, AI has often had to make technical advances in pursuit of its mission. David Waltz [1996], a past President of the American Association of Artificial Intelligence and a past President of the Association for Computing Machinery (ACM) Special Interest Group on Artificial Intelligence provides the following examples:

> "Laboratories whose focus was AI first conceived and demonstrated such well-known technologies as the mouse, time-sharing, high-level symbolic programming languages (Lisp, Prolog, Scheme), computer graphics, the graphical user interface (GUI), computer games, the laser printer, object-oriented programming, the personal computer, email, hypertext, symbolic mathematics systems (Macsyma, Mathematica, Maple, Derive), and, most recently, the software agents which are now popular on the World Wide Web."

The problem is that without adopting the mission statement of AI as constructing smarter computers, success in this area does not mean progress. Works constructed in this vision of AI are not necessarily consistent with either AI_{IB} or AI_{HB}. Recall the survey results reported earlier. Many of these works are not supported in either the survey of what people say is AI or in the survey of what people are doing when they are doing AI research. So even though AI researchers have contributed significantly to computer science and other fields along these lines, by all the previous characterizations of AI, no such work is progress.

Example (Speech Recognition).

> One area that appears to counter some of these points is the development of speech recognition systems. The task of speech recognition is to map a digitally encoded signal to a string of words. Over the past 12 years speech recognition technology has emerged as a success story, evolving into large vocabulary continuous speech systems capable of transcribing naturally spoken sentences on specific topics from typical human talkers. Surprisingly, virtually all the early attempts were considered AI works, but now, virtually all speech recognition projects are considered general computer science.
>
> In the 1970s, the Advanced Research Projects Agency (ARPA) of the United States Department of Defense conducted its Speech Understanding Research Project (ARPA SUR) which catapulted speech recognition research from speaker-dependent, small-wordlist recognition to the large-scale language model systems available today. The earlier efforts were ad hoc approaches primarily produced by AI researchers. They were not fast or robust enough for ARPA's standards.
>
> Eventually, one system met ARPA SUR's original mandate, Harpy [Klatt, 1977], and the success of Harpy's statistical modeling techniques continues to have a profound effect as researchers seek to build larger statistical knowledge bases in an attempt to overcome problems and extend the performance of systems. The use of Hidden Markov Models required rigorous mathematics but provided a large vocabulary model. Slices of sound are regularly sampled and compared to a statistical network of stored sounds within the machine to determine the most likely utterance. As computers themselves got faster and had more memory storage available, commercial speech systems using this approach emerged and today several continuous speech systems are commercially available.

> Despite the success of speech recognition systems, achieving human-like performance remains distant. Lippmann [1996] points out: (1) the error rates of machines are more than an order of magnitude greater than those for humans under the most ideal circumstances for the machines; (2) machine performance plummets much faster than humans when operating in noise and other degraded conditions; (3) humans exhibit much more powerful types of adaptation and incorporate newly learned words; and (4) humans rely on context much less than machines and can accurately recognize nonsense words, which are words that sound like English words but in reality have no meaning.
>
> The human brain tends to automatically categorize speech sounds. Fletcher and his colleagues, who studied the principles behind human speech recognition from 1918 to 1950 at Bell Labs [Allen, 1994], concluded that humans decode speech sounds into independent units at an early stage, before semantic context is used. These findings imply that humans use phoneme-like units and classification of sound into these units happens at an early processing stage. If so, this would dramatically reduce the computation required and may be responsible for the faster, flexible and more robust performance of humans.

As the example above discussed, the evolution of speech recognition systems was driven by performance standards determined by ARPA. Development moved from ad hoc systems to those using more rigorous mathematical models following AI_{IB}. While this brought some success, driving forces such as the one ARPA provided are rare. The likelihood that pursuing a market-driven walk would ever achieve Turing's dream or a superior human is highly unlikely because resulting machines would be driven by short-term payoff with no long-term overall research strategy.

So far in this section, the notion of AI as constructing a superior human and the notion of AI as constructing smarter computers have been discussed. In the next section the notion of AI as constructing a human-like machine is discussed. This paper then ends with by comparing human judgment to statistical approaches, thereby further contrasting AI_{IB} and AI_{HB}.

AI as constructing a human-like machine

The notion of AI as constructing human-like machines (AI_{HB}) has some advantages. First, humans provide an existence proof against which to compare and test. Humans perform a wide array of tasks in complex and complicated environments. Among other features, humans are adaptive, can anticipate, can take initiative, can make suggestions and can provide explanations.

A second advantage is the ability to evolve learners from simple to complex forms operating in the same environments as humans and to compare machine results to human performance. For example, imagine a machine learner that has evolved to third grade mathematics and first grade reading comprehension; such performance levels can be determined by the standardized tests given to human learners. Such evolution has the constant requirement of integration because machine learners would be expected to continue to master previously learned material, as humans typically are. This is not to say that given a machine M_k, which performs at educational level k, and another machine M_{k+1}, which achieves an educational level of k+1, that M_{k+1} does not require completely different hardwiring from M_k. Rather, the requirement for constant integration of performance measures progress.

There are some concerns with the notion of AI as constructing human-like machines (AI_{HB}). A disadvantage is that some human characteristics may not be desirable. A researcher can only avoid

modeling what might be considered an undesirable characteristic if it provably has no bearing on the task. Examples of human characteristics that may not be desirable include emotions, competitiveness, and desire for revenge. On the other hand, we know so little about the human mind these seemingly undesirable characteristics may be advantageous. A second concern, as mentioned earlier, is that human performance is not always ideal.

From the survey results reported earlier, it is not surprising that there are few development examples to provide in this pursuit. The closest works appear in cognitive science where machines are often constructed to explain observed human phenomena consistent with a psychological model of the mind. This additional constraint of complying to a psychological model of the mind makes these works a subset of AI_{HB}.

Example (Soar).

> Cognitive science integrates expertise on intelligence from numerous fields including psychology, linguistics, anthropology and artificial intelligence. One cognitive science pursuit is Soar, whose aim is to develop and apply a unified theory of human and machine intelligence. Soar is realized as a computational architecture (a machine) that has a fixed set of hardwired mechanisms. These mechanisms form the basis for accomplishing problem solving, planning, learning and other cognitive tasks. Implications constitute its primary internal representation. Soar began in 1983 and the 21st North American Soar Workshop was held in 2001. Papers presented included works on visual perception, game playing with a simulated opponent, and emotional agents. Despite its longevity, Soar and its mechanisms have not been readily adopted within traditional AI research, presumably because its fixed mechanisms perform a limited number of low-level functions. What Soar does not provide directly are the higher-level capabilities of an intelligent machine such as problem solving, planning and learning. These must be hardwired atop of Soar.

While traditional AI works on constructing human-like machines have been lacking, there has been considerable attention paid to testing humans and machines in order to determine if they are comparable in performance. The Chess Test, mentioned earlier, is an example, though some now consider it obsolete after Kasparov's defeat. But, the Turing Test, the Loebner Prize and Human Interactive Proofs deserve some attention.

Turing Test

In Turing's 1950 paper he considers the question "Can machines think?" Rather than answering the question directly, he proposes an "imitation game" which is now commonly termed *the Turing Test* or the unrestricted Turing Test. Here is how the game is played. It involves a human interrogator, a human player and a machine. A human interrogator is given two communication channels on which to ask written questions and receive written responses. One channel communicates with the human player and the other with the machine. After some time, the human interrogator must determine which channel had responses provided by the human and which channel had responses provided by the machine. Presumably, the machine is trying to convince the interrogator it is human, while the human is trying to convince the interrogator he is a machine[2].

Turing considered the machine's performance comparable to the human's performance if the machine played the game so well that an average human interrogator would not have more than a 70% chance of

[2] Other interpretations of the intended goals of the human player and the machine are possible in Turing's paper, thereby leading to variations of the game.

correctly identifying the machine after 5 minutes of questioning. Even though today's machines show proficiency in some very specific, well-defined tasks, machines are still far behind humans in general cognitive abilities. It is not surprising therefore, that no machine to date has passed the unrestricted Turing Test.

Loebner Prize

The survey results reported earlier showed that very little work has been done on constructing human-like machines specifically. In an attempt to drive development in this area, Hugh Loebner underwrote an annual contest to compare human and machines in a limited Turing Test [Loebner 1993]. Each year an annual prize of $2000 and a medal is awarded to the "most" human-like machine among the contestants. As long as there is more than one contestant, a prize is awarded. The test used in the contest is far more limited than the unrestricted Turing Test in order to give machines a starting place[3]. The contest began in 1991. Over the last 10 years, 6 individuals have won, some winning more than once.

Despite the longevity and motivation for the contest, winning works are rarely, if ever, cited in AI research literature. There are numerous reasons for this, but one fundamental reason concerns the nature of the kinds of machines that tend to win. These machines often engage in whimsical conversation because they substitute English syntax for semantics. That is, these machines are void of understanding the meaning conveyed in the communications they receive or produce. Their goal after all is to trick the interrogator. So, it is the human interrogator that provides meaning to the garbled text in an attempt to classify it as being written either by a machine trying to act like a human or by a human trying to act like a machine.

Human Interactive Proofs – CAPTCHAs

An interesting and seemingly powerful twist comes from the cryptography community by way of human interactive proofs (HIPs). One kind of HIP is a program that attempts to tell humans apart from machines by grading a computerized test, which most humans can pass but which current machines cannot pass. This kind of HIP is called a CAPTCHA, which stands for "Completely Automated Public Turing test to tell Computers and Humans Apart" [Ahn et al. 2002]. An inherent goal of a CAPTCHA is for a machine to eventually pass the test, but in order to do so machines will have to improve to be comparable to human performance. In the meantime, the test itself is useful. A necessary requirement is that the source code for the program providing the test be publicly available. An example of a CAPTCHA is Gimpy, which was developed in collaboration with Yahoo! to keep automated programs out of their on-line chat rooms. Gimpy works by displaying an obscured image of a word, which is to be typed in order to gain entry to the chat room. The image is obscured in such a way that current optical character recognition algorithms cannot reliably decipher it, but humans generally have no difficulty identifying the word and then typing it. If a machine passes Gimpy's test, then the machine has improved optical character recognition. The legitimacy of this claim is based on correctness proofs of Gimpy's source code, which is publicly available.

CAPTCHAs can be quite useful to AI. For example, a set of CAPTCHAs designed around tasks of interest to AI could provide economic and technical driving forces for developing works consistent with AI_{HB}. Progress could be measured by how many CAPTCHAs over time were no longer able to detect a difference between humans and machines. The ultimate goal of AI_{HB} is achieved when there exists no CAPTCHA that can distinguish humans from machines. Unlike the Loebner Prize and the Turing Test, the tasks that form the basis for comparisons are not fixed in CAPTCHAs. Like the Loebner Prize,

[3] For a critique of the Loebner Prize and its test construction, see Shieber 1994.

CAPTCHAs operationally drive ongoing assessments. As a final comparison to the Loebner Prize and the Turing Test, the judge in a CAPTCHA is itself a machine.

In summary, the developmental approaches that have been discussed do show progress in AI. Common patterns found in successful works are: (1) the use of more rigorous statistics; and/or, (2) the ability to take advantage of more powerful hardware. Machine performance is hardly near general human cognitive ability. It is an open question as to whether there is a limit on what can be achieved using rigorous statistical approaches. Even though humans provide an existence proof against which to compare, virtually no work in AI research is actually modeled after human performance specifically.

Human judgment versus probabilistic approaches

Before concluding, the distinction between works that model human behavior (AI_{HB}) and those that model an ideal standard (AI_{IB}) must be addressed further. This division seems to pit AI's original goal and what others expect of AI against the work primarily conducted. Human judgments versus probabilistic approaches are the subject of the comparison.

People use intuitive strategies in estimating the likelihood of events, testing hypothesis, and updating beliefs in the face of new evidence. On the other hand, applying mathematical models that use probability, stochastic processes, or Bayesian inference, often yield answers that are quite different than their intuitive counterpart.

The psychologists Tversky and Kahneman [1982] posed the following example[4]. It seems to illustrate how humans view base rate information incidentally and not as a contributing or causal factor. As a result, humans provide answers different from those found in statistics.

> A cab was involved in a hit and run accident at night. Two cab companies, the Brown and the Blue, operate in the city. Consider the following information:
> (a) 85% of the cabs in the city are Brown and 15% are Blue.
> (b) A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

When potential jurors were then asked the probability the cab involved in the accident was Blue rather than Brown, typical answers were around 80 percent. The correct mathematical answer is 41%. That is, the hit-and-run cab is more likely to have been Brown and not Blue. Base rate misconceptions are not limited to the average person without an advanced mathematics education either. Similar results have been found involving physicians and graduate students [Cassells et al. 1978].

As shown, human judgment and decision making under uncertainty in real-world situations often lead to conclusions that completely contradict statistical analysis leading to a debate as to whether probabilistic approaches can ultimately model human thinking and behavior. Probabilistic approaches are important in resolving uncertainty, but humans often behave differently and far more creatively [Gigerenzer 2000].

Intuitive judgment based on assessing likelihood, plays a crucial role in many cognitive areas for humans. Examples include: learning [Hilard and Bower 1966], attribution theory [Kelley 1967],

[4] The example provided herein was modified to use Brown and Blue cabs rather than Green and Blue cabs in order to heighten the ambiguity for the reader.

judgments of causality [Einhorn and Hogarth 1986], decision making [Burden 1997], clinical assessment [Chapman and Chapman 1969], implicit personality theories [Fiske and Neuberg 1990], stereotyping [Fiske 1999], scientific reasoning [Mynatt, Doherty and Tweney 1978], and categorization [Smith and Medlin 1981]. These are all areas in which humans are currently far superior to machines and many believe the difference is due in great part to the human intuitive process. If so, should AI model human performance or an ideal standard in order to achieve its goals?

Human versus Machine

When given an interesting task to achieve, the approach of prescribing superior behavior involves: (1) using human ingenuity and the laws of logic and probability to find a way for a machine to achieve the task rationally; and then, (2) hardwiring that solution into the machine. The approach of modeling human behavior involves: (1) studying how humans achieve the task; (2) using human ingenuity to find a way for a machine to achieve the task consistent with observed human performance; and then, (3) hardwiring those behaviors into the machine. This later approach of modeling human behavior may not always be as correct or as fast, but is believed to scale more easily into larger, complex tasks than prescription allows. In the spirit of this belief, here is a comparison of these two approaches in evolving a game-playing machine.

First, consider the game of Tic-Tac-Toe. This game involves two players and begins with an empty 3 x 3 matrix. Each player takes turns writing a single letter in an unused cell. One player writes X's and the other player writes O's. The objective is to get three of the same letters in adjacent cells. In this case, a machine is to be constructed that plays the role of one of the players. After playing numerous games, humans tend to develop a set of rules that are guaranteed to win or draw (which is the case of no winner), but never lose. These rules are hardwired into the machine thereby prescribing superior behavior. Alternatively, human ingenuity is used to hardwire a machine that attempts to learn the same rules in much the same way a human appears to learn them.

Now consider expanding these two machines to also play Hangman. This game also involves two players. One player selects an English word and displays an empty vector whose length is the same as the number of letters in the word. The other player (the "guesser") attempts to guess the word, a letter at a time. Each time the letter guessed appears in the word, the cell(s) in the vector that represent that letter in the word is changed to display the letter. There are a limited number of wrong guesses allowed. If all the letters of the word are revealed before the maximum number of wrong guesses is encountered, the guesser wins. Otherwise, the guesser loses.

Again, human ingenuity prescribes superior behavior by hardwiring a strategy for good guessing. The resulting machine would then have two hardwired games and able to play only these two games. The machine based on modeling human behavior has its hardwiring expanded to learn strategies for playing Hangman. However, this can be done in such a way that the same machine can also learn strategies for playing Find-a-Word with no additional hardwiring. The Find-a-Word game has a 2-dimensional matrix displaying letters in all its cells. The machine is to find English words that are spelled across adjacent cells.

When this discussion is generalized, it gets recursive. Prescribing superior behavior is the result of human ingenuity. Humans determine an optimal set of deterministic or probabilistic rules and then hardwire those rules into the machine. Imagine constructing a machine based on this human pursuit. The humans who prescribe such behaviors become the subjects of those who seek to model them. If successful, a machine would be constructed that prescribes superior behaviors to other machines, thereby

achieving this human task. But this machine itself would be hardwired by humans, thereby resulting from some deterministic and/or probabilistic rules. Notwithstanding the recursion, there is a fundamental division between those who believe the laws of logic and probability rule the realm of sound reasoning and those who believe psychology is indispensable for sound reasoning in a complex and uncertain world.

Final remarks on the meaning of AI.

Many people's dream of artificial intelligence concerns itself with intelligent behavior -- the things that make us seem intelligent. In an ultimate view such people want artificial intelligence researchers to re-create a perception of humankind by building a machine in the human image. This is a strong statement, but it describes the underlying motivation of many expectations of AI. Clearly this is how AI began, consistent with Turing's dream, but available technology, financial opportunity, scientific know-how and obstacles seem to have pushed AI research somewhere else.

AI may need to re-define itself to better match the actual work conducted and thereby ensure viability for its name. Doing so would be prudent to protect funding streams because progress could be shown consistent to the effort provided and the community's focus of attention. On the other hand, some may view these findings as a call to fund works that model human behavior specifically where progress may be easier to measure. No matter the outcome, the desire to construct a machine in the image of a human will not die. A similar desire is present in art, whether expressed as a painting, a sculpture, or even as a film, and seems intricately linked to the human search for immortality and need for self-reflection. Therefore, the quest will rekindle itself over and over again as new technology and opportunities burst forth. The cultural dream of AI will never die.

Acknowledgments

This work is dedicated to those who pursue the thinking machine. I thank Herbert Simon and Marvin Minsky for the time fragments each spent separately with me, in public and in private, in animated debate and vibrant conversation discussing various aspects of AI. While they neither agreed with each other on the best direction for AI research, nor with me, they each gave me a little time and energy to engage in spirited discussions. I also thank Gerald Sussman and Patrick Winston for their pockets of time and passionate discussions on yet other perspectives of AI. At MIT and at Carnegie Mellon University (CMU), these titans fostered constructive and intellectually stimulating environments where my views as a newcomer were eagerly discussed along with their veteran experiences, and for that, I am very grateful. The experience gave me the confidence to develop these ideas, which are my own. I also thank Sherice Livingston for her assistance in typing bibliographical information for the databases. Finally, I thank Sylvia Barrett, Pat Larkey, Elaine Newton, Lawrence Beasley, and Bradley Malin for suggestions and comments. This work was supported in part by the Laboratory for International Data Privacy in the School of Computer Science at CMU.

References

Ahn, L., Blum, M. and Langford, J. Telling Humans and Computers Apart (Automatically) or How Lazy Cryptographers do AI. CMU Technical Report CMU-CS-02-117. February, 2002. Also available at http://reports-archive.adm.cs.cmu.edu/anon/2002/CMU-CS-02-117.pdf

Allen, J. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*: 1994, 2(4).

Bellman, R. *An introduction to artificial intelligence: can computers think?* San Francisco: Boyd and Fraser Publishing Company. 1978.

Brooks, R. Intelligence without representation. North-Holland 1991. Reprinted in *Computation and Intelligence* by G. Luger, ed. MIT Press, Cambridge, 1995, p.343-362.

Burden, B.C. "Deterministic and Probabilistic Voting Models" *American Journal of Political Science*. 1997, 41: 1150-69.

Cassells, W., Schoenberger, A., and Grayboys, T. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*: 299, pp.999-1001.

Chapman, L., and Chapman, J. *Disordered thought in schizophrenia*. Englewood Cliffs, N.J.: Prentice-Hall. 1973.

Charniak, E. and McDermott, D. *Introduction to artificial intelligence.* Reading: Addison-Wesley. 1985.

Davis, R., Buchanan, B., and Shortliffe, E. Production rules as a representation for a knowledge based consultation program. *Artificial intelligence.* 1977, v8, pp. 15-45.

Doyle, J. Cleaving (unto) artificial intelligence. *ACM Computing Surveys* 1996, v28, n4.

Einhorn, H.J. and Hogarth, R.M. Judging probable cause. *Psychological Bulletin*. 1986, 99(1) 3-19.

Fiske, S. T. Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Lindzey, (Eds.) *Handbook of social psychology*. New York: McGraw- Hill. 1998, 4th ed., 357-411.

Fiske, S. T., and Neuberg, S. L. A continuum of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology*. New York: Academic. 1990, 23, pp. 1-74.

Gigerenzer, G. *Adaptive thinking: rationality in the real world*. New York: Oxford. 2000.

Haugeland, J., editor. *Artificial intelligence: the very idea.* Cambridge: MIT Press. 1985.

Hayes-Roth, F. Artificial intelligence: what works and what doesn't. *AI Magazine* 1997, v18, n2, p99-113.

Hilard, E. R. and Bower, G. H. *Theories of learning*. New York: Appleton-Century-Crofts. 1966.

House, A. On the learning of speechlike vocabularies. *Journal of Verbal and Learning Verbal Behavior*, 1, 133-143.

Kelley, H. H. Attribution theory in social psychology. In D. Levine, Ed., *Nebraska Symposium on Motivation*. Lincoln: University of Nebraska Press. 1967.

Klatt, D. Review of the ARPA Speech Understanding Project. *Journal of the Acoustic Society of America*, 62, 1345-1366.

Knight, T. and Sussman, G. *Cellular Gate Technology.* MIT Artificial Intelligence Laboratory, Working Paper, 1997. Available at http://www.ai.mit.edu/people/tk/ce/cellgates.ps.

Kurzweil, R. *The age of intelligent machines.* Cambridge: MIT Press. 1990.

Lenat, D. From 2001 to 2001: common sense and the mind of HAL. In Stork, D., editor, *2001's Computer as Dream and Reality*. 1998, pp. 193-210.

Liberman, A.M. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

Lippmann, R. Recognition by humans and machines: miles to go before we sleep. *Speech Communication*: 1996, 18(3), 247-248.

Loebner, H. *In response*, 1993 at http://www.loebner.net/Prizef/In-response.html.

Luger, G. and Stubblefield, W. *Artificial intelligence: structures and strategies for complex problem solving.* Redwood City: Benjamin/Cummings. 1993, 2nd ed.

Miller, R., Shultz, E. and Harrison, J. Is there a role for expert systems in diagnostic anatomic pathology? *Human Pathology*, 1997, 28(9): 999-1001.

Myers, J., Pople, H., and Miller, R. CADUCEUS: a computerized diagnostic consultation system in internal medicine. In, *Proceedings, Symposium on Computer Applications in Medicine*. 1982, pp.44-47.

Mynatt, C. R., Doherty, M.E., and Tweney, R.D. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*. 1977, 29:85-95.

Newborn, M. *Kasparov versus Deep Blue: computer chess comes of age.* New York: Springer-Verlag. 1996.

Newell, A. and Simon, H. GPS, a program that simulates human thought. In Billing, H., editor, *Lerende Automaten.* 1961, pp. 109-124. Reprinted in Feigenbaum and Feldman, *Computers and Thought* 1963.

Prescott, T. and Ibbotson, C. A robot trace-maker: modeling the fossil evidence of early invertebrate behavior. *Artificial Life*: 3(4), pp.289-306.

Pauker, S. et al. Towards the simulation of clinical cognition taking a present illness by computer. In, W. Clancey and E. Shortliffe, eds. *Readings in Medical Artificial Intelligence: the first decade*. Reading: Addison Wesley. 1984.

Reddy, R. To dream the possible dream. *Communications of the ACM*, 1996, v39, n5, p105-112.

Rich, E. and Knight, T. *Artificial intelligence.* New York: McGraw-Hill. 1991, 2nd ed.

Russell, S. and Norvig, P. *Artificial intelligence: a modern approach.* Englewood Cliffs: Prentice-Hall, 1995.

Russell, S. and Norvig, P. *Adoptions of AI: A Modern Approach.* http://cs.berkeley.edu/~russell/adoptions.html, January 18, 1999.

Schalkoff, R. *Artificial intelligence: an engineering approach.* New York: McGraw-Hill. 1990.

Shieber, S. Lessons from a restricted Turing test. *Communications of the ACM*, 1994, v37, n6, p70-78.

Simon, H. Heuristic problem solving: the next advance in operations research. *Operations Research.* 1958, Jan-Feb, pp.1-10.

Smith, E. E., and Medin, D. L. Categories and concepts. Cambridge: Harvard University Press. 1981.

Spielberg, S. *E.T.: the extra terrestrial.* Universal Pictures. 1982.

Spielberg, S. *A.I.: artificial intelligence.* Warner Brothers and Dream Works. 2001.

Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. In: Cimino, JJ, ed. *Proceedings, Journal of the American Medical Informatics Association*. Washington, DC: Hanley and Belfus, Inc, 1996:333-337.

Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Creemers, A., Dellaert, F., Fox, D., Hahnel, D., Rosenberg, C., Roy, N., Schulte, J. and Schultz, D. Probabilistic algorithms and the interactive museum tour-guide robot Minerva. *International Journal of Robotics Research.* 2000, November 19(11), pp.972-999.

Turing, A. Computing machinery and intelligence. *Mind* 1950. Reprinted in *Computation and Intelligence* by G. Luger, ed. MIT Press, Cambridge, 1995, pp.23-46.

Tversky, A. and Kahneman, D. Evidential impact of base rates. In Kahneman, Slovic and Tversky, eds., *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press, 1982:153-162.

Waltz, D. Artificial Intelligence: realizing the ultimate promises of computing. NEC Research Institute and the Computing Research Association. 1996. Available May 2001 at http://www.cs.washington.edu/homes/lazowska/cra/ai/html.

Wegner, P. and Doyle, J. Editorial: strategic directions in computing research. *ACM Computing Surveys* 1996, v28, n4.

Weiss, S., et al. A model-based method for computer aided medical decision making. *Artificial Intelligence* 1978, v11, n1.

Weizenbaum, J. ELIZA-- a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 1966, v9, pp.36-45.

Wilkes, M. Artificial intelligence as the year 2000 approaches. *Communications of the ACM* 1992, v35, n8, pp.17-20.

Winston, P. *Artificial intelligence.* Reading: Addison-Wesley. 1977, 1st ed.

Winston, P. *Artificial intelligence.* Reading: Addison-Wesley. 1985, 3rd ed.

Appendix 1 *Definitions of AI found in Other References*

Human thinking

Academic Press Dictionary of Science Technology, 1996

> AI is a field of study concerned with the development and use of computer systems that have some resemblance to human intelligence.

U.S. Dept. Commerce; National Telecommunications and Information Administration, 1998

> AI is the branch of computer science that attempts to approximate the results of human reasoning.

whatis.com, 1998

> [whatis® is a knowledge exploration tool about information technology, especially about the Internet and computers. It contains over 1,500 individual encyclopedic definition/topics and a number of quick-reference pages. The topics contain over 5,000 hyperlinked cross-references and over 3,000 links to others sites for further information.]
>
> > Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using the rules to reach approximate or definite conclusions), and self-correction.

Funk and Wagnall's Encyclopedia, 1986

> AI would indicate the ability of an artifact to perform the same kinds of functions that characterize human thought.

Dictionary of Computing, 1994, Peter Collin Publishing

> AI is the design and device of computer programs that attempt to imitate human intelligence and decision making functions, providing basic reasoning and other human characteristics.

On-line Dictionary of Computing. Dennis Howe, 1998

> AI is an attempt to model aspects of human thought on computers. It is also sometimes defined as trying to solve by computer any problem that a human can solve faster.

Human behavior

PC Webopedia, 1998

[used by Yahoo and rated 1 of best 100 sites by PC magazine]

> AI is the branch of computer science concerned with making computers behave like humans.

Webster's Dictionary, 1991

> The capability of a machine to imitate intelligent human behavior; a branch of computer science dealing with the simulation of intelligent behavior in computers.

Dictionary of Computer Terms, Barron's New York, 1989.

> The branch of computer science that deals with using computers to simulate human thinking.

Ideal thinking

The Hutchinson Dictionary of Science, 1994, TSP Edition

> AI concerns the creation of computer programs that can perform actions comparable with those of an intelligent human