

Materials: Transparency of Russell and Norvig page 5  
Demonstration of conversational blocks program  
Demonstration of Project 3  
Demonstration of Eliza program  
Film series handout

I. What is AI?

- ---- - -

- A. Suppose one of your friends were to ask you what you are taking this semester, and when you mentioned "Artificial Intelligence" asked you "What's that all about?" What would you say?

ASK CLASS TO WRITE OUT A RESPONSE - THEN DISCUSS

1. Typically, on the first day of a course, the professor attempts to give students a good idea of what the field being considered is all about.
  2. However, there is a real sense - in this course in particular - where one might say that a major goal of the course is that, by the end of the course, you might have an idea of what AI is all about.
  3. For example, the author of our text speaks of "the paradoxical notion of a field of study whose major goals include its own definition" (page 2).
- B. There is no universal agreement as to what is encompassed by the term "Artificial Intelligence", even among those working in the field.
1. It is standard for AI textbooks to begin with a chapter that seeks to answer this question - but with varying answers depending on the book.
  2. One widely-used text by Russell and Norvig collects eight different definitions from other writers:

TRANSPARENCY - Russell and Norvig page 5

Key questions:

- a. Does what we humans are define what intelligence is, or is there some larger concept of "rationality" or "pure logic" of which human intelligence is just one expression?

Corollary: is possession of a human (or human-like) body and undergoing human (or human-like) experiences essential to intelligence? (at least of the sort we experience?)

Clearly, this question leads into some fascinating "integration of faith and learning" type questions - such as the nature of the soul. I have deliberately chosen to put this at the end of the course, so that we can draw on what we have learned about AI research throughout the semester.

- b. Is the essential feature of intelligence process ("thinking") or behavior ("acting")?

Corollary: if a machine could perform the way we do at some task, but should do so using a very different process, would we be right to call it intelligent?

Example: "Deep Blue", a chess-playing system developed by IBM, has been able to defeat the world chess champion - Garry Kasparov. Is Deep Blue intelligent? (IBM's official web site on Deep Blue says no, but others would say yes.)

Key Question: if a machine were to be intelligent, how would we know it is except by observing its behavior?

This is key issue behind the "Turing Test". We will read and discuss Turing's paper next week, along with a paper by Searle that takes an opposing view the following week.

- C. One reason why there are a wide variety of definitions of what "artificial intelligence" is all about is that the notion of "intelligence" is itself an ill-understood concept.
1. Consider some of the ways we use the term in ordinary speech:
    - a. "I wonder if there is intelligent life on other planets".
    - b. "Suzy is such an intelligent person".
    - c. "Among dogs, German shepherds are known for their intelligence."
  2. We may distinguish both an absolute and a relative usage of the term.
    - a. When we wonder if there is "intelligent" life on other planets, we are using the term in an absolute sense. Among earthly creatures, intelligence, in this sense, is an attribute possessed only by humans. (Though we recognize that there are non earthly intelligent beings, including - at least - God and the angels.)
    - b. On the other hand, we use intelligent in a relative sense, recognizing degrees of intelligence among humans and also some degree of intelligence in sub-human creatures. (Though if a space probe discovered a German Shepherd on Mars we would not regard that as the discovery of intelligent life.)
  3. Most people would have little problem with the notion that computers can exhibit intelligence in the relative sense. The question is whether they are or can ever be "intelligent" in the absolute sense.
    - a. On the one hand, even the simplest functions of the computer - such as arithmetic - are "intelligent" in a real sense. Simple addition is a skill that even the most intelligent dog will apparently never possess, one that human children only begin to master when they have reached school age, and one that adults often make errors in.
    - b. Sophisticated computer programs exist today which are capable of performing at a very high degree of expertise in such areas as playing chess, diagnosing bacterial infections, and configuring other computer systems.

- c. Yet today's computer systems fail miserably at tasks (such as interpreting visual data) which are no problem for even relatively unintelligent animals.
  - d. While the level of (relative) intelligence exhibited by computer systems has been steadily increasing, it is an open question whether it will ever be appropriate to speak of a computer as intelligent in the absolute sense. This question will be in the background of several of our discussions during the course.
- D. Related to this issue is a distinction some people draw between "intelligence" and "consciousness".
- 1. Is it possible to think without being having consciousness (a "self")? (Evidently, we ourselves do this - sometimes we do things that require intelligence without being conscious that we are doing them.)
  - 2. Does consciousness emerge from thought, or is it something distinct?
  - 3. Perhaps this distinction is related to two ways in which we use the word "intelligent" - e.g. perhaps the absolute use of "intelligent" is associated with the idea of self-consciousness. Humans are conscious of themselves; it is not clear that dogs, say, are.
- E. Another factor which complicates defining "artificial intelligence" is philosophical in nature.
- 1. Strong AI (the upper right-hand corner of the chart we looked at earlier) is rooted in a Western philosophical tradition that goes back to Plato - the idea that truth can ultimately be formalized in the form of propositions or rules.
  - 2. However, not all philosophers subscribe to this tradition, and not all the work we will study in this course is based on it (though much is.) For example, work along the lines suggested by the definitions in the lower left-hand corner of the chart we looked at earlier does not necessarily require any assumptions about the ability to formalize everything.
  - 3. Sections 1.1.1 to 1.1.3 of the introduction to our book (which I've assigned you to skim) deals with this question, as does the final part of the book (which we will come to at the end of the course.)
  - 4. One of the most outspoken critics of strong AI - Hubert Dreyfus - bases his critique of strong AI on his assertion that strong AI is rooted in fundamentally wrong philosophy. (Dreyfus was a philosophy instructor at MIT in the early days of AI work, and has become one of its most prolific critics. Though we will read and discuss an article by him during the semester, it does not deal directly with this specific issue.)

## II. The Goals of AI

-- --- ----- -- --

- A. Rather than attempting to define AI, we perhaps should consider some of the goals aimed at by those who regard themselves as working in AI.

1. Notice the phrase I used: "regard themselves as working in AI". There is considerable diversity of opinion even over the matter of what properly constitutes subject matter for AI research.
2. This is basically the approach taken by the author of our text in the introduction to part 1, where he defines AI as "the collection of problems studied by artificial intelligence researchers".

B. People work in AI to accomplish a variety of goals.

1. One possible goal is to UNDERSTAND human intelligence (or perhaps intelligence in general - human or otherwise.) In this sense, AI is closely related to psychology. (In fact, one of the reserve books for the course is catalogued in the psychology section of the library.)
  - a. In any science, one good way to check out the validity of a theory is to build a model from the theory and test it to see if it performs as one expected. Such a model can also help to refine the theory itself.

Thus, a theory about the nature of intelligence might be tested by incorporating it in a computer program and seeing if the program's behavior matches the behavior of living creatures. In some cases, this has been successfully done.
  - b. Some writers refer to this approach as COGNITIVE SIMULATION or SIMULATION-MODE AI.
  - c. Today, there is an emerging discipline known as "Cognitive Science" involving researchers from the fields of psychology, philosophy, linguistics, neuroscience and computer science. This discipline aims at understanding intelligence wherever it is found.
  - d. The author of one widely-used AI text (Nilsson), refers to this as "AI as Science".
2. A second possible goal for AI work is to IMITATE human intelligent activity, without necessarily understanding how humans do what they do. Note that an AI worker who succeeds in some aspect of goal 1 would also be successful in terms of this goal; but the reverse would not necessarily be true.
  - a. Successful computer chess programs like Deep Blue are a good example of this approach. A human chess player typically considers only a small number of positions for any move, while a computer player uses the speed of the computer to consider tens of 100's of 1000's. But human masters have an intuitive sense of which positions ought to be considered that programs lack.
  - b. Some writers reserve the term "Artificial Intelligence" for this approach, as differentiated from Cognitive Simulation - i.e. the stress is on "artificial". Alternately, we may speak of this approach as PERFORMANCE MODE AI as distinguished from simulation mode AI: the emphasis is on performing like a human, not simulating how he does what he does.

ex: Contrast human efforts to fly like the birds. Daedalus and Icarus, Leonardo da Vinci etc = simulation mode; modern airplanes = performance mode.

- c. Nilsson refers to this as "AI as Engineering".
- d. This approach can be further analyzed in terms of ultimate goals.
  - i. Some whose goal is to produce systems that imitate intelligence have, as an ultimate goal, producing systems that are truly intelligent in the most general sense of the word.
  - ii. Others are content with developing systems that perform at a high level in a specific, well-defined domain, without regard to whether such systems might some day be developed into human-level intelligence.

This is sometimes referred to as APPLIED AI or KNOWLEDGE ENGINEERING.

- e. One focus of applied AI work concerns improving the interface between human users and computer systems through the use of natural language for both input and output. We will work with some examples of this in course projects. (This is a good domain for us to work with, because it is something we are all familiar with and requires no special hardware.)

At this point, I'd like to show a brief demonstration of two programs, one of which you will work with later in the course.

- i. Demo: blocks program.

```
Start prologj from the command line.  
[blocks].  
go.  
Demo one or more commands.  
quit.
```

- ii. Demo: isa hierarchy program

```
Restart prologj.  
[myproj3].  
proj3.  
Demo with "Rocco/Alexander" conversation.  
quit.
```

- f. Another example of applied AI work is the field of robotics, which has generated a number of important problems for applied AI research - including:
  - i. Vision.
  - ii. Speech understanding (as opposed to processing of natural language ASCII text)
  - iii. Planning movement  
etc.

3. Summary: Several different goals have historically been subsumed under the heading of "AI":
  - a. Producing a machine that embodies theories about how humans think in such a way as to effectively simulate human intelligence: cognitive simulation.
  - b. Producing a machine whose performance matches or rivals that of humans, without regard to whether or not its internal workings resemble those of the brain: performance mode AI.
  - c. Making computers more useful or useful to more people
    - i. Producing a system that exhibits human-like expertise in a specific domain (and so can serve to substitute for scarce human experts): knowledge engineering.
    - ii. Improving the friendliness of the interface between humans and computers - thus making computer technology available to a broader spectrum of users.
    - iii. Supporting work in robotics.
4. Patrick Henry Winston, then the Director of the AI Lab at MIT, defined the goals of AI this way:
  - "One central goal of Artificial Intelligence is to make computers more useful."
  - "Another central goal is to understand the principles that make intelligence possible."

### III. Approaches to AI

---       -----   --   --

- A. As the author of one widely-used text (Nilsson), points out, historically there have been two quite distinct approaches to AI.
- B. One approach, which he calls the "TOP-DOWN" or "SYMBOLIC" approach, is guided by the Physical Symbol System Hypothesis elucidated by Newell, Simon, and Shaw in an article we will read and discuss in a few weeks.
  1. This approach equates intelligence with symbol manipulation, and contends that wherever intelligence is found, symbol manipulation is taking place. Since digital computers are ultimately symbol manipulators, there is no reason why a sufficiently complex digital computer cannot be programmed to be intelligent.
  2. Historically, this approach has guided much of the work in AI - to the point where one writer describes it as GOFAI ("Good Old Fashioned AI").
- C. A second approach, which Nilsson calls the "BOTTOM-UP" or "SUBSYMBOLIC" approach, looks to build up intelligent systems from very simple elements functioning in a complex environment (much like the human brain is composed of a large number of fairly symbol cells called neurons.)

1. Historically, this approach arose in parallel with the first approach, but then went into eclipse for a long time.
2. Recently, interest in this approach has been revived in the form of several lines of research:
  - a. Neural networks.
  - b. Genetic algorithms.
  - c. "Embodied" intelligence.
- D. In terms of the structure of this course, most of our material will be within the symbolic framework - because, historically, that's where most AI work has been done.

#### IV. Historical Survey

-- -----

- A. To close out this introduction to AI, we will briefly survey the history of the field: a period of roughly 50 years. Our goal here is to note some names of people and programs that occur again and again in the literature.
  1. As we have noted, the philosophical roots of AI, go back through over 2000 years of Western history. We will not cover that ground here, but the issues involved do become important when one gets into philosophical discussions about the potential capabilities and limitations of AI.
  2. For further study in the history of AI, you may wish to consult McCorduck, Pamela. *Machines Who Think* (NY: W.H. Freeman, 1979; 2nd ed Natick, MA: A.K. Peters, 2004). (By the way, note in the title - machines WHO think.)
- B. The idea of building an "artificial man" is an old one.
  1. The Greek myth of Pygmalion and Galatea.
  2. Ancient and medieval stories about talking statues etc.
  3. Rabbi Loew of 16th century Prague was said to have created an artificial man, Joseph Golem, to spy on the Gentiles to protect the Jewish community and to perform chores around the temple in his off hours. Interestingly, several leading workers in modern AI come from families regarding themselves as descendants of Rabbi Loew - including Marvin Minsky and Joel Moses of MIT, as well as John VonNeumann and Norbert Wiener.
  4. In the 19th century, Von Kempelen built a "chess-playing machine" which aroused interest in Europe for decades until it was shown that the game was actually played by a human skillfully concealed inside it.
  5. The darker side of man-made men is explored in Mary Shelley's novel *Frankenstein*.

6. In 1923, Karel Cepak brought his play RUR (Rossum's Universal Robots) to London. This play is the origin of the term "robot", derived from a Slavic word meaning "slave".
7. In 1950, Isaac Asimov formulated his "three laws of robotics" which have occupied a prominent place in subsequent science fiction:
  - a. A robot may not injure a human being, or through inaction allow a human to come to harm.
  - b. A robot must obey the orders given it by a human being except when those orders would conflict with the First Law.
  - c. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
8. Recent movies continue to explore this issue - e.g.
  - a. The Thirteenth Floor
  - b. The Bicentennial Man
  - c. Artificial Intelligence (Spielberg)
  - d. The various movies in the Matrix series

#### MENTION FILM SERIES

- C. The field of AI - as we know it - was born in 1956. A summer conference held at Dartmouth College (hence known as the Dartmouth conference) brought together a number of scientists who had developed an interest in artificial intelligence. As an outgrowth of this conference, research projects were established at three universities:
  1. MIT - under the leadership of Marvin Minsky
  2. Carnegie-Mellon University - under the leadership of Herbert Simon and Allan Newell, later joined by Cliff Shaw.
  3. Stanford - under the leadership of John McCarthy (formerly of MIT)

These, plus Stanford Research Institute, were the major centers of AI work in the early years, and remain leading centers today. Many of the leading workers in the field were trained at one of them.

Note that the approach coming out of the Dartmouth Conference was based on the symbolic approach.
- D. At about the same time, a second stream of work developed, working with perceptrons, which were devices modelled after the neurons in the brain (a subsymbolic approach).
  1. Perceptron networks were intriguing because they could LEARN to solve problems - they were TRAINED by examples, rather than being programmed.

2. Perceptron work went into eclipse following the publication of a paper by Minsky and Pappert showing that there were severe limits to what perceptron networks of the sort being used could do.
  3. In the 1980's, this approach became viable again as the result of additional discoveries which overcame one of the key limitations of the networks that the Minsky and Pappert book critiqued. (Neural networks).
- E. Early in its history, the new field saw a number of rather spectacular successes:
1. Newell and Simon developed a program known as Logic Theorist that was able to generate proofs for the theorems in Whitehead and Russell's Principia Mathematica. One of the proofs used a non-conventional "creative" method which surprised even the program's authors, and delighted Russell.
  2. A program by Joseph Weizenbaum of MIT, called ELIZA, emulated a Rogerian psychologist in dialogue with a patient so successfully that human users would sometimes reveal their deepest thoughts to it (and some may even have mistakenly thought they were talking with a human!)

Demo:

```
cd Project 1 - mine
Restart prologj.
[myproj1].
eliza.
```

(Actual sample dialogue from Weizenbaum book)

```
Men are all alike.
They're always bugging us about something or other.
Well, my boyfriend made me come here.
He says I'm depressed much of the time.
It's true; I am unhappy.
I need some help, that much seems certain.
Perhaps I could learn to get along with my mother.
My mother takes care of me.
My father.
You are like my father in some ways.
You are not very aggressive, but I think you don't want me to notice that.
You don't argue with me.
You are afraid of me.
My father is afraid of everybody.
Bullies.
```

- a. This led one psychologist to seriously suggest that computers might help alleviate the shortage of trained psychotherapists.
  - b. It also led Weizenbaum to question his own work in AI and to become one of the field's leading critics, as you saw in the reserve reading.
3. A series of chess playing programs by various authors progressed to expert-level play. Today, there is an annual computer-chess championship at which computers play at the master level, and the rules

of the US Chess Federation have been rewritten to include provision for handling computer programs as "members" of the federation and contestants in regular tournaments. Just a couple of years ago, a chess playing program defeated the world champion of chess.

4. Shakey - a mobile robot that could navigate around obstacles etc - was developed at SRI. It was the subject of a Life Magazine article entitled "Meet Shakey: The First Electronic Person" (1970). Today, special-purpose robots are widely used in various industries.

F. However, there were failures as well.

1. Newell and Simon sought to extend Logic Theorist to handle any kind of problem solving, producing a program known as General Problem Solver or GPS. Alas, GPS was not as general as they had hoped; but its design influenced many subsequent projects though it is now regarded as outdated.
2. One of the major focuses of early AI work was on automated translation of foreign language texts into English or vice-versa. This project was of interest to the DOD, and so received very sizeable funding (which may have accounted for the degree of interest on the part of researchers!) However, despite lofty promises the work was largely unsuccessful, and the failure led to a period when AI was in disrepute, which lasted until fairly recently.

(One of the foibles of this project came from feeding the phrase "the spirit is willing but the flesh is weak" into an English to Russian translator, then feeding the result back into a Russian to English translator to yield "the vodka is strong but the meat is rotten.")

G. The early successes and failures of AI seem to have typified a pattern that has continued to this day.

1. There have been many successful systems that have performed well in in specially-conceived "micro-worlds", or in limited problem domains.
2. However, whenever there have been attempts to generalize these successes to more realistic domains or more general domains, at some point the attempt to generalize has broken down. (The successes failed to "scale up".)

Note: you will become painfully aware of this when you work with the two natural language programs we looked at earlier

3. This phenomenon of systems "breaking" when generalized can lead to one of several possible conclusions:
  - a. We can ultimately produce generally-intelligent systems, but we just haven't figured out how to do it yet - we need to keep trying.
  - b. The whole goal of producing generally-intelligent systems is wrong-headed, and people should stop trying to do it.

- c. The goal is correct, but the means of going about it (typically via symbolic AI) is wrong. What is needed is an altogether different approach.

Many writers have taken this tack, going down different lines. The paper by Rodney Brooks we will read and discuss toward the end of the course is one such argument.

4. Some of the philosophical articles we will read and discuss deal with issues related to this.