ARTIFICIAL INTELLIGENCE

The Very Idea

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science, Bioinformatics & Genomics and Neuroscience
Director, Artificial Intelligence Research Laboratory
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Core ideas behind AI






- Working hypothesis of AI – Hobbes, Leibniz, Boole, Turing
  - Cognition is, or at least can be modeled by Computation
- Computation is tantamount to executing an algorithm or a finite sequence of instructions on a Turing Machine
  - Transforming one sequence of symbols into another sequence of symbols according to precise step-by-step instructions
- Church-Turing Thesis: Any function that is computable can be computed by a Turing Machine
- Working hypothesis of AI + Church-Turing Thesis:
  - Cognition can be modeled by algorithms executable on a Turing machine

Core ideas behind AI

- The working hypothesis of AI implies a computational theory of minds
- We will have a theory of
  - Problem-solving if we can devise algorithms that can solve problems
  - Game-playing if we can devise algorithms that play games
  - Reasoning if we can devise algorithms that reason from assumptions to conclusions
  - Learning if we can devise algorithms for learning from experience
  - Language if we can devise algorithms that effectively communicate using language
  - Cooperation if we have a computational model of multi-agent collaboration
  - Creativity if we can devise algorithms that exhibit creativity

Core ideas behind AI

- David Marr: Computational theory of minds implies levels of analysis

Levels of analyses

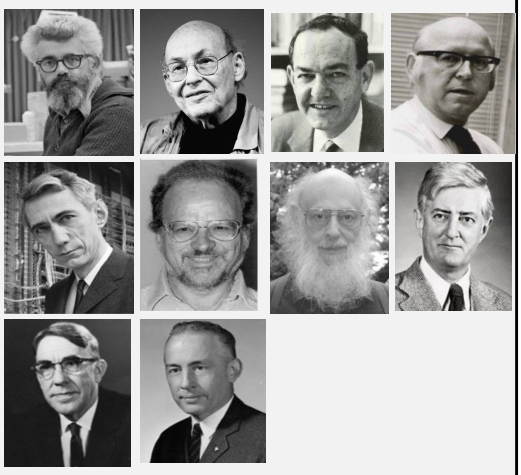
Computation	1	What is computed – functional specification
Algorithm	2	How it is computed – algorithm
Implementation	3	How exactly – program or implementation

The focus of this course

- In this course, our focus is on
  - what is being computed by machines that behave as if they have minds
  - how it is computed by an algorithm
  - but not necessarily how exactly it is being computed – that is, realized by a program
- Church-Turing thesis says that a given algorithm can have many implementations
- Details of implementation of the algorithms of the mind on specific physical substrates is left to other courses
  - AI – implementation using programs written in a specific programming language executed on computers
  - Neuroscience – implementation in brains

Artificial Intelligence – The very idea

- John McCarthy organized a meeting of the group at Dartmouth in the summer of 1956
- "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."



A brief (recent) history of AI

- Birth of artificial intelligence (1956)
- Early demonstrations of artificial intelligence and the publication of Computers and Thought (1959)
- 1960-1970
  - Optimism fueled by early success on some problems thought to be hard (e.g., theorem proving)
  - Slow progress on many problems thought to be easy (e.g., vision, language);
  - Fragmentation of AI into sub-areas focused on problem-solving, knowledge representation and inference, vision, planning, language processing, learning, etc.

A brief (recent) history of AI

- 1970-mid 1980s
  - Advent of practical tools such as expert systems
  - Codification of expert knowledge in specific domains into programs
  - Realization of the difficulty of knowledge engineering
  - Snow shift in focus to systems that learn from experience

A brief (recent) history of AI

- Mid 1980s-mid 1990s
  - Renewed interest in biologically inspired neural network models
  - Modest success on vision, text-to-speech conversion
  - Emergence of machine learning as a promising and practical alternative to knowledge engineering
  - Progress in the various sub-fields of AI directs attention on the design of intelligent agents that exhibit multiple aspects of intelligence

A brief (recent) history of AI

- Mid 1990s-2010
  - The advent of the World-Wide-Web, big data, and computing advances lead to AI systems trained on massive amounts of data
  - Major breakthroughs in learning theory offer insights that lead to practical advances, e.g., kernel machines, in machine learning
  - Large data coupled with machine learning yield breakthroughs in many applications: information retrieval, computer vision, information extraction, robotics, among others
  - Successes in design of intelligent agents shifts attention to multi-agent systems – inter-agent communication, coordination, and multi-agent organizations

A brief (recent) history of AI

- 2010-2020
  - Large data sets and hardware advances, e.g., graphical processing units, and deep learning algorithms lead to rapid progress on computer vision, natural language processing
- 2020-present
  - Large language models capture public interest and imagination
  - Advances in powerful AI systems that could automate aspects of human intellectual work highlight ethical concerns
  - Goals of AI shift from automating intelligent behavior to augmenting and extending human intellect and abilities

Goals of AI

- For some, the long-term dream of AI is to build machines that are intelligent, self-aware, conscious, and autonomous in the same way that people like you and I are
- This is the science fiction vision of AI you see in movies or read about in popular press
- We don't really understand
  - What such it takes to build machines that are intelligent in all the way humans are
  - Or how we can know when we have succeeded
    - Recall the various tests of intelligence
- There is little consensus on whether human-like AI is feasible or even desirable

Scientific goal of AI

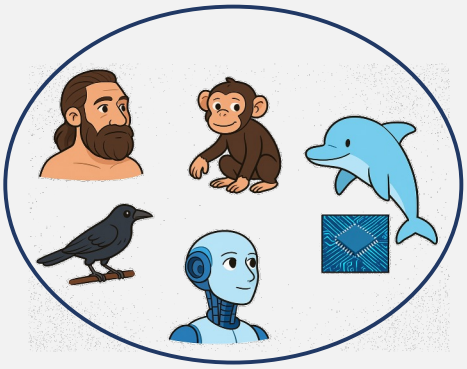
The central scientific goal of AI is to understand intelligence by building and studying computational models of intelligence

- We will have a theory of
  - Problem-solving if we can devise algorithms that can solve problems
  - Game-playing if we can devise algorithms that play games
  - Reasoning if we can devise algorithms that reason from assumptions to conclusions
  - Learning if we can devise algorithms for learning from experience
  - Language if we can devise algorithms that effectively communicate using language
  - Cooperation if we have a computational model of multi-agent collaboration
  - Creativity if we can devise algorithms that exhibit creativity

Scientific goal of AI

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

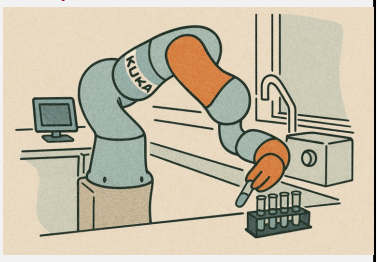
- The intelligence of humans and animals provide existence proofs or examples of designs for intelligent systems found in nature



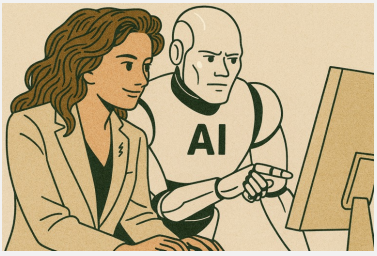
- Exploring this design space entails
  - characterizing the existing designs
  - conceiving and evaluating alternative designs
    - with the desired characteristics and performance
    - Under various design constraints

Practical goal of AI

- AI is is the enterprise of the automating tasks that are believed to require intelligence when performed by humans
  - proving theorems
  - planning trips
  - recognizing faces
  - diagnosing diseases
  - designing computers
  - composing music
  - discovering scientific laws
  - proving mathematical theorems
  - playing chess, writing stories
  - teaching physics
  - negotiating contracts
  - providing legal advice



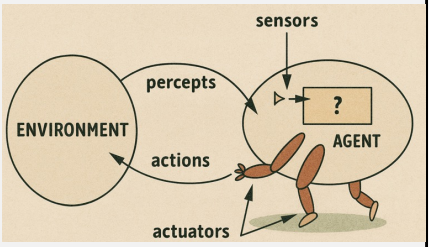
Practical goal of AI



- AI is about augmenting and extending human intelligence, problem-solving and creativity
  - An AI physician's assistant helps medical practitioners make better decisions
  - A search engine augments human memory
  - Natural language translation systems help people communicate across linguistic barriers
  - An AI writer's assistant can act as a muse
  - AI-powered scientist's assistants can help identify promising hypotheses to pursue, optimal experiments to run, and help analyze and interpret experimental results

Practical goal of AI

- AI is about the analysis and synthesis of agents that behave as if they have minds
- An agent is an entity that senses or perceives the environment acts on the environment



- The agent is considered intelligent
  - to the extent that its actions are appropriate given its circumstances and perceptual and computational limitations
  - and its behavior is adaptive to changes in its circumstances and in the environment
  - and its behavior improves with experience

Success measures

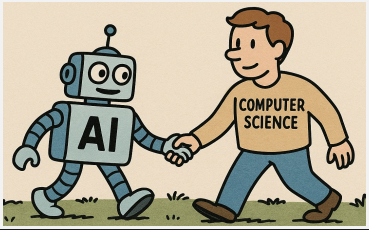
Not everything that can be counted counts and not everything that counts can be counted

- Success measures depend on the goals
  - Successful automation of some aspect of intelligent behavior simply requires that the AI system perform the tasks at hand with a level of competence that makes it useful in practice
  - Effectively augmenting or extending the capabilities of a physician in an emergency room calls for complementing and not duplicating what the physician is good at so that the human-AI team achieves outcomes that are superior to what either could on its own
- Choosing appropriate measures is critical to progress in AI

Relationship pf AI to other disciplines

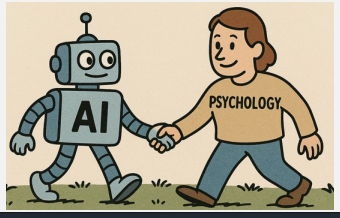
AI in relation to computer science

- AI has a special relationship to computer science because computation : mind :: calculus : physics
- AI takes advantage of advances in computer science
- AI contributes concepts and tools for computer science
- AI stimulates advances in computer science
- AI is not a subset of computer science.
- Computer science is not a subset of AI



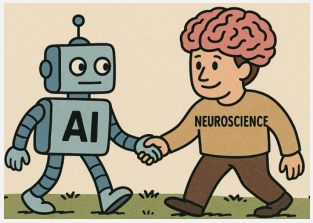
AI in relation to psychology

- AI is often seen as a sibling of psychology
- Psychology is concerned with studies of human and animal behavior
- AI is concerned with computational models or artifacts that exhibit intelligent behavior
- AI is not committed to human-like mechanisms or human-like implementation of such mechanisms
- Computational models from AI influence research in psychology
- Findings and insights from psychology inform the design of AI models



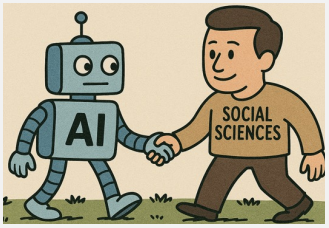
AI in relation to neuroscience

- AI has deep connections to neuroscience
- Neuroscience is about brains
- AI is about computational theories of brain function
- AI is informed by or at least inspired by findings in neuroscience
- Advances in machine learning offer new perspectives on old questions in neuroscience



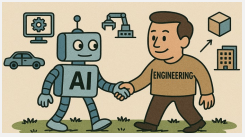
AI in relation to social sciences

- Insofar as AI is concerned with the design of intelligent agents that
  - act rationally
  - interact with other agents
- AI draws on ideas from economics, game theory, organizational theory, decision theory, and other areas of social sciences
- AI systems can inform studies that drive advances in the social sciences



AI in relation to engineering

- Insofar as AI is concerned with the design of intelligent artifacts, it both contributes to, and draws on advances in engineering
- AI advances have resulted in practical tools for
  - configuring computer systems
  - Diagnosing faults in machinery
  - software agents that scour the Internet for information on demand
  - Intelligent systems for planning and scheduling
  - Computer-aided design tools in many engineering disciplines
  - Self-driving automobiles, Smart buildings, smart robots, etc.



Some lessons from AI

Easy problems for AI

- Arithmetic, algebra, logic
- We have precise algorithms to instruct computers

Somewhat hard problems for AI

- Board games like Chess, Backgammon, etc.
- Effective play requires looking ahead many moves – search space too large – need heuristics

Moderately hard problems for AI

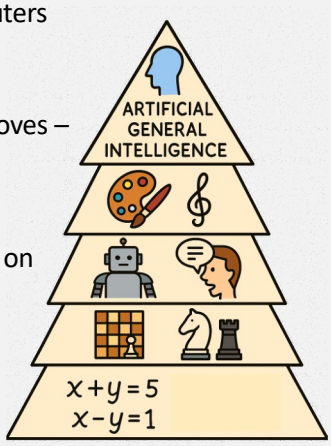
- Vision, language etc.
- No known algorithm – train machine learning on large data

Even harder problems for AI

- Creativity

Hardest problem for AI

- Artificial General Intelligence



Moral Imperative of AI

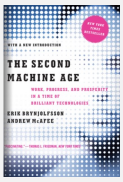
Example: Automated driving

- Each year, 1 million people die in auto accidents worldwide
- Over 90% of auto accidents are caused by human error
- Automating driving could save countless lives, reduce injuries
- Automating driving could result in significant job loss
- Should we automate driving?

Similar examples abound in other areas

- AI and robotics have unleashed the 4th industrial revolution
- Majority of jobs we have today will be lost or dramatically changed due to advances in AI, robotics, and automation

AI and economy



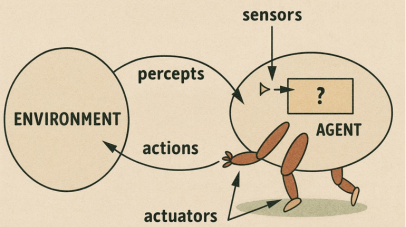
- AI will disrupt all areas of life
  - Transportation, accounting, healthcare, journalism, banking
  - What will happen to the workers who once occupied those jobs?
  - Will new jobs be created?
  - How can workers get trained for the new jobs?
  - If AI automates almost all work, what will humans do?
  - Do we need new economic models?

Moral Imperative of AI

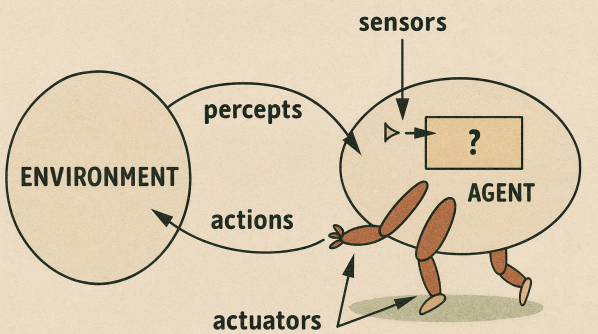
- How can we maximize the benefits of AI while minimizing the harms of AI?
- How do we build AI systems that can augment and extend human abilities?
- How do we build AI systems that can explain their decisions?
- How do we ensure that AI systems are safe?
- How do we ensure that AI systems are accountable?
- How do we ensure that AI systems are fair?
- How do we ensure that AI systems adhere to human ethical and social norms?

Intelligent agents – A practical avenue for AI advances

An agent can
- perceive its environment through its sensors
- represent aspects of its environment
- reason and predict consequences of its actions
- act on its environment through its effectors
- make rational choices
- learn from experience
- communicate with other agents
- interact with other agents
- display autonomy
- display creativity
- exhibit purposeful behavior

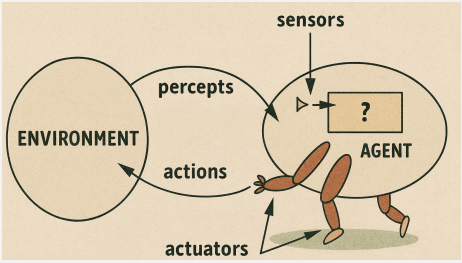


Agents and environments



- Agents include humans, animals, robots, soft-bots, thermostats, etc.
- An agent senses the state of the environment through its sensors, obtaining *percepts*
- The *agent function* maps percept sequence to actions

Agents and environments

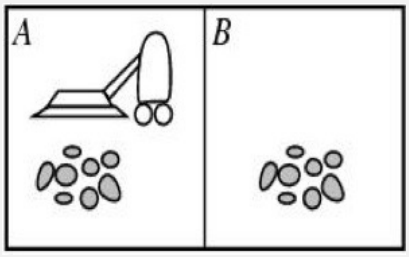


- The *agent function* f is encoded by the *agent program*.
- The agent program runs on a physical *architecture* (robot, human, etc.) to produce f.

AI Thermostat

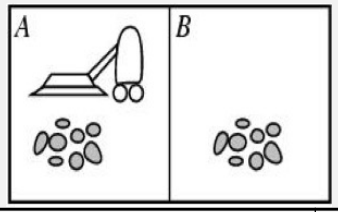
- One of the simplest agents we can imagine is a simple thermostat
- The thermostat senses the ambient temperature using its temperature sensor
- If the sensed temperature is greater than the preset temperature, it turns on the air conditioner
- If the sensed temperature is less than equal to the preset temperature, it shuts off the air conditioner if it is on
- If the sensed temperature is less than the preset temperature, it turns on the heater
- If the sensed temperature is greater than equal to the preset temperature, it shuts off the heater if it is on

The vacuum-cleaner agent



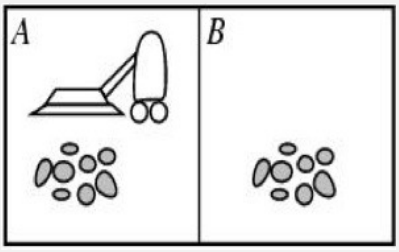
- Environment: rooms A and B
- Percepts: [location, content] e.g. [A, Dirty]
- Actions: left, right, cleanup, and no-op

The vacuum-cleaner agent



Percept sequence	Action
[A,Clean]	Right
[A, Dirty]	Cleanup
[B, Clean]	Left
[B, Dirty]	Cleanup
[A, Clean],[A, Clean]	Right
[A, Clean],[A, Dirty]	Cleanup

The vacuum-cleaner agent



if *status is Dirty* then *Cleanup*
else if *location is A* then move *Right*
else if *location is B* then move *Left*

How does agent ought to behave?

- How agents ought to behave is a topic of debate in moral philosophy
- AI adopts a form of consequentialism
  - whether or not an action is the right one depends on its consequences relative to what we value
- Example:
  - Most people would agree that lying is wrong
  - But if telling a lie would help save a person's life, consequentialism says it's the right thing to do

How can we relate actions to consequences?

- An agent it generates actions in response to the percepts it receives, thereby causing the environment to go through a sequence of states
- Example: the thermostat agent would ensure that the room temperature is maintained to its preset value
- If the consequence of the agent's actions is desirable, then the agent has performed well
- How does an agent know the desirability of its actions?
  - Performance measure
- An agent's performance in its environment is evaluated by the chosen performance measure

Humans versus AI agents

- Humans have desires and preferences of their own which determines their performance measure
- If your goal is to become a world's leading expert in AI, you should be assessed according to whether you become a leading AI expert
- If your brother's goal was to win an Olympic gold medal in gymnastics, he should be assessed according to whether he wins an Olympic gold medal

Humans versus AI agents

- Machines, unlike humans, do not have desires and preferences of their own
- The performance measure for an AI agent must be, initially at least, specified by
  - the AI agent's designer, or
  - its users, or more broadly,
  - the society at large

  ``If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively, we had better be quite sure that the purpose put into the machine is the purpose which we really desire.''
- How do we ensure that the AI system adheres to human and societal values?

Specifying good performance measures is hard

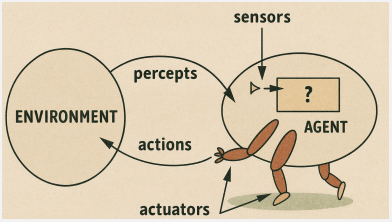
- Specifying the performance measure to reflect precisely how the agent out to behave from an individual or societal point of view is highly non-trivial
- Consequently, the more autonomous and more powerful an AI agent is, the greater the concern about ensuring that its performance measure is aligned with human and societal values

.

Specifying good performance measures is hard

- When AI agents are designed to serve the needs of multiple users,
  - we end up with a piece of software, copies of which will serve different individuals
- We cannot possibly anticipate in advance the goals and preferences of each individual
- Even if we could, custom-designing of agents for each individual is likely to be highly impractical
- AI agents should
  - accommodate uncertainty about the actual performance measure against which they will be assessed and
  - refine the measure over time
  - through their interactions with their respective users

How an agent should behave



What an agent ought to do at any given time depends on:

- The performance measure that defines the criterion of success against which the agent is evaluated
- The agent's prior knowledge of its environment
- The actions that the agent has at its disposal
- The agent's percept sequence up to that time

Rational Agents

- For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given
  - the evidence provided by the percept sequence
  - whatever built-in knowledge the agent has and
  - the actions at its disposal
- Examples of performance measures
  - the amount of dirt cleaned within a certain time
  - how clean the floor is
  - the amount of dirt cleaned per unit of electricity used
- What is your performance measure?
- Who decides what the performance measure should be?
  - Internal drives
  - External rewards

Rationality

- What is rational at a given time depends on:
  - The performance measure
  - What the agent knows
  - The actions that the agent can perform
  - What the agent has observed (through its sensors)

Rationality ≠ omniscience

- You are stopped at a red light at an intersection
- You watch it turn yellow and then green
- Being a rational driver who knows the traffic rules, you take your foot off the breaks and start driving across the intersection
- Meanwhile, at 30,000 feet in the air,
  - the cargo door falls off an airliner in flight,
  - comes crashing down, and
  - lands right on top of you,
  - pinning you down while
  - other vehicles crash into you causing a pileup
- Was your behavior rational?
  - If so, why?
  - If not, why not?

Rationality ≠ omniscience

- An omniscient agent knows everything there is to know
- A rational agent knows only
  - the current state of the world seen through its sensors and
  - the expected state of the world resulting from its action
- Your behavior at the intersection was rational based on what you knew
  - You are not omniscient
  - You couldn't possibly be expected to know that the cargo door of the airplane flying overhead would come crashing down on you

Rationality ≠ clairvoyance

- A clairvoyant agent knows the actual effects of each actions before it is performed
- A rational agent knows only the expected effect of its action

Rationality ≠ perfection

- Rational agent maximizes
  - *expected* performance
- Perfect agent maximizes
  - *actual* performance

Is the thermostat agent rational?

- The performance measure awards one point for each hour the temperature is maintained within ± 2 degrees Celsius from the preset temperature
- What the agent knows:
  - The preset temperature in degrees Celsius.
  - If the air conditioner is turned on, it cools the room, causing a decrease in the temperature;
  - if the heater is turned on, it heats the room, causing an increase in the temperature;
  - The heater, once turned on, will remain on until it is turned off;
  - The air conditioner and the heater cannot both be on at the same time.
- The available actions are to turn the air conditioner on or off, and the heater on or off, or do nothing (leave things the way they are).

Is the thermostat agent rational?

- The thermometer or room temperature sensor is operational and provides the correct temperature readings
- An air conditioner status sensor tells the thermostat whether the air conditioner is on
- The heater status sensor tells the thermostat whether the heater is on
- If the sensed temperature is greater than the preset temperature, the thermostat turns on the air conditioner
- If the sensed temperature is less than equal to the preset temperature, it shuts off the air conditioner if it is on
- If the sensed temperature is less than the preset temperature, it turns on the heater
- If the sensed temperature is greater than equal to the preset temperature, it shuts off the heater if it is on

Is the new thermostat agent rational?

- Suppose the thermostat is additionally equipped with room occupancy sensor that tells it whether the room is occupied by people
- The performance measure awards the agent
  - one point for each hour the temperature is maintained within $\pm$ 2 degrees Celsius of the preset temperature whenever the room is occupied; and
  - -10 points if either the heater or air conditioner are found to be on when the room is unoccupied by people
- Under these circumstances, does the agent function described previously ensure that the agent is rational? Why or why not?
- How can you restore rationality?