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

Goals of AI

The long-term dream of AI is

- To build machines that have the full range of capabilities for intelligent action that people have
- To build machines that are self-aware, conscious, and autonomous in the same way that people like you and I are.''

Michael Wooldridge in "A Brief History of Artificial Intelligence"

- Building machines that are intelligent in all the ways humans are is the science fiction vision of AI you see in movies
- We don't really understand what such an effort entails - 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 the information processing principles and information processing mechanisms that underlie intelligent behavior.
  - Precise descriptions of inputs and outputs of different components of intelligent systems - perception, memory, reasoning, decision-making ..
  - Precise descriptions of their interactions
  - Their algorithmic realizations
  - Theoretical and empirical investigations of alternative theories of intelligence

In short, the goal is to understand intelligence by building and studying computational models of intelligence

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 and conceiving and evaluating alternative designs with the desired characteristics and performance.

Practical goal of AI – Automating intelligent behavior

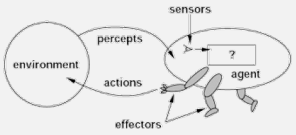
- AI is is the enterprise of the automating tasks that are believed to require intelligence when performed by humans
  - proving theorems
  - planning trips
  - recognizing faces
  - diagnosing diseases
  - designing computers
  - composing music
  - discovering scientific laws
  - proving mathematical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

Practical goal of AI –Augmenting Human Intelligence

- AI is about augmenting and extending human intelligence, problem solving abilities, 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 help identify promising hypotheses to pursue, optimal experiments to run, and help analyze and interpret experimental results

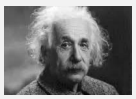
Practical goal of AI – Building intelligent agents

- AI is about the analysis and synthesis of agents that exhibit intelligent behavior
- An agent is an entity that 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 (through learning).



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 practical settings.
  - 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

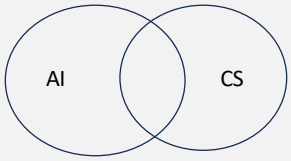
Relationship pf AI to other disciplines

AI in relation to computer science

- AI has a special relationship to computer science due to the centrality of computational theory of mind to AI
- The relationship of AI to Computer Science is like the relationship of physics to mathematics
- Mathematics offers essential tools for physics, but physics is not just mathematics
- The objects of study in mathematics can be entirely abstract with no relationship to the physical world or experimental measurements
- When mathematics is employed in physics, the definitions are abstractions or idealizations of entities in the natural world
- Consequently, physical theories must be verified through experiments

AI in relation to computer science

- Computational theories of intelligence in AI must be verified through experiments
- AI takes advantage of advances in computer science
- AI contributes concepts and tools to 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 studies of computational models or artifacts that exhibit intelligent behavior
- AI is not committed to human-like mechanisms or any particular implementation of such mechanisms
- Computational models from AI have influenced contemporary research in psychology
- Findings and insights from psychology often have informed the design of AI models

Relationship of AI to Artificial Life

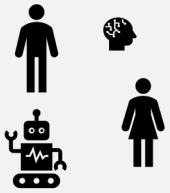
- Artificial life is is to living systems what AI is to cognitive systems
- ALIFE takes a functionalist view of life
- ALIFE uses computational models to study living systems
- In the case of living systems that display intelligent behavior, AI can be seen as a subset of ALIFE

AI in relation to neuroscience

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

AI in relation to social sciences

- AI is concerned with the design of intelligent agents that act rationally and interact with other agents that make up multi-agent organizations
- AI draws on ideas from economics, game theory, organizational theory, decision theory, and other areas of social sciences
- Conversely, algorithmic realizations of these ideas in AI systems offer new insights that can inform 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.

Progress in AI

- First wave of AI applications – Knowledge based systems (1975 - 1985)
- Second wave of AI applications – Neural Networks (1985-1995)
- Third wave of AI applications –networked intelligence (1995-2005)
- Fourth wave of AI applications – Machine learning, computer vision, social media, NLP (2005-2015)
- Fifth wave of AI applications – (big data enabled health, security, AI for social good, AI for augmenting and extending human abilities (2015 – present)

Some lessons from AI – or why is it hard to realize 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 translation, etc.
- No known algorithm – train machine learning on large data

Even harder problems for AI

- Creativity

Hardest problem for AI

- Artificial General Intelligence

AI and society

- AI will disrupt all areas of life
  - Automated driving can reduce accidents and save lives
  - Automated driving will result in significant job loss
  - Similar dilemmas are presented by other areas – 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?
  - How can we design systems and prepare society to best leverage the complementary strengths of humans and AI?
  - How can we ensure that AI systems do not become instruments of injustice, human rights violations, and oppression?
  - How can we maximize the benefits of AI while minimizing its potential for harm?

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
- A majority of the jobs we have today will be lost or dramatically changed as a result of advances in AI, robotics, and automation

Moral Imperative of AI

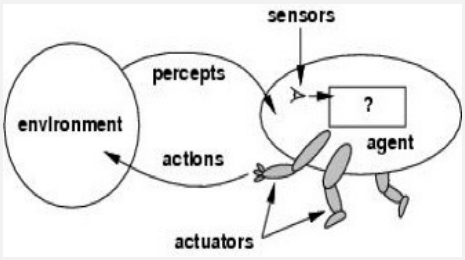
- How can we maximize the benefits of AI while minimizing the harms of AI?
- How do we build AI systems that can augment and extend human abilities?
- How do we build AI systems that can explain their decisions?
- How do we ensure that AI systems are safe?
- How do we ensure that AI systems are accountable?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

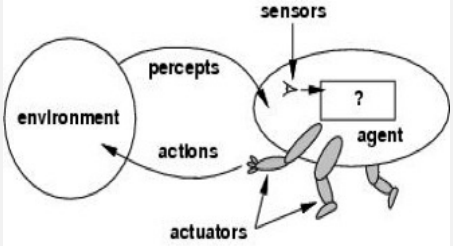
- perceive its environment through its sensors
- represent aspects of its environment and reason with the representation to predict consequences of its actions
- act on its environment through its effectors
- make rational choices
- learn from experience
- communicate with other agents – signals, signs, language
- interact with other agents – cooperation, competition
- display autonomy
- 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

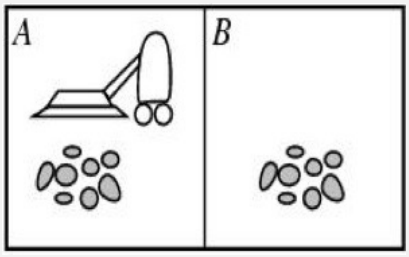


- The *agent function* is encoded by the *agent program*.
- The agent program runs on a physical *architecture* (robot, human, etc.) to produce *f*.

AI Thermostat

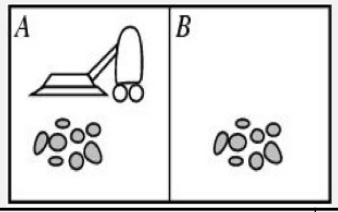
- One of the simplest agents we can imagine is a simple thermostat
- The thermostat senses the ambient temperature using its temperature sensor
- If the sensed temperature is greater than the preset temperature, it turns on the air conditioner
- If the sensed temperature is less than equal to the preset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



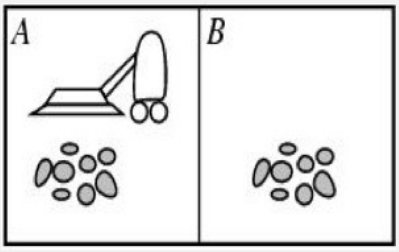
- Environment: rooms A and B
- Percepts: [location, content] e.g. [A, Dirty]
- Actions: left, right, cleanup, and no-op

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 a sequence of actions in response to the percepts it receives, thereby causing the environment to go through a sequence of states
- For example, the thermostat agent in our example, would ensure that the state transitions ensure that temperature in the room is maintained to its preset value
- If the consequence of the agent's actions is desirable, then the agent has performed well
- The desirability of effects of an agent's actions is specified by a 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 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 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 individuals, 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.
- We will need build AI agents that can 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 should an agent 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 the time when it has to choose an action to perform.

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
  - E.g. the amount of dirt cleaned within a certain time
  - E.g. how clean the floor is
  - E.g. 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 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 the everything there is to know about the world
- 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 about the cargo door of the airplane flying overhead come crashing down on you

Rationality ≠ clairvoyance

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

Rationality ≠ perfection

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

Discuss: Is or 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 agent knows the preset temperature in degrees Celsius.
  - If the air conditioner if 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).

Discuss: Is or thermostat agent rational?

- 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.
- The thermometer or room temperature sensor is operational and provides the correct temperature readings.
- An air conditioner status sensor tells the thermostat whether or not the air conditioner is on.
- The heater status sensor tells the thermostat whether the heater is on.

Discuss: Can you identify a setting where the thermostat agent is not rational?

- What if the thermostat is equipped with room occupancy sensor that tells it whether or not the room is occupied by people.
- The performance measure awards it
  - one point for each hour the temperature is maintained within ± 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?