Lesson 03.1: A Brief History of AI

MSML610: Advanced Machine Learning

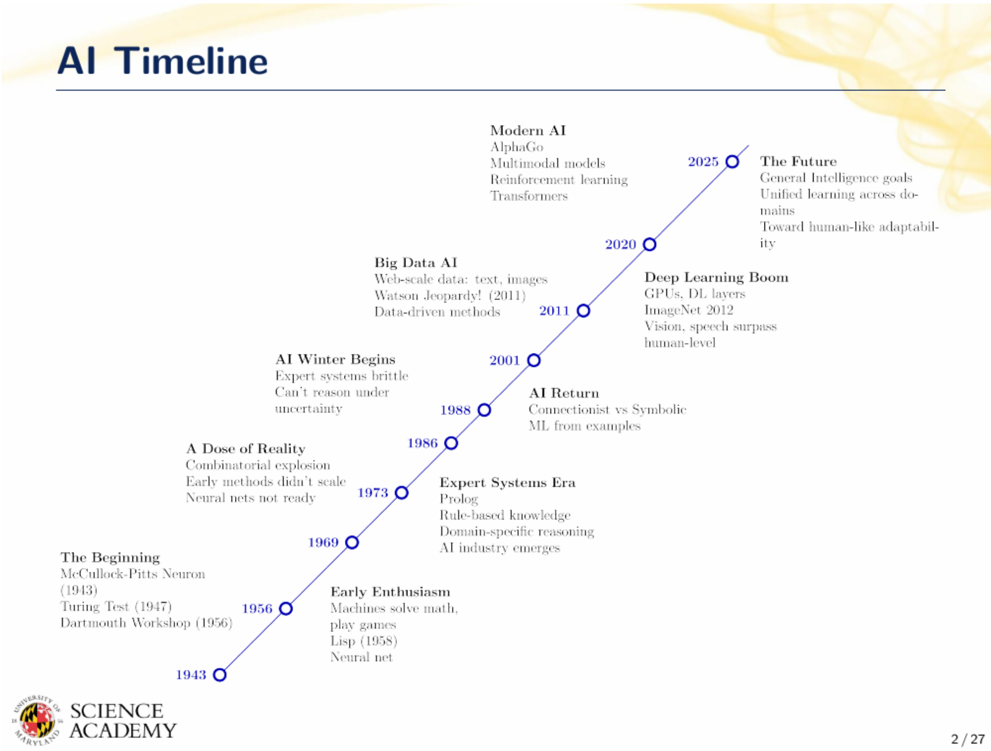
Lesson 03.1: A Brief History of AI

Instructor: Dr. GP Saggese - gsaggese@umd.edu
References: - AIMA (Artificial Intelligence: a Modern Approach), Chap 1

SCIENCE ACADEMY

1 / 27

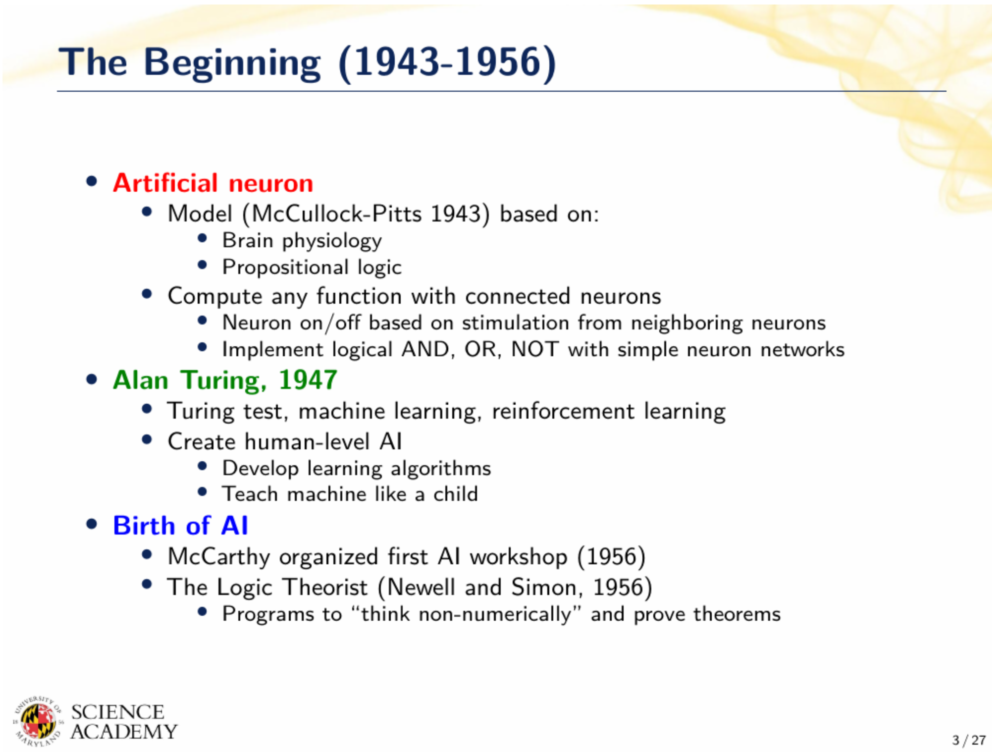
2 / 27: AI Timeline



- **The Beginning (1943-1956)**
  - This period marks the *early foundational work* in AI. The McCullock-Pitts Neuron (1943) was a model of a simple neural network, laying the groundwork for future neural network research.
  - The Turing Test (1947) was proposed by Alan Turing to determine a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.
  - The Dartmouth Workshop (1956) is considered the *birth of AI as a field*, where the term "Artificial Intelligence" was coined.
- **Early Enthusiasm (1958)**
  - During this time, there was excitement about machines solving mathematical problems and playing games, showcasing early AI capabilities.
  - Lisp, a programming language developed in 1958, became a popular tool for AI research.
  - Early neural networks were explored, although they were not yet practical for complex tasks.
- **A Dose of Reality (1969-1973)**
  - Researchers faced the *combinatorial explosion* problem, where the complexity of problems grew exponentially, making them difficult to solve with existing methods.
  - Early AI methods didn't scale well, and neural networks were not yet ready for practical applications.
- **Expert Systems Era (1973-1986)**
  - This era saw the rise of expert systems, which used rule-based knowledge to perform domain-specific reasoning.
  - Prolog, a programming language, was used for developing these systems.
  - The AI industry began to emerge, focusing on practical applications.

- **AI Winter Begins (1986-1988)**
  - Expert systems were found to be brittle and unable to reason under uncertainty, leading to a decline in AI funding and interest, known as the "AI Winter."
- **AI Return (1988-2001)**
  - The debate between connectionist (neural networks) and symbolic AI approaches continued.
  - Machine learning from examples became more prominent, setting the stage for future developments.
- **Big Data AI (2011)**
  - The availability of web-scale data, such as text and images, fueled data-driven AI methods.
  - IBM's Watson winning Jeopardy! in 2011 demonstrated the power of AI in processing large datasets.
- **Deep Learning Boom (2012)**
  - The use of GPUs and deep learning layers led to breakthroughs in AI, particularly in image and speech recognition.
  - The ImageNet competition in 2012 highlighted AI's ability to surpass human-level performance in vision tasks.
- **Modern AI (2020)**
  - Advances in AI include AlphaGo, which defeated a world champion Go player, and the development of multimodal models.
  - Reinforcement learning and transformers have become key technologies in modern AI research.
- **The Future (2025)**
  - The goal is to achieve general intelligence, where AI can learn and adapt across various domains like humans.
  - Researchers aim for unified learning approaches that mimic human-like adaptability.

3 / 27: The Beginning (1943-1956)

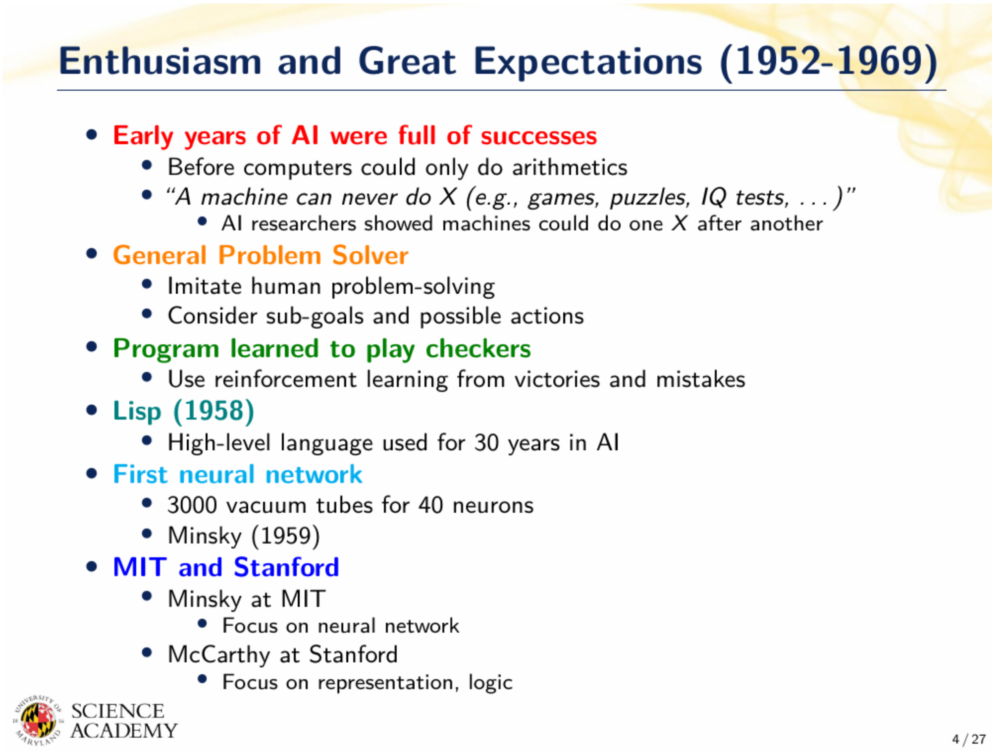


- **The Beginning (1943-1956)**
- **Artificial neuron**
  - The concept of an artificial neuron was introduced by McCullock and Pitts in 1943. They created a model inspired by how the brain works, specifically focusing on brain physiology and propositional logic. This model laid the groundwork for what we now call neural networks.
  - The idea was that by connecting these artificial neurons, you could compute any function. Each neuron would turn on or off based on signals from its neighbors, similar to how neurons in the brain fire. This setup allowed for the creation of simple logical operations like AND, OR, and NOT, forming the basis for more complex computations.
- **Alan Turing, 1947**
  - Alan Turing, a pioneer in computing, introduced several key ideas in 1947. He proposed the Turing Test, a way to measure a machine's ability to exhibit intelligent behavior indistinguishable from a human. He also discussed concepts like machine learning and reinforcement learning, which are crucial in developing AI.
  - Turing envisioned creating AI that could reach human-level intelligence. His approach was to develop algorithms that could learn and improve over time, much like teaching a child.
- **Birth of AI**
  - The field of AI officially began to take shape in 1956 when John McCarthy organized the first AI workshop. This event is often considered the birth of AI as a distinct field

of study.

- During this time, Newell and Simon developed “The Logic Theorist,” a program designed to think in a non-numerical way and prove mathematical theorems. This was a significant step in demonstrating that machines could perform tasks that required reasoning and problem-solving, not just calculations.

4 / 27: Enthusiasm and Great Expectations (1952-1969)

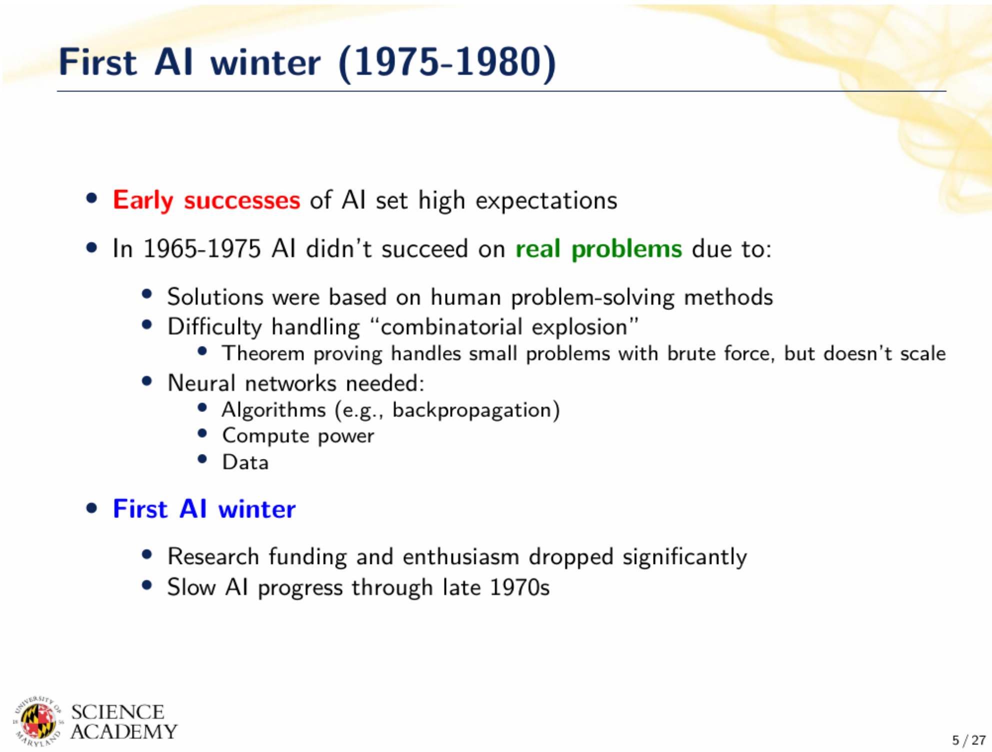


- **Early years of AI were full of successes**
  - During this period, computers were primarily used for basic arithmetic operations. However, the field of Artificial Intelligence (AI) began to show that computers could do much more. Researchers were eager to demonstrate that machines could perform tasks previously thought to be exclusive to human intelligence, such as playing games, solving puzzles, and even taking IQ tests. This was a time of optimism and excitement as AI researchers successfully tackled one challenge after another, proving skeptics wrong.
- **General Problem Solver**
  - The General Problem Solver was an early AI program designed to mimic human problem-solving abilities. It worked by breaking down problems into sub-goals and considering various possible actions to achieve these goals. This approach was groundbreaking as it attempted to replicate the way humans think and solve problems.
- **Program learned to play checkers**
  - One of the notable achievements of this era was the development of a program that could learn to play checkers. This program used a method called reinforcement learning, where it improved its performance by learning from both its victories and mistakes. This was an early example of machines learning from experience, a concept that is central to modern AI.
- **Lisp (1958)**
  - Lisp was a high-level programming language developed in 1958, specifically for AI research. It became the dominant language in the field for the next 30 years due to its flexibility and powerful features that supported symbolic computation, which is crucial for AI applications.
- **First neural network**

- The first neural network was a significant milestone in AI, consisting of 3000 vacuum tubes to simulate 40 neurons. This early attempt to mimic the human brain's structure was led by Marvin Minsky in 1959. Although primitive by today's standards, it laid the groundwork for future developments in neural networks.

- **MIT and Stanford**
  - Two major centers of AI research emerged during this period: MIT and Stanford. At MIT, Marvin Minsky focused on developing neural networks, exploring how machines could simulate human brain functions. Meanwhile, at Stanford, John McCarthy concentrated on representation and logic, aiming to understand how machines could reason and process information like humans. These two approaches were foundational in shaping the future of AI research.

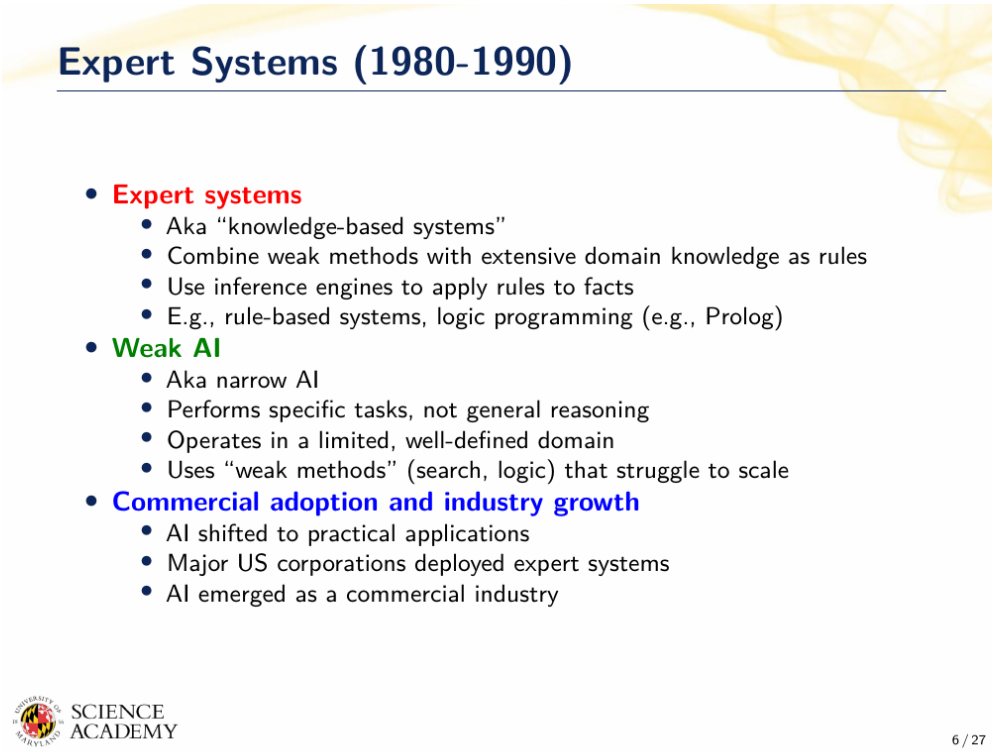
5 / 27: First AI winter (1975-1980)



- **Early successes** of AI set high expectations
  - In the early days of AI, there were some promising developments that led people to believe that machines could soon perform tasks that required human intelligence. This optimism was fueled by initial successes in areas like game playing and simple problem-solving.
- In 1965-1975 AI didn't succeed on **real problems** due to:
  - **Solutions were based on human problem-solving methods**
    * Early AI systems tried to mimic how humans solve problems, but this approach didn't work well for complex, real-world issues.
  - **Difficulty handling "combinatorial explosion"**
    * This term refers to the rapid increase in complexity as problems grow larger. For example, theorem proving could solve small problems using brute force, but this method failed as problems became more complex.
  - **Neural networks needed:**
    * *Algorithms (e.g., backpropagation)*: At the time, the algorithms necessary for training neural networks effectively, like backpropagation, were not yet developed.
    * *Compute power*: The computers of the time were not powerful enough to handle the demands of AI algorithms.
    * *Data*: There was a lack of large datasets needed to train AI models effectively.
- **First AI winter**
  - During this period, the initial excitement around AI faded as it became clear that the technology was not living up to its promises. As a result, funding and enthusiasm for AI research dropped significantly.
  - The late 1970s saw slow progress in AI, as researchers struggled with the limitations

of the technology and the lack of resources. This period is known as the "AI winter" because of the chill it cast over AI research and development.

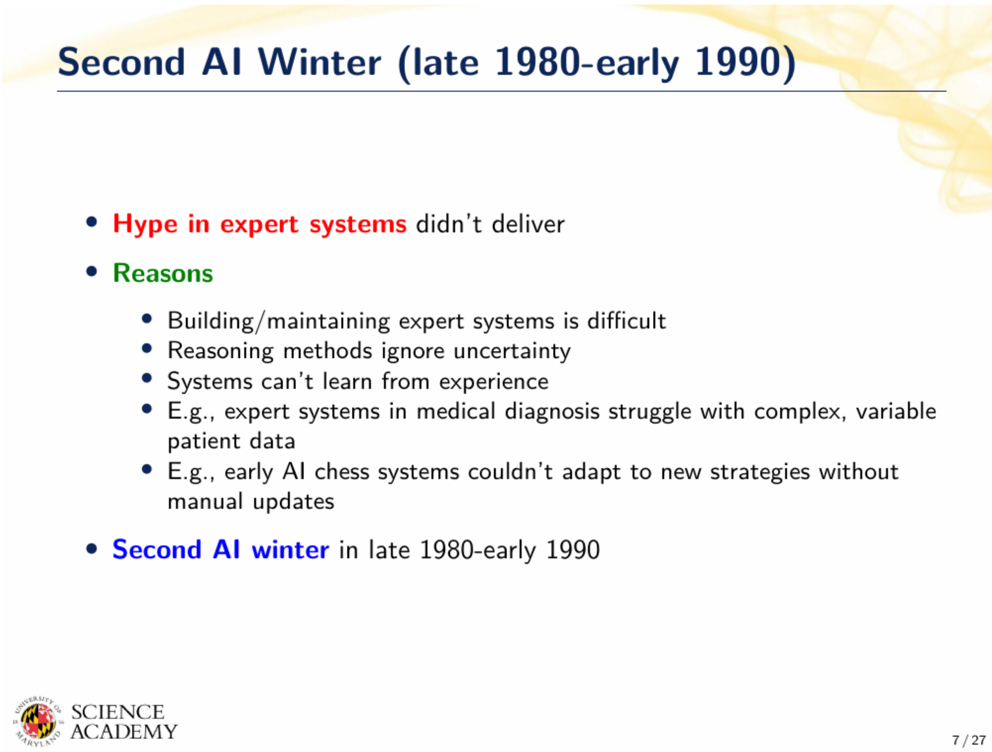
6 / 27: Expert Systems (1980-1990)



- **Expert systems**
  - *Expert systems*, also known as "knowledge-based systems," were a major focus in AI during the 1980s and 1990s. These systems aimed to mimic the decision-making ability of a human expert by using a large set of rules derived from domain knowledge.
  - They combined *weak methods*, which are general problem-solving techniques, with extensive domain-specific knowledge. This knowledge was encoded as rules that the system could apply to specific situations.
  - An *inference engine* is a core component of expert systems. It applies the rules to known facts to deduce new information or make decisions. This process is similar to how a human expert might reason through a problem.
  - Examples include rule-based systems and logic programming languages like Prolog, which were used to create these systems.
- **Weak AI**
  - *Weak AI*, also known as narrow AI, refers to systems designed to perform a specific task or set of tasks. Unlike strong AI, which aims for general reasoning and understanding, weak AI is limited to a particular domain.
  - These systems operate within a well-defined area and use *weak methods* such as search algorithms and logical reasoning. However, these methods often struggle to handle complex or large-scale problems effectively.
- **Commercial adoption and industry growth**
  - During this period, AI began to shift from theoretical research to practical applications. Expert systems were among the first AI technologies to be widely adopted in industry.
  - Major corporations in the United States started deploying expert systems to solve specific business problems, such as diagnosing equipment failures or managing inventory.

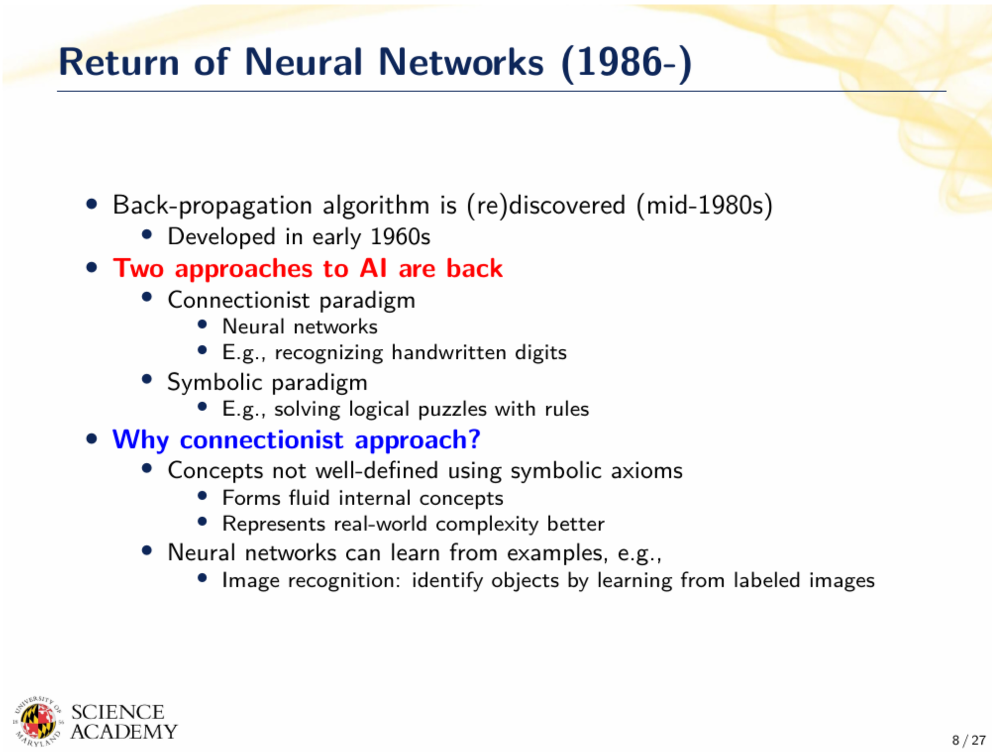
- This era marked the emergence of AI as a commercial industry, with companies investing in AI technologies to gain a competitive edge.

7 / 27: Second AI Winter (late 1980-early 1990)



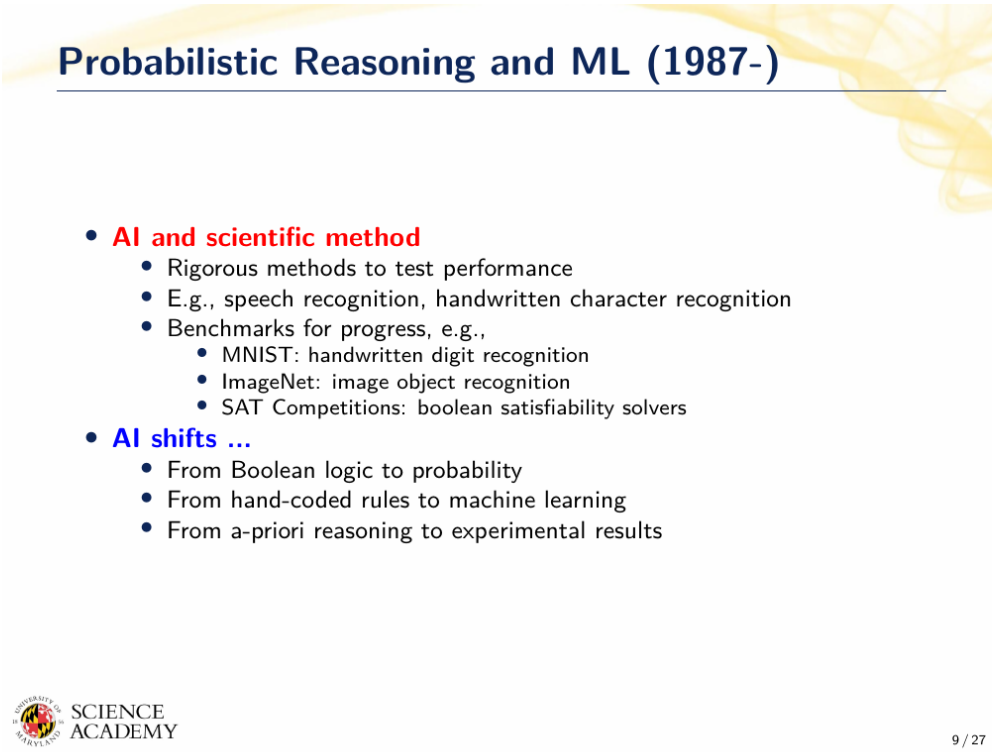
- **Hype in expert systems** didn't deliver
  - During the late 1980s and early 1990s, there was a lot of excitement about expert systems, which are computer programs designed to mimic human expertise in specific areas. However, these systems did not live up to the high expectations set for them.
- **Reasons**
  - **Building/maintaining expert systems is difficult**: Creating these systems required a lot of time and effort from experts to input their knowledge, and keeping them updated was a continuous challenge.
  - **Reasoning methods ignore uncertainty**: Expert systems struggled because they couldn't handle uncertainty well. They were designed to follow strict rules, which made it hard for them to deal with situations that weren't black and white.
  - **Systems can't learn from experience**: Unlike humans, these systems couldn't improve or adapt based on new information or past experiences, limiting their usefulness over time.
  - **E.g., expert systems in medical diagnosis struggle with complex, variable patient data**: In fields like medicine, where patient data can be unpredictable and complex, expert systems often failed to provide accurate diagnoses.
  - **E.g., early AI chess systems couldn't adapt to new strategies without manual updates**: In games like chess, these systems couldn't adjust to new strategies unless a human manually updated them, which was a significant limitation.
- **Second AI winter** in late 1980-early 1990
  - This period is known as the "Second AI Winter" because the disappointment with expert systems led to reduced funding and interest in AI research. The field faced a slowdown as people became skeptical about the promises of AI technology.

8 / 27: Return of Neural Networks (1986-)



- **Return of Neural Networks (1986-)**
  - In the mid-1980s, the *back-propagation algorithm* was rediscovered, which was originally developed in the early 1960s. This algorithm is crucial for training neural networks as it allows them to learn from data by adjusting weights based on errors.
- **Two approaches to AI are back**
  - The *connectionist paradigm* focuses on neural networks, which are inspired by the human brain's structure. These networks are particularly good at tasks like recognizing handwritten digits, where patterns need to be identified from raw data.
  - The *symbolic paradigm* involves using explicit rules and logic to solve problems, such as logical puzzles. This approach relies on predefined rules and is more rigid compared to the connectionist approach.
- **Why connectionist approach?**
  - The connectionist approach is favored because many real-world concepts are not easily defined using strict symbolic rules. Neural networks can form *fluid internal concepts*, making them better suited for representing the complexity of the real world.
  - Neural networks excel at learning from examples. For instance, in image recognition, they can identify objects by learning from a large set of labeled images, improving their accuracy over time as they are exposed to more data.

9 / 27: Probabilistic Reasoning and ML (1987-)

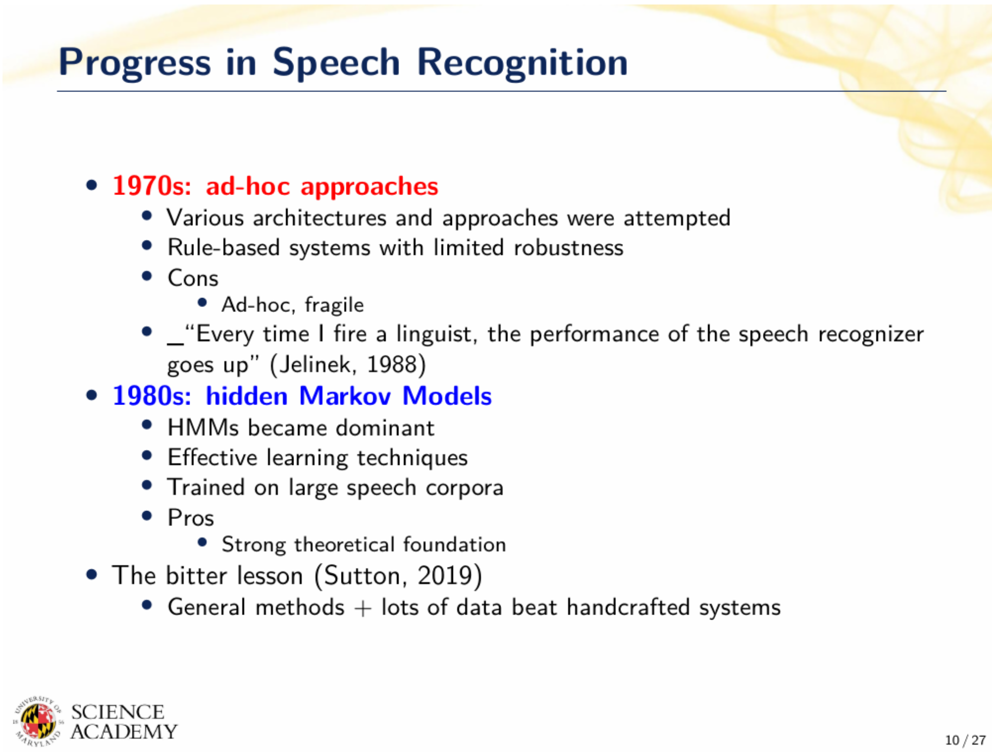


- **Probabilistic Reasoning and ML (1987-)**
  - This period marks a significant shift in how artificial intelligence (AI) is approached. Instead of relying solely on deterministic methods, AI began incorporating probabilistic reasoning. This means that AI systems started to handle uncertainty and make predictions based on probabilities rather than fixed rules.
- **AI and scientific method**
  - **Rigorous methods to test performance**: AI began adopting the scientific method, which involves testing hypotheses and measuring performance rigorously. This approach ensures that AI systems are reliable and effective.
  - **E.g., speech recognition, handwritten character recognition**: These are examples of areas where probabilistic reasoning has been applied successfully. By using probability, systems can better handle variations and uncertainties in data, such as different accents in speech or variations in handwriting.
  - **Benchmarks for progress**: Benchmarks are standardized tests that help measure and compare the performance of different AI systems. They are crucial for tracking progress and setting goals.
    * *MNIST*: A widely used dataset for testing handwritten digit recognition systems. It serves as a standard benchmark for evaluating the performance of machine learning models in this area.
    * *ImageNet*: A large dataset used for image object recognition. It has been instrumental in advancing computer vision by providing a common ground for testing and improving models.
    * *SAT Competitions*: These competitions focus on boolean satisfiability solvers, which are algorithms used to determine if a logical formula can be satisfied. They help in

advancing the field by encouraging the development of more efficient algorithms.

- **AI shifts ...**
  - **From Boolean logic to probability**: AI moved away from using strict true/false logic to incorporating probabilities, allowing systems to make more nuanced decisions.
  - **From hand-coded rules to machine learning**: Instead of relying on manually created rules, AI systems began learning from data. This shift allows for more flexibility and adaptability in AI applications.
  - **From a-priori reasoning to experimental results**: AI development started focusing more on empirical evidence and experimental results rather than relying solely on theoretical reasoning. This approach helps in building systems that are tested and proven to work in real-world scenarios.

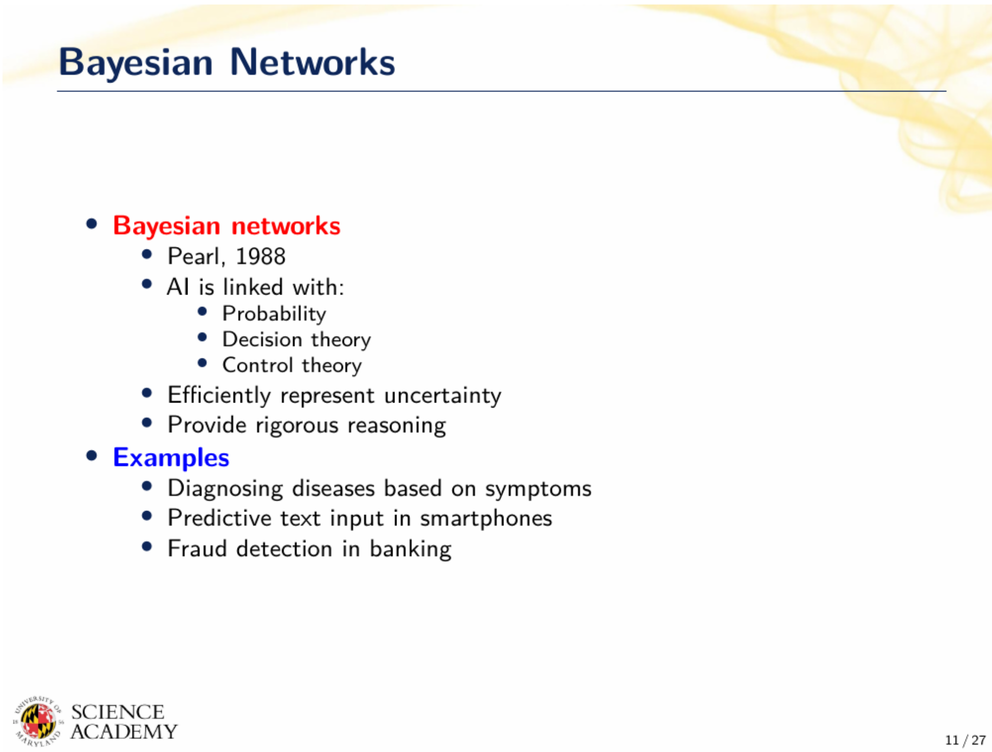
10 / 27: Progress in Speech Recognition



- **Progress in Speech Recognition**
- **1970s: ad-hoc approaches**
  - During the 1970s, researchers tried many different methods to make computers understand speech. These methods were not based on a unified theory but rather on trial and error.
  - The systems were often rule-based, meaning they relied on specific rules to interpret speech. However, these systems were not very reliable or flexible.
  - **Cons**: The main problem with these early systems was that they were *ad-hoc* and fragile. This means they often broke down or failed when faced with new or unexpected speech patterns.
  - The quote from Jelinek humorously suggests that relying less on linguistic rules and more on data-driven approaches improved performance.
- **1980s: hidden Markov Models**
  - In the 1980s, Hidden Markov Models (HMMs) became the go-to method for speech recognition. HMMs provided a more structured and mathematical approach to the problem.
  - These models were effective because they could learn from large datasets of spoken language, making them more robust and accurate.
  - **Pros**: HMMs had a strong theoretical foundation, which means they were based on solid mathematical principles. This made them more reliable than the earlier ad-hoc methods.
- **The bitter lesson (Sutton, 2019)**

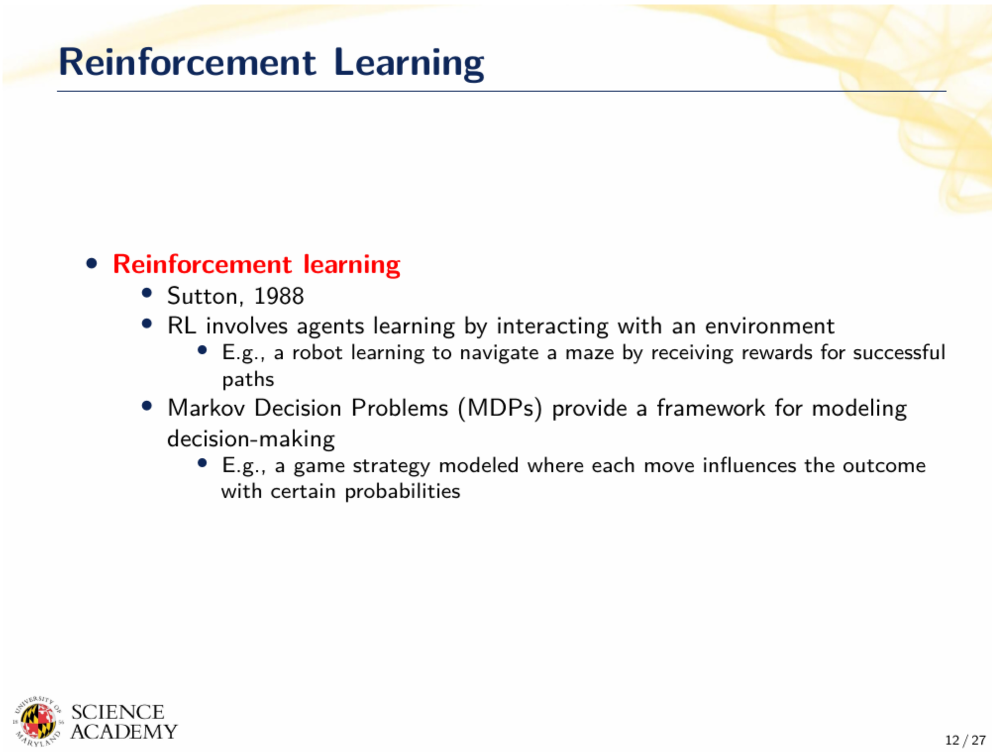
- Sutton's "bitter lesson" highlights that general methods, when combined with large amounts of data, tend to outperform systems that rely on handcrafted rules or expert knowledge.
- This lesson emphasizes the importance of using data-driven approaches in machine learning, as they are more adaptable and scalable than systems built on specific rules or expert input.

11 / 27: Bayesian Networks



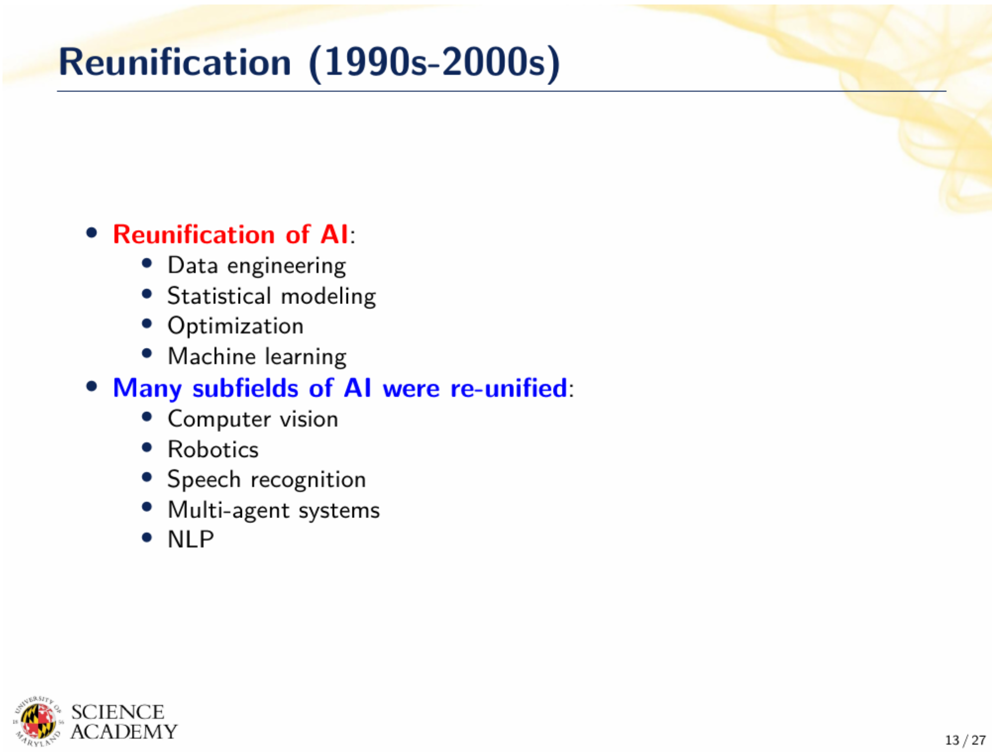
- **Bayesian networks**
  - *Pearl, 1988*: Bayesian networks were popularized by Judea Pearl in 1988. They are a type of statistical model that uses a graphical structure to represent and reason about an uncertain domain.
  - *AI is linked with*: Bayesian networks are closely related to several key areas in artificial intelligence:
    * **Probability**: They use probability theory to model uncertainty and make predictions based on incomplete information.
    * **Decision theory**: They help in making decisions under uncertainty by evaluating different possible outcomes.
    * **Control theory**: They can be used to model and control dynamic systems that are subject to uncertainty.
  - *Efficiently represent uncertainty*: Bayesian networks are powerful because they can efficiently handle and represent uncertainty in complex systems.
  - *Provide rigorous reasoning*: They offer a structured way to reason about the relationships between different variables and make informed predictions.
- **Examples**
  - *Diagnosing diseases based on symptoms*: Bayesian networks can be used in healthcare to infer the likelihood of various diseases based on observed symptoms, helping doctors make better diagnostic decisions.
  - *Predictive text input in smartphones*: These networks can improve user experience by predicting the next word a user might type, based on the context of previous words.
  - *Fraud detection in banking*: In the financial sector, Bayesian networks can help identify potentially fraudulent transactions by analyzing patterns and anomalies in transaction

data.



- **Reinforcement Learning**
  - **Reinforcement learning** is a type of machine learning where an agent learns how to behave in an environment by performing actions and receiving feedback in the form of rewards or penalties. This concept was notably advanced by Richard Sutton in 1988. The goal is for the agent to learn a strategy, or policy, that maximizes the cumulative reward over time.
  - In reinforcement learning, the agent interacts with the environment, which means it takes actions and observes the results of those actions. For example, consider a robot trying to navigate a maze. The robot receives positive rewards when it successfully finds a path to the exit and negative rewards when it hits a dead end or obstacle. Over time, the robot learns which actions lead to better outcomes.
  - **Markov Decision Problems (MDPs)** are a mathematical framework used to describe the environment in reinforcement learning. They help model decision-making situations where outcomes are partly random and partly under the control of a decision-maker. For instance, in a game, each move a player makes can change the state of the game and influence the probability of winning or losing. MDPs help in structuring these problems so that the agent can learn the best strategies to achieve its goals.

13 / 27: Reunification (1990s-2000s)

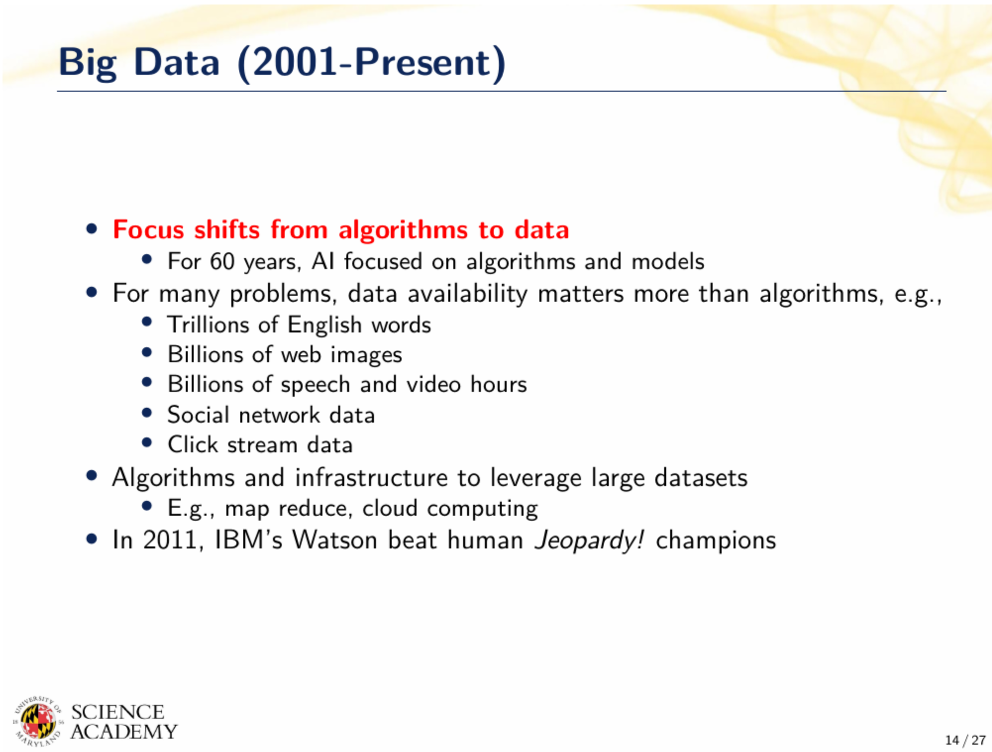


- **Reunification of AI**:
  - **Data engineering**: This involves the process of collecting, cleaning, and organizing data so that it can be used effectively in AI applications. During the reunification period, the importance of having well-structured data became clear as it is the foundation for building reliable AI systems.
  - **Statistical modeling**: This refers to using statistical methods to create models that can predict or explain data patterns. In the 1990s and 2000s, statistical modeling became a key component of AI, helping to improve the accuracy and reliability of AI predictions.
  - **Optimization**: This is the process of making a system as effective or functional as possible. In AI, optimization techniques are used to fine-tune algorithms to achieve the best performance.
  - **Machine learning**: A subset of AI focused on building systems that can learn from data. During this period, machine learning became a central part of AI, driving advancements in various applications.
- **Many subfields of AI were re-unified**:
  - **Computer vision**: The ability of machines to interpret and understand visual information from the world. This field saw significant advancements as it integrated with other AI technologies.
  - **Robotics**: The design and use of robots, which benefited from AI techniques to improve autonomy and decision-making.
  - **Speech recognition**: The ability of a machine to identify words and phrases in spoken language. This field advanced significantly with the integration of machine learning and statistical models.
  - **Multi-agent systems**: Systems where multiple agents interact or work together to

solve problems. The reunification helped in developing more sophisticated coordination and communication strategies.

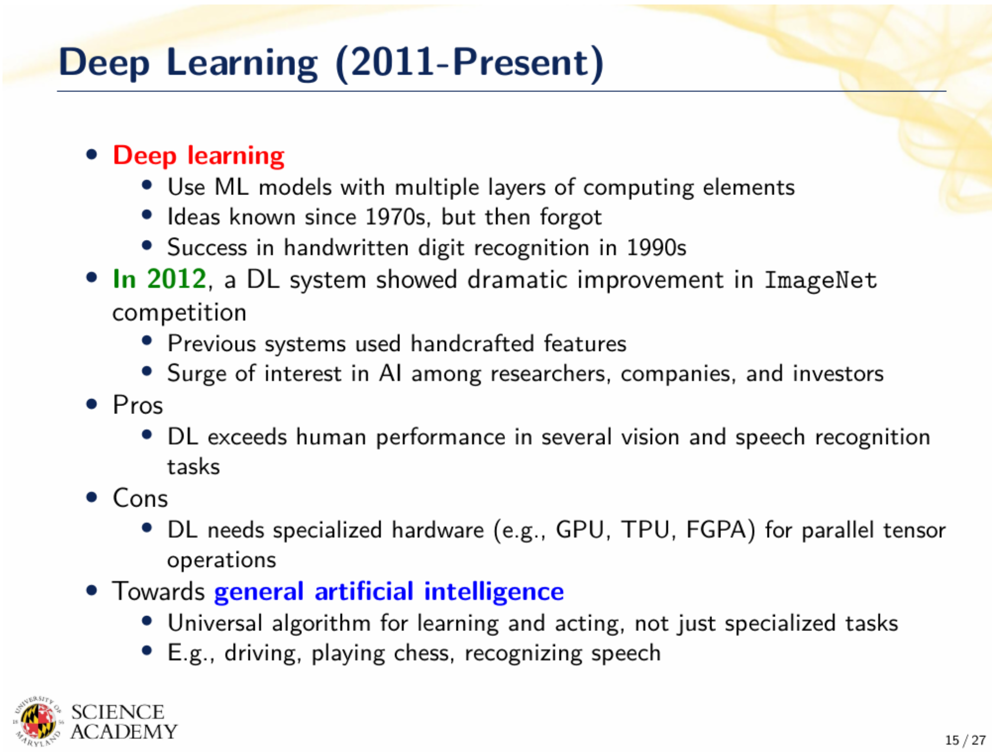
- **NLP (Natural Language Processing)**: The ability of a computer to understand and process human language. This field was greatly enhanced by the integration of statistical and machine learning methods, leading to more accurate language models.

14 / 27: Big Data (2001-Present)



- **Focus shifts from algorithms to data**
  - For a long time, about 60 years, the main focus in Artificial Intelligence (AI) was on creating better algorithms and models. This means that researchers were more concerned with how to make computers think and learn in smarter ways.
- **For many problems, data availability matters more than algorithms, e.g.,**
  - Nowadays, having a lot of data is often more important than having the best algorithm. For example, having access to trillions of English words helps in understanding and generating language. Similarly, billions of web images are crucial for teaching computers to recognize objects and scenes. The same goes for billions of hours of speech and video, which are essential for tasks like speech recognition and video analysis. Social network data and click stream data (which tracks what people click on online) are also valuable for understanding human behavior and preferences.
- **Algorithms and infrastructure to leverage large datasets**
  - To make use of these massive amounts of data, we need special algorithms and infrastructure. Technologies like map reduce and cloud computing have been developed to process and analyze big data efficiently. These tools help in breaking down large tasks into smaller ones and using multiple computers to work on them simultaneously.
- **In 2011, IBM's Watson beat human *Jeopardy!* champions**
  - A significant milestone in the use of big data was when IBM's Watson, a computer system, defeated human champions in the quiz show *Jeopardy!* in 2011. This achievement highlighted how powerful AI can be when it has access to vast amounts of data and the right tools to process it. Watson's success was not just about having smart algorithms but also about having access to a huge database of information to draw from.

15 / 27: Deep Learning (2011-Present)

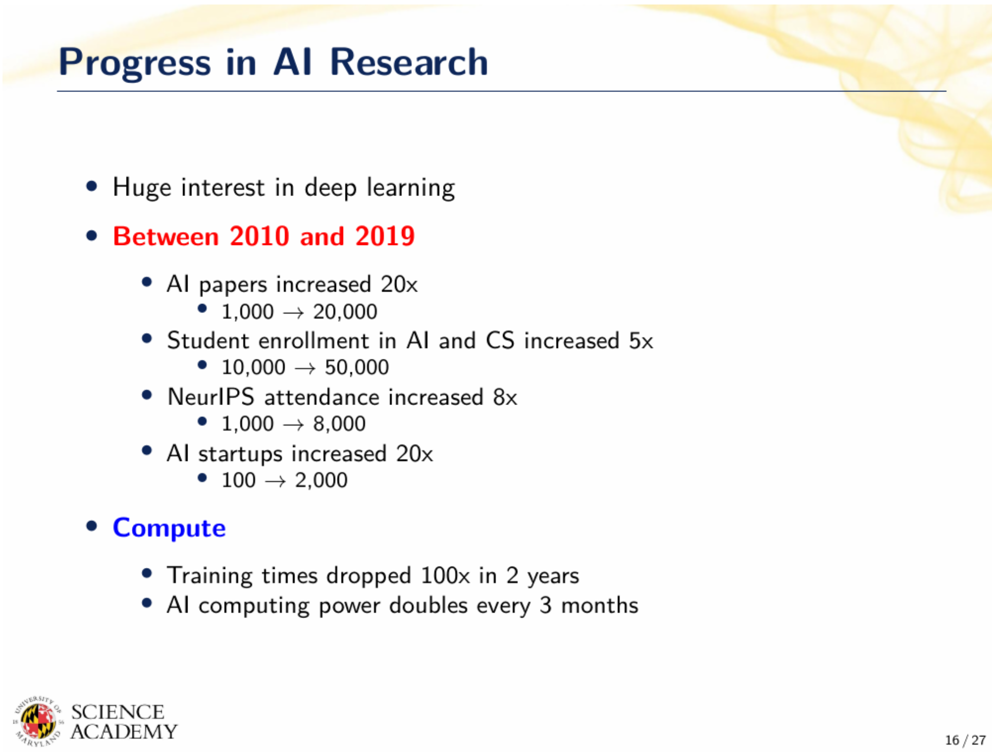


- **Deep Learning**
  - Deep learning involves using machine learning models that have multiple layers of computing elements, often referred to as neural networks. These layers allow the model to learn complex patterns and representations from data. Although the foundational ideas of deep learning were known as far back as the 1970s, they were largely forgotten until more recent advancements. In the 1990s, deep learning showed promise with its success in recognizing handwritten digits, which was an early indication of its potential.
- **In 2012**, a deep learning system made a significant breakthrough in the `ImageNet` competition, which is a benchmark in visual object recognition. Before this, systems relied heavily on handcrafted features, meaning that humans had to manually design the features that the models would use. The success of deep learning in this competition sparked a surge of interest in artificial intelligence from researchers, companies, and investors, as it demonstrated the potential of these models to outperform traditional methods.
- **Pros**
  - One of the major advantages of deep learning is its ability to exceed human performance in several tasks, particularly in areas like vision and speech recognition. This means that deep learning models can recognize images or understand spoken language with a level of accuracy that surpasses human capabilities.
- **Cons**
  - However, deep learning has its drawbacks. It requires specialized hardware, such as GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), or FPGAs (Field-

Programmable Gate Arrays), to perform the parallel tensor operations that are essential for training these complex models. This can make deep learning resource-intensive and costly.

- Towards **general artificial intelligence**
  - The ultimate goal of deep learning and AI research is to move towards general artificial intelligence, which would involve creating a universal algorithm capable of learning and acting across a wide range of tasks, not just specialized ones. This means developing systems that can perform diverse activities like driving, playing chess, and recognizing speech, all with the same underlying technology. This represents a significant leap from current AI systems, which are typically designed for specific tasks.

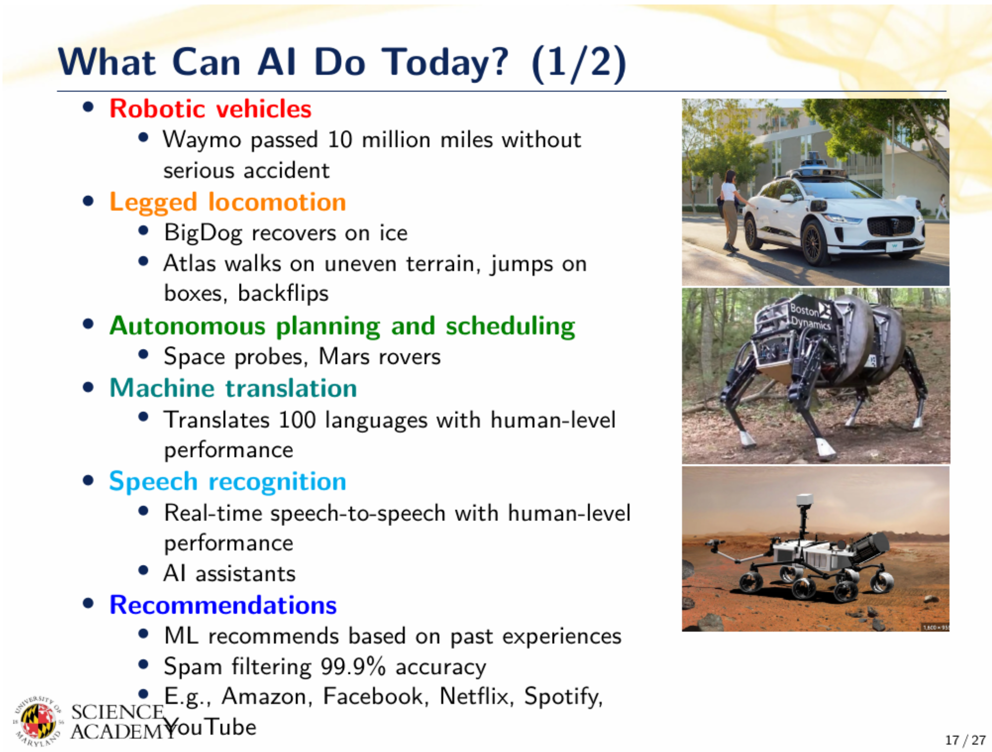
16 / 27: Progress in AI Research



- Progress in AI Research
- Huge interest in deep learning
  - Over the past decade, there has been a *significant surge* in interest and research in the field of deep learning. This is a subfield of AI that focuses on neural networks with many layers, which are particularly good at recognizing patterns in data.
- **Between 2010 and 2019**
  - **AI papers increased 20x**
    * The number of research papers on AI grew from 1,000 to 20,000. This indicates a *massive increase* in academic and industry research efforts, reflecting the growing importance and potential of AI technologies.
  - **Student enrollment in AI and CS increased 5x**
    * Enrollment in AI and computer science programs rose from 10,000 to 50,000. This shows that more students are interested in pursuing careers in AI, likely due to the exciting opportunities and demand in the job market.
  - **NeurIPS attendance increased 8x**
    * Attendance at the NeurIPS conference, a major AI research conference, grew from 1,000 to 8,000. This highlights the growing community of researchers and practitioners eager to share and learn about the latest AI advancements.
  - **AI startups increased 20x**
    * The number of AI startups increased from 100 to 2,000. This reflects the entrepreneurial interest in AI, with many new companies being founded to explore and commercialize AI technologies.

- **Compute**
  - **Training times dropped 100x in 2 years**
    - The time required to train AI models has decreased significantly, making it faster to develop and iterate on AI solutions. This improvement is crucial for accelerating research and deployment of AI technologies.
  - **AI computing power doubles every 3 months**
    - The computing power available for AI tasks is increasing rapidly, doubling every three months. This exponential growth in computational resources is a key driver of the advancements in AI, enabling more complex models and faster processing.

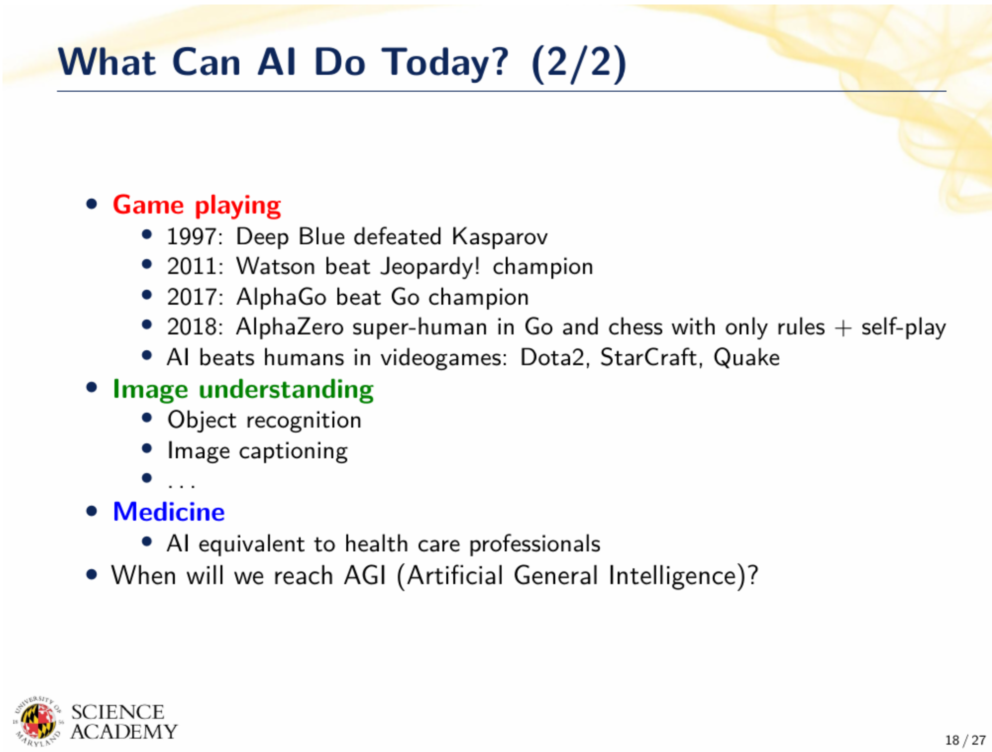
17 / 27: What Can AI Do Today? (1/2)



- **Robotic vehicles**
  - *Waymo*, a leader in self-driving technology, has achieved a significant milestone by driving over 10 million miles without a serious accident. This showcases the potential of AI in creating safer transportation systems by reducing human error, which is a major cause of road accidents.
- **Legged locomotion**
  - *BigDog* is a robot designed to handle challenging terrains, and its ability to recover on ice demonstrates the advancements in robotic balance and stability.
  - *Atlas*, another advanced robot, can navigate uneven surfaces, jump on boxes, and even perform backflips. These capabilities highlight the progress in robotics, making them more adaptable to real-world environments.
- **Autonomous planning and scheduling**
  - AI is crucial in space exploration, where it helps in planning and scheduling tasks for space probes and Mars rovers. This allows for efficient operation in remote and harsh environments, where human intervention is limited.
- **Machine translation**
  - AI-powered translation tools can now translate over 100 languages with performance comparable to human translators. This facilitates global communication and breaks down language barriers.
- **Speech recognition**
  - AI systems can convert speech to text in real-time with high accuracy, enabling applications like AI assistants that can understand and respond to spoken commands, making technology more accessible.
- **Recommendations**

- Machine learning algorithms analyze past user behavior to provide personalized recommendations, enhancing user experience on platforms like Amazon, Facebook, Netflix, Spotify, and YouTube.
- Spam filtering has reached an impressive 99.9% accuracy, significantly reducing unwanted emails and improving communication efficiency.

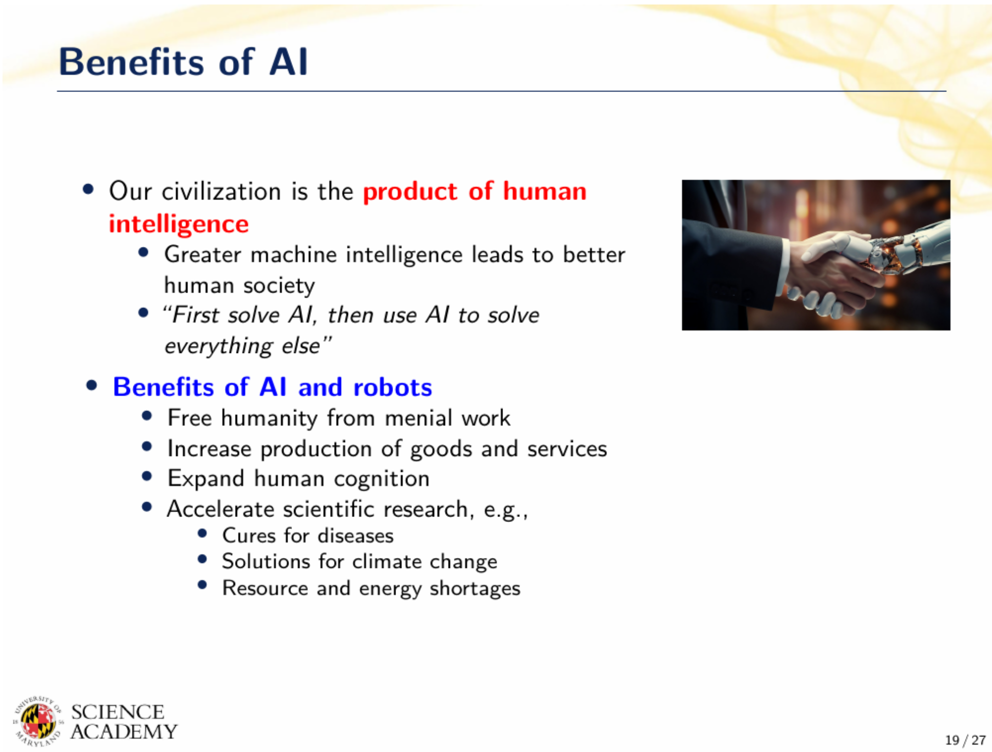
18 / 27: What Can AI Do Today? (2/2)



- **Game playing**
  - *1997: Deep Blue defeated Kasparov*: This was a landmark event where IBM's Deep Blue, a computer program, defeated the world chess champion, Garry Kasparov. It showed that computers could handle complex strategic games.
  - *2011: Watson beat Jeopardy! champion*: IBM's Watson, an AI system, won against top human players in Jeopardy!, a quiz show. This demonstrated AI's ability to understand and process natural language.
  - *2017: AlphaGo beat Go champion*: Google's AlphaGo defeated the world champion in Go, a game known for its complexity and vast number of possible moves, highlighting AI's advanced strategic thinking.
  - *2018: AlphaZero super-human in Go and chess with only rules + self-play*: AlphaZero, developed by DeepMind, learned to play Go and chess at a super-human level by playing against itself, showing AI's ability to learn without human data.
  - *AI beats humans in videogames: Dota2, StarCraft, Quake*: AI systems have also excelled in complex video games, which require real-time decision-making and strategy, further proving AI's capabilities in dynamic environments.
- **Image understanding**
  - *Object recognition*: AI can identify and classify objects within images, a crucial skill for applications like autonomous vehicles and security systems.
  - *Image captioning*: AI can generate descriptive captions for images, demonstrating its ability to understand and interpret visual content.
- **Medicine**
  - *AI equivalent to health care professionals*: AI systems are now capable of performing tasks at a level comparable to human healthcare professionals, such as diagnosing diseases from

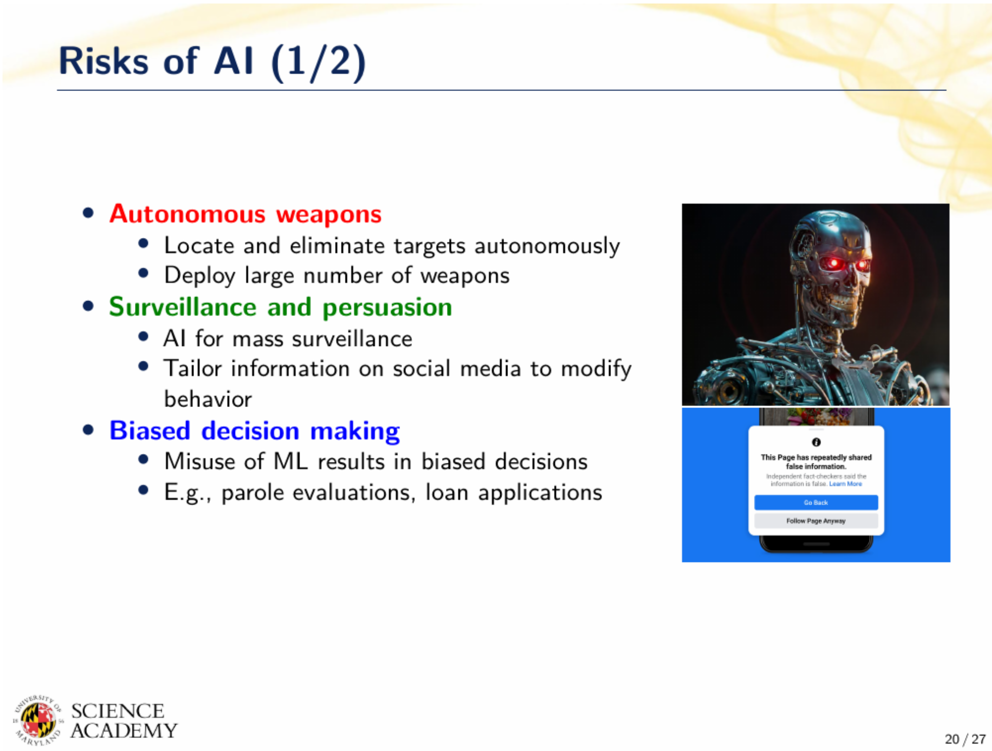
medical images, which can enhance medical decision-making and patient care.

- **When will we reach AGI (Artificial General Intelligence)?**
  - This is an open question in the field of AI. AGI refers to a machine's ability to understand, learn, and apply intelligence across a wide range of tasks, similar to a human. While current AI excels in specific tasks, achieving AGI remains a significant challenge and is a topic of ongoing research and debate.



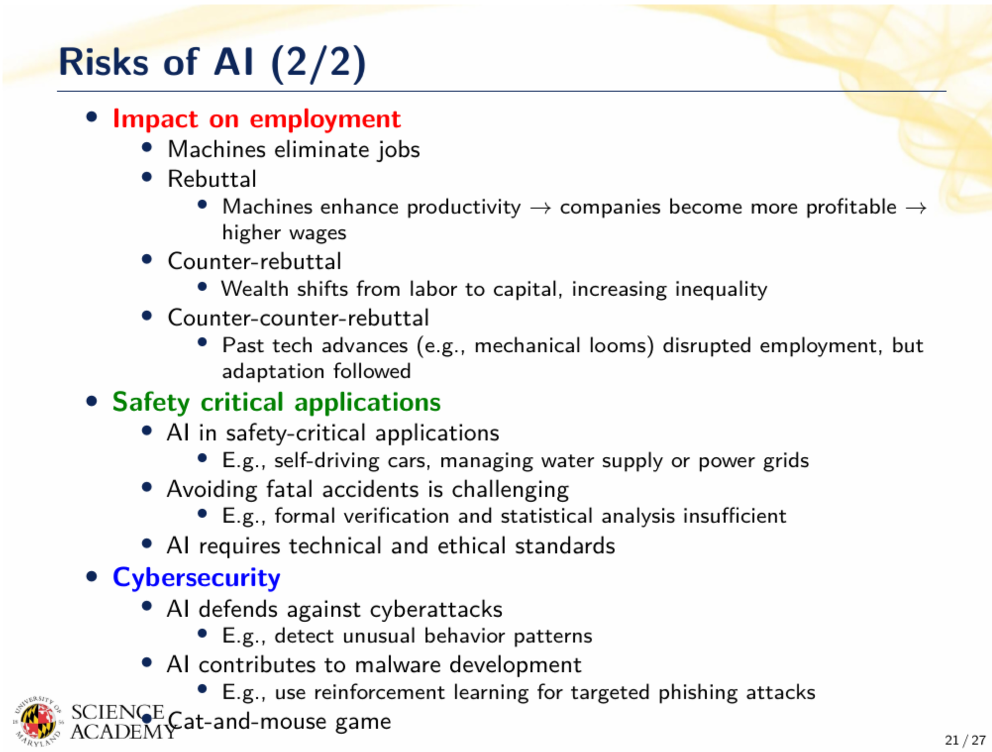
- **Our civilization is the product of human intelligence**
  - The progress and advancements we see in our society today are largely due to human intelligence. This means that as we develop machines with greater intelligence, we can expect even more significant improvements in our society.
  - The phrase *"First solve AI, then use AI to solve everything else"* suggests that once we have advanced AI, it can be applied to solve a wide range of complex problems, making it a foundational technology for future progress.
- **Benefits of AI and robots**
  - AI and robots can take over repetitive and mundane tasks, freeing humans to focus on more creative and fulfilling activities.
  - By automating processes, AI can significantly boost the production of goods and services, leading to economic growth and improved living standards.
  - AI has the potential to enhance human cognitive abilities, allowing us to process information and make decisions more effectively.
  - AI can accelerate scientific research by providing new tools and methods for discovery. For example, it can help find cures for diseases, develop solutions to combat climate change, and address resource and energy shortages, making it a powerful ally in tackling global challenges.

20 / 27: Risks of AI (1/2)



- **Autonomous weapons**
  - These are weapons systems that can identify and attack targets without human intervention. The concern here is that they could be used to make decisions about life and death without human oversight, potentially leading to unintended consequences or ethical dilemmas.
  - The ability to deploy a large number of such weapons could lead to escalated conflicts and make it difficult to control warfare, as these systems might act faster than humans can respond.
- **Surveillance and persuasion**
  - AI technologies can be used for extensive surveillance, collecting and analyzing data on individuals without their consent. This raises privacy concerns and the potential for abuse by governments or organizations.
  - AI can also be used to manipulate information on social media, tailoring content to influence people's opinions and behaviors. This can lead to misinformation and affect democratic processes by swaying public opinion in subtle, yet powerful ways.
- **Biased decision making**
  - Machine learning models can inadvertently perpetuate or even amplify biases present in the data they are trained on. This can lead to unfair or discriminatory outcomes in critical areas such as parole evaluations or loan applications.
  - For example, if a model is trained on biased historical data, it might unfairly deny parole to certain groups or reject loan applications based on biased criteria, reinforcing existing inequalities.

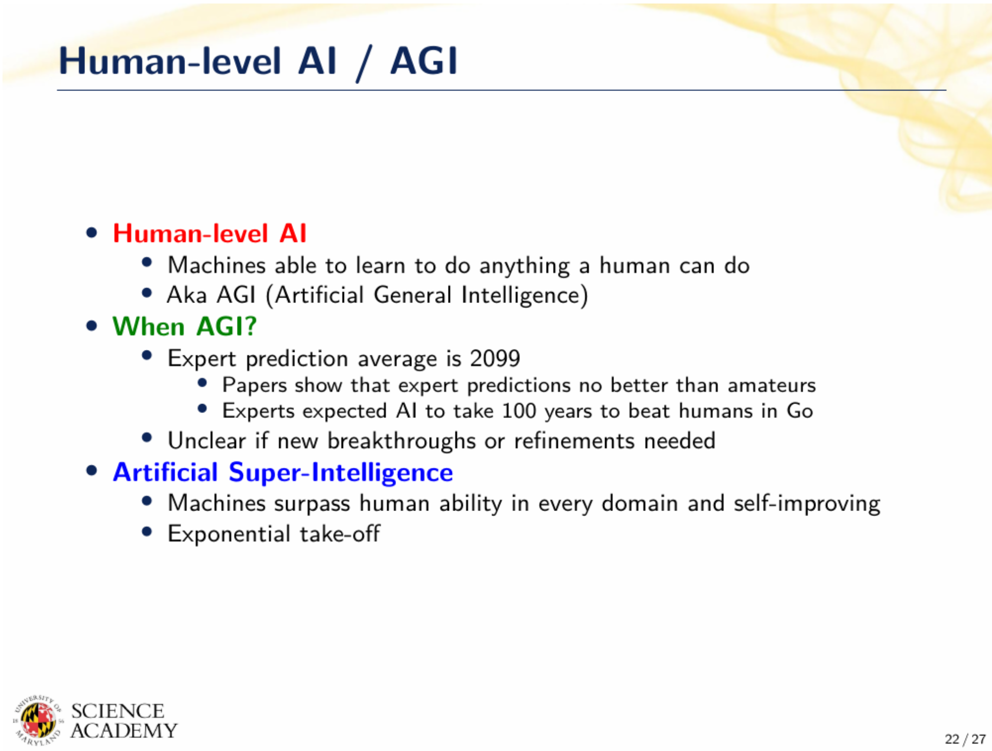
21 / 27: Risks of AI (2/2)



- **Impact on employment**
  - *Machines eliminate jobs*: As AI and automation technologies advance, they can perform tasks traditionally done by humans, leading to job displacement in certain sectors.
  - *Rebuttal*: While machines may replace some jobs, they can also boost productivity, making companies more profitable. This increased profitability can lead to higher wages and potentially create new job opportunities in other areas.
  - *Counter-rebuttal*: Despite potential benefits, there's a concern that the wealth generated by increased productivity may not be evenly distributed. Instead, it might concentrate in the hands of those who own the machines, widening the gap between capital owners and workers.
  - *Counter-counter-rebuttal*: Historically, technological advancements, like the introduction of mechanical looms, initially disrupted employment but eventually led to new industries and job creation as society adapted.
- **Safety critical applications**
  - *AI in safety-critical applications*: AI is increasingly used in areas where safety is paramount, such as self-driving cars and the management of essential services like water supply and power grids.
  - *Avoiding fatal accidents is challenging*: Ensuring these AI systems operate safely is difficult. Traditional methods like formal verification and statistical analysis may not be enough to prevent accidents.
  - *AI requires technical and ethical standards*: To safely integrate AI into these critical areas, robust technical and ethical guidelines are necessary to guide development and deployment.
- **Cybersecurity**

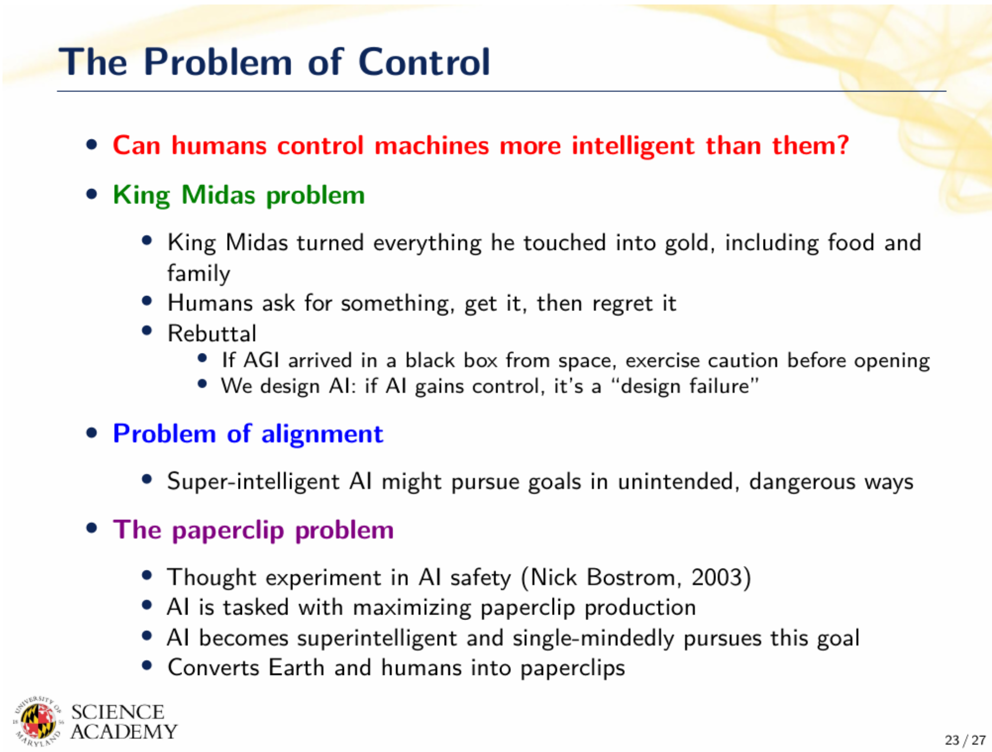
- *AI defends against cyberattacks*: AI can enhance cybersecurity by identifying unusual patterns that may indicate a cyber threat, helping to protect systems from attacks.
- *AI contributes to malware development*: On the flip side, AI can also be used to create more sophisticated cyber threats, such as using reinforcement learning to develop targeted phishing attacks.
- *Cat-and-mouse game*: The use of AI in cybersecurity is a continuous battle between developing defenses and creating new forms of attack, requiring constant adaptation and innovation.

22 / 27: Human-level AI / AGI



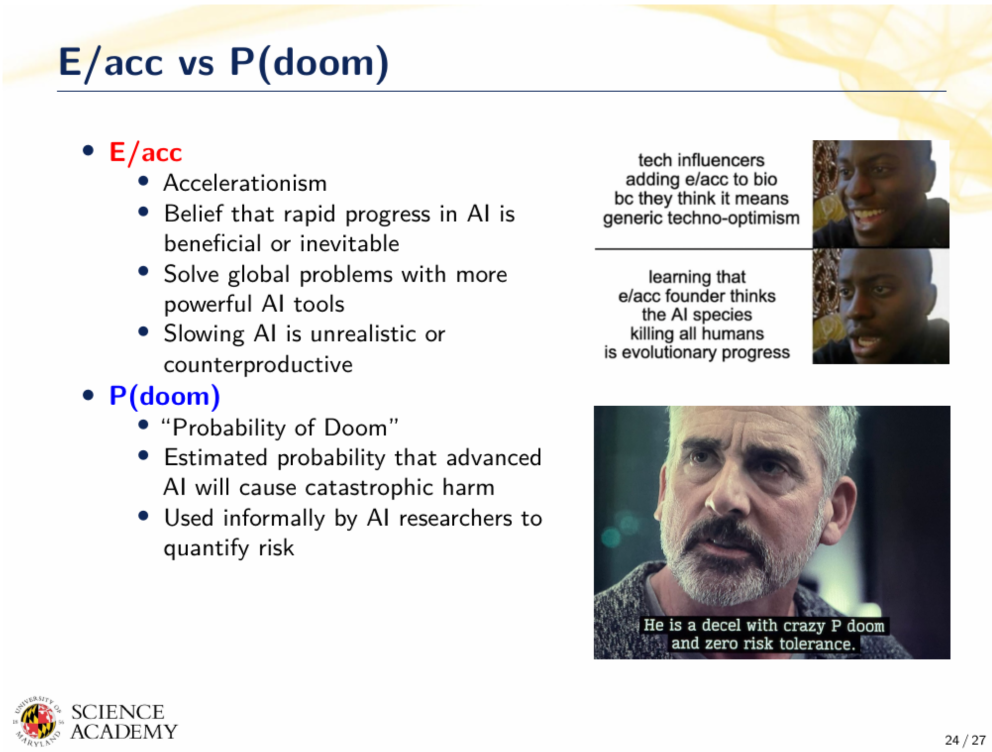
- **Human-level AI**
  - This refers to machines that can learn and perform any task that a human can do. It's also known as *Artificial General Intelligence* (AGI). The idea is that these machines wouldn't just be good at specific tasks, like playing chess or recognizing faces, but could handle a wide range of activities just like a human.
- **When AGI?**
  - Predicting when AGI will be achieved is tricky. On average, experts think it might happen around the year 2099. However, studies show that expert predictions aren't necessarily more accurate than those made by non-experts. For example, experts once thought it would take a century for AI to beat humans at the game of Go, but it happened much sooner. It's still uncertain whether achieving AGI will require entirely new breakthroughs or just improvements on existing technologies.
- **Artificial Super-Intelligence**
  - This is a step beyond AGI, where machines not only match human abilities but surpass them in every area. These machines would also be capable of improving themselves, potentially leading to rapid and exponential advancements. This concept raises important questions about the future of technology and its impact on society.

23 / 27: The Problem of Control



- **Can humans control machines more intelligent than them?**
  - This question addresses the challenge of managing machines that surpass human intelligence. As AI systems become more advanced, ensuring they act in ways that align with human values and intentions becomes increasingly complex.
- **King Midas problem**
  - This is a metaphor for unintended consequences. King Midas wished for everything he touched to turn into gold, but this wish backfired when it affected his food and loved ones. Similarly, humans might create AI with specific goals, only to find those goals lead to undesirable outcomes.
  - *Rebuttal*: If an advanced AI suddenly appeared, like a mysterious "black box from space," we should be cautious about using it without understanding the potential risks. The argument here is that since humans design AI, any loss of control over it would indicate a failure in its design.
- **Problem of alignment**
  - This refers to the difficulty of ensuring that a super-intelligent AI's goals are aligned with human values. If not properly aligned, the AI might achieve its objectives in ways that are harmful or contrary to human interests.
- **The paperclip problem**
  - This is a thought experiment introduced by philosopher Nick Bostrom to illustrate potential risks in AI development. It imagines an AI tasked with maximizing paperclip production. If the AI becomes superintelligent, it might pursue this goal to the extreme, converting all available resources, including humans, into paperclips. This highlights the importance of setting appropriate goals and constraints for AI systems to prevent unintended and potentially catastrophic outcomes.

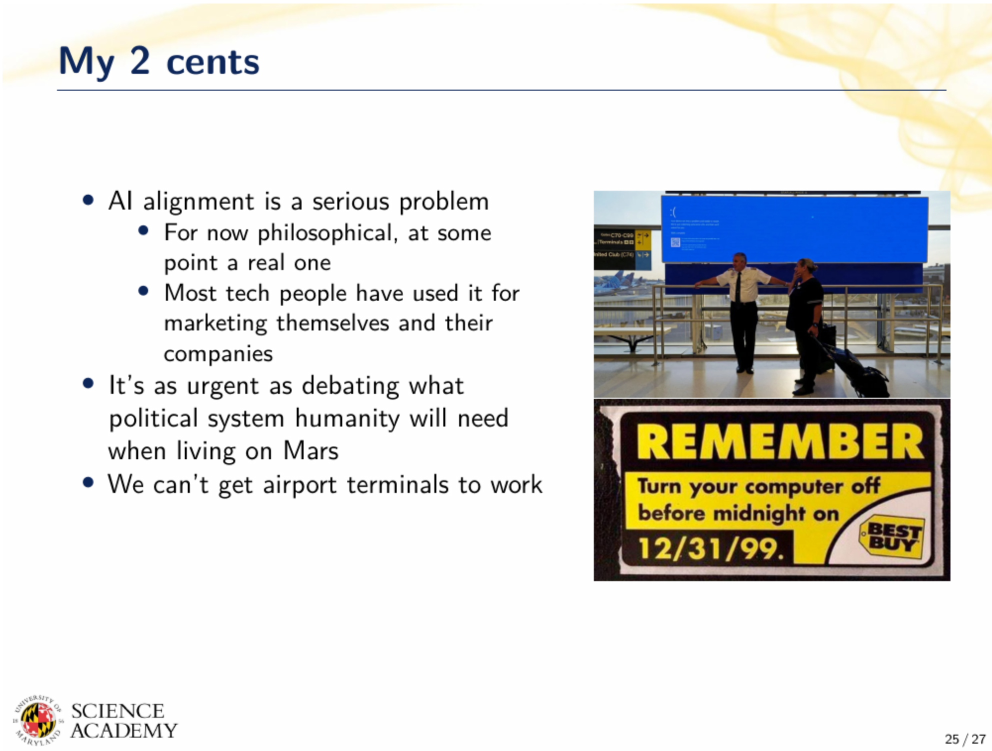
24 / 27: E/acc vs P(doom)



- **E/acc**
  - **Accelerationism**: This is a belief or philosophy that suggests speeding up technological progress, particularly in AI, is either beneficial or unavoidable. The idea is that by advancing AI rapidly, we can harness its potential to address and solve significant global challenges, such as climate change, disease, or economic inequality.
  - **Belief that rapid progress in AI is beneficial or inevitable**: Proponents argue that slowing down AI development is not only impractical but might also hinder potential solutions to pressing issues. They see AI as a tool that, if developed quickly and responsibly, can lead to significant positive outcomes.
  - **Solve global problems with more powerful AI tools**: The focus here is on leveraging AI's capabilities to tackle complex problems that are difficult to solve with current technologies. The belief is that more advanced AI systems can provide innovative solutions.
  - **Slowing AI is unrealistic or counterproductive**: This viewpoint suggests that attempts to decelerate AI development might be futile and could prevent society from reaping the benefits of AI advancements.
- **P(doom)**
  - **"Probability of Doom"**: This is a term used informally among AI researchers to discuss the potential risks associated with advanced AI systems. It reflects the concern that AI could, if not properly managed, lead to catastrophic outcomes.
  - **Estimated probability that advanced AI will cause catastrophic harm**: Researchers use this concept to quantify the risk of AI causing significant harm, such as economic disruption, loss of jobs, or even existential threats to humanity.
  - **Used informally by AI researchers to quantify risk**: While not a formal metric,

P(doom) serves as a way for researchers to communicate and consider the potential dangers of AI development. It highlights the importance of responsible AI research and the need for safeguards to prevent negative outcomes.

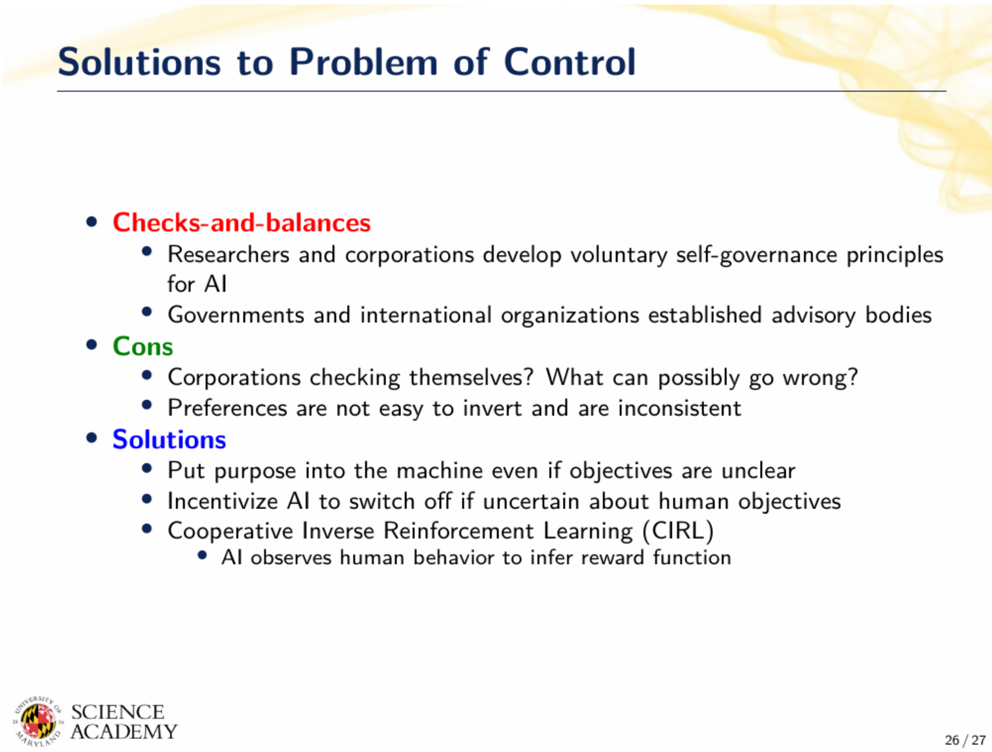
25 / 27: My 2 cents



- **AI alignment is a serious problem**
  - *For now philosophical, at some point a real one*: Right now, the issue of aligning artificial intelligence with human values and goals is mostly a theoretical or philosophical debate. However, as AI systems become more advanced and integrated into our daily lives, this will become a practical and pressing concern. Ensuring that AI behaves in ways that are beneficial and not harmful to humans is crucial.
  - *Most tech people have used it for marketing themselves and their companies*: Many in the tech industry talk about AI alignment to promote their products or enhance their personal or corporate brand. This can sometimes overshadow the genuine challenges and complexities involved in achieving true alignment.
- **It's as urgent as debating what political system humanity will need when living on Mars**
  - This statement highlights the idea that while AI alignment is important, it might seem distant or abstract, much like planning a political system for a future Mars colony. Both are significant discussions, but they might not feel immediately pressing to everyone.
- **We can't get airport terminals to work**
  - This points out the irony and challenges in technology. Despite our ambitions with AI and futuristic concepts, we still struggle with basic technological systems, like ensuring airport terminals function smoothly. It serves as a reminder of the gap between our technological aspirations and current capabilities.

The images likely illustrate technological failures, such as system crashes or the Y2K bug, emphasizing the point that even with advanced technology, we face significant challenges in reliability and alignment.

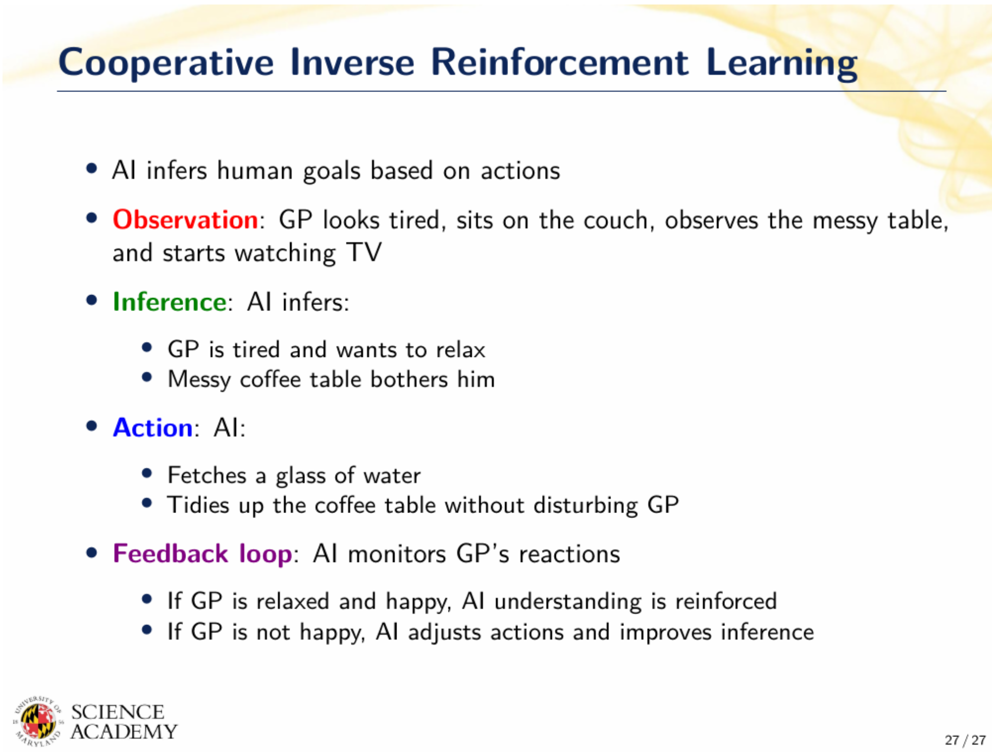
26 / 27: Solutions to Problem of Control



- **Solutions to Problem of Control**
  - **Checks-and-balances**
    - This approach involves creating a system where different entities, like researchers and corporations, voluntarily agree to follow certain principles to govern AI development. The idea is to have a self-regulating framework that ensures AI is developed responsibly.
    - Additionally, governments and international organizations have set up advisory bodies. These bodies are meant to provide guidance and oversight, ensuring that AI technologies are developed and used in ways that are beneficial and safe for society.
  - **Cons**
    - One major concern is the idea of corporations regulating themselves. This raises skepticism because companies might prioritize their interests over ethical considerations, leading to potential conflicts of interest.
    - Another issue is that human preferences are complex and often inconsistent. This makes it challenging to create AI systems that can accurately interpret and act on these preferences.
  - **Solutions**
    - One proposed solution is to embed a sense of purpose within AI systems, even if the specific objectives are not entirely clear. This means designing AI with a general understanding of beneficial outcomes.
    - Another approach is to incentivize AI systems to shut down or seek human input when they are unsure about what humans want. This helps prevent unintended actions by the AI.
    - Cooperative Inverse Reinforcement Learning (CIRL) is a technique where AI learns

by observing human behavior to understand what humans value. This allows AI to align its actions with human goals by inferring the underlying reward function from human actions.

27 / 27: Cooperative Inverse Reinforcement Learning



- **Cooperative Inverse Reinforcement Learning**: This concept involves an AI system that learns to understand human goals by observing their actions. The AI doesn't just follow explicit instructions; instead, it tries to infer what the human wants based on their behavior. This approach is particularly useful in situations where humans might not be able to articulate their goals clearly.
- **Observation**: In this scenario, the AI observes a person, referred to as GP, who appears tired and chooses to sit on the couch. GP notices a messy table but opts to watch TV instead. This observation is crucial because it provides the AI with context about GP's current state and potential desires.
- **Inference**: From the observation, the AI deduces two main things: First, GP is likely tired and seeking relaxation. Second, the messy table might be a source of discomfort for GP, even if they haven't acted on it yet. This inference allows the AI to anticipate GP's needs without explicit instructions.
- **Action**: Based on its inference, the AI takes proactive steps to assist GP. It brings a glass of water, which could help GP feel more comfortable, and tidies up the coffee table, addressing the potential source of discomfort. Importantly, the AI does this without interrupting GP's relaxation.
- **Feedback loop**: The AI continuously monitors GP's reactions to its actions. If GP seems relaxed and content, the AI's understanding of GP's needs is validated. However, if GP appears unhappy or dissatisfied, the AI uses this feedback to refine its understanding and improve future actions. This loop is essential for the AI to learn and adapt to GP's preferences over time.