THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

AIAA 2290: Ethics, Privacy and Security in AI

Foundations of AI Ethics

Xuming HU

xuminghu@hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

2025 Spring

About Me and Course

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

群聊: AIAA2290课程群

- Xuming HU
- Assistant Professor@AI Thrust

- Research Interests
  - Natural Language Processing
  - Large Language Models

- Contact Info
  - Email: xuminghu@hkust-gz.edu.cn
  - Office: Room E3-303
  - Office Hour: Every Monday & Wednesday 6:30 PM-7:30 PM

- TAs
  - Yuanhuiyi LYU (ylyu650@connect.hkust-gz.edu.cn)
  - Yibo YAN (yyan047@connect.hkust-gz.edu.cn)
  - Xiuze ZHOU (xzhou154@connect.hkust-gz.edu.cn)

About Me

WeChat Group

Anonymous Feedback

About Me and Course

Xuming HU's research interests include natural language processing and large language models, specifically in trustworthy large language model exploration, multi-modality large language models, and the training and development of vertical large language models. In the past five years, Dr. Hu has published over 20 first-author papers in top international journals and conferences in the fields of data mining and natural language processing, such as KDD, ICLR, ACL, and TKDE. He has served as an Area Chair for top natural language processing conferences like ACL, EMNLP, NAACL, and EACL, as well as the Action Editor for ACL Rolling Review

PhDs

刘书良
清华大学硕士

薛海威
清华大学硕士

刘翔
香港大学硕士

严一博
新加坡国立大学硕士

邓浩霖
哈尔滨工业大学（深圳）硕士

李俊卓
天津大学硕士

吕原汇一
东北大学本科

郑旭
东北大学硕士

徐亦捷
香港科技大学（广州）硕士

- **Ethics, Privacy and Security in AI**

- **Content:**
  - 13 Week Teaching
    - AI Ethics (Week 1-2)
    - AI Privacy (Week 3-7)
    - AI Security (Week 8-13)
  - The **first class** of the week is a regular class, and the **second class** consists of practice, classroom activities and **student presentations**.

Course Requirements and Grading Breakdown

Total: 100 Points

- **Attendance:** 20 Points
- **Student Presentations:** 25 Points
- **Final Report:** 55 Points

Course Requirements and Grading Breakdown

Attendance (20 Points)

- **Description**: ***Regular attendance is essential for successful participation and understanding of course material***.
- **Mechanism**:
  - The instructor will take attendance **12 times** throughout the semester.
  - Each attendance session is worth **2.5 points**.
  - Full Attendance Score (20 points): Achieved by attending **at least 8** out of 12 attendance sessions.

Student Presentations (25 Points)

- **Description**: ***Each student is required to deliver a presentation on a relevant topic within the course scope***.
- **Format**:
  - Duration:
    - **10 minutes** for the presentation.
    - **2-3 minutes** for a Q&A session with peers and the instructor.
  - Content: Presentations should demonstrate a clear understanding of ethical, privacy, or security issues in AI, supported by relevant case studies or research.
- **Evaluation Criteria**:
  - Content Mastery (**10 points**): Depth of understanding and accuracy of the information presented.
  - Presentation Skills (**10 points**): Clarity, organization, and delivery of the presentation.
  - Engagement and Q&A (**5 points**): Ability to engage the audience and effectively respond to questions.

Course Requirements and Grading Breakdown

Student Presentations (25 Points)

- **Description**: ***Each student is required to deliver a presentation on a relevant topic within the course scope***.

- **Notes**:
  - **Starting from the second class of Week 3 (26th February)**
  - **4 students will deliver their presentations during each Wednesday class session**
  - **The order of presentations will be randomly generated and announced at this Wednesday.**

Final Report (55 Points)

- **Description**: ***A comprehensive written report that explores a specific topic related to Ethics, Privacy, or Security in AI.***
- **Requirements**:
  - Length: limited to **8 pages**, double-spaced.
  - Format: Follow the given template.
  - Content:
    - Introduction and background of the chosen topic.
    - Detailed analysis and discussion of relevant issues.
    - Case studies or examples to illustrate key points.
    - Conclusion and recommendations.
- **Evaluation Criteria**: We will show the details on Canvas.
- **Late Submission Policy**: Late submissions will be penalized. We will deduct **3%** of the overall score for every **24 hours** after the deadline (2025/05/26).

Workload

For each student:

- **Attend 8 times (20 Points)**
- **10-min presentations (25 Points)**
- **final report (8 pages at most) (55 Points)**

Part I: Ethics in AI (Week 1-2)

Week 1: Foundations of AI Ethics

Week 2: Ethical Challenges and Responsibilities

Part II: Privacy in AI (Week 3-7)

Week 3: Foundations of Data Privacy

Week 4: Privacy Across AI Domains: NLP

Week 5: Privacy Across AI Domains: CV

Week 6: Sector-Specific Privacy Concerns and Innovations: AI-Driven Healthcare

Week 7: Sector-Specific Privacy Concerns and Innovations: Social Media

Part III: Security in AI (Week 8-13)

Week 8: Introduction to AI Security

Week 9: Defending AI Systems

Week 10: Security in Natural Language Processing

Week 11: Security in Computer Vision and Robotics

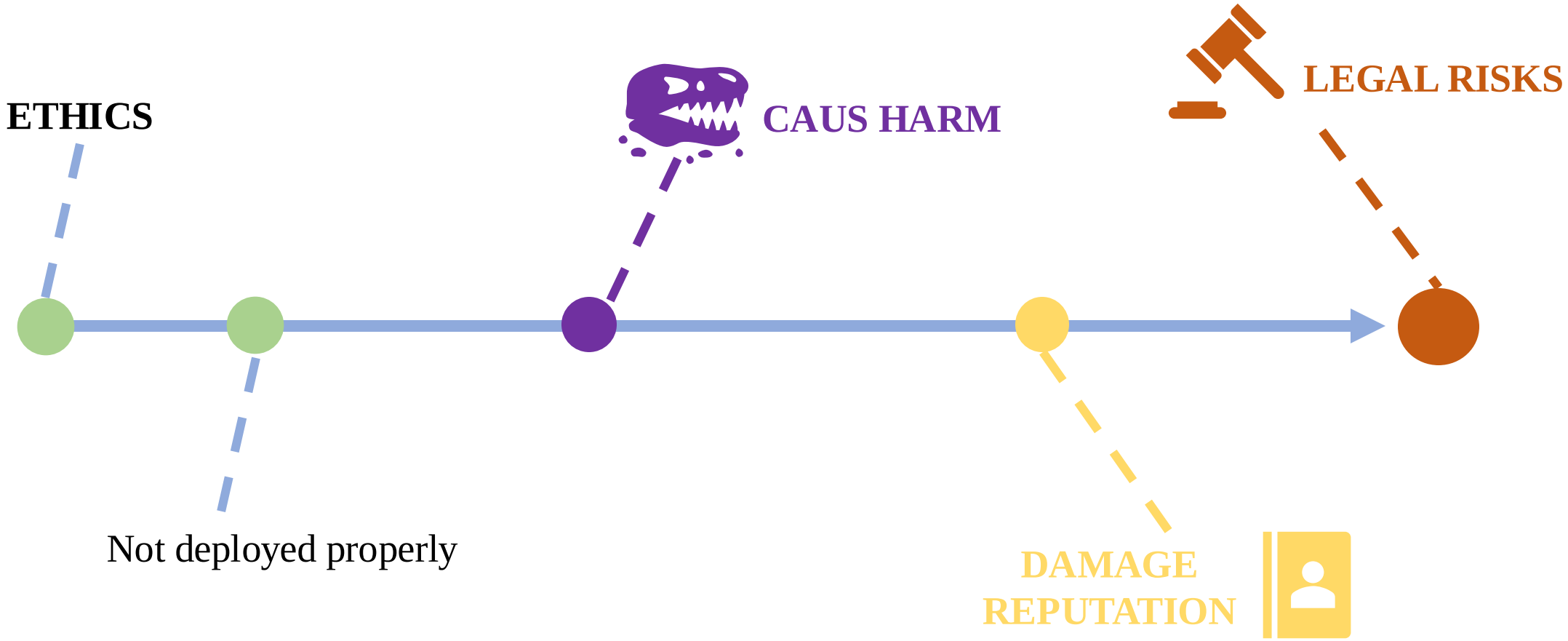
Week 12: Security in Autonomous Systems and Healthcare

Week 13: Future Directions and Comprehensive Review

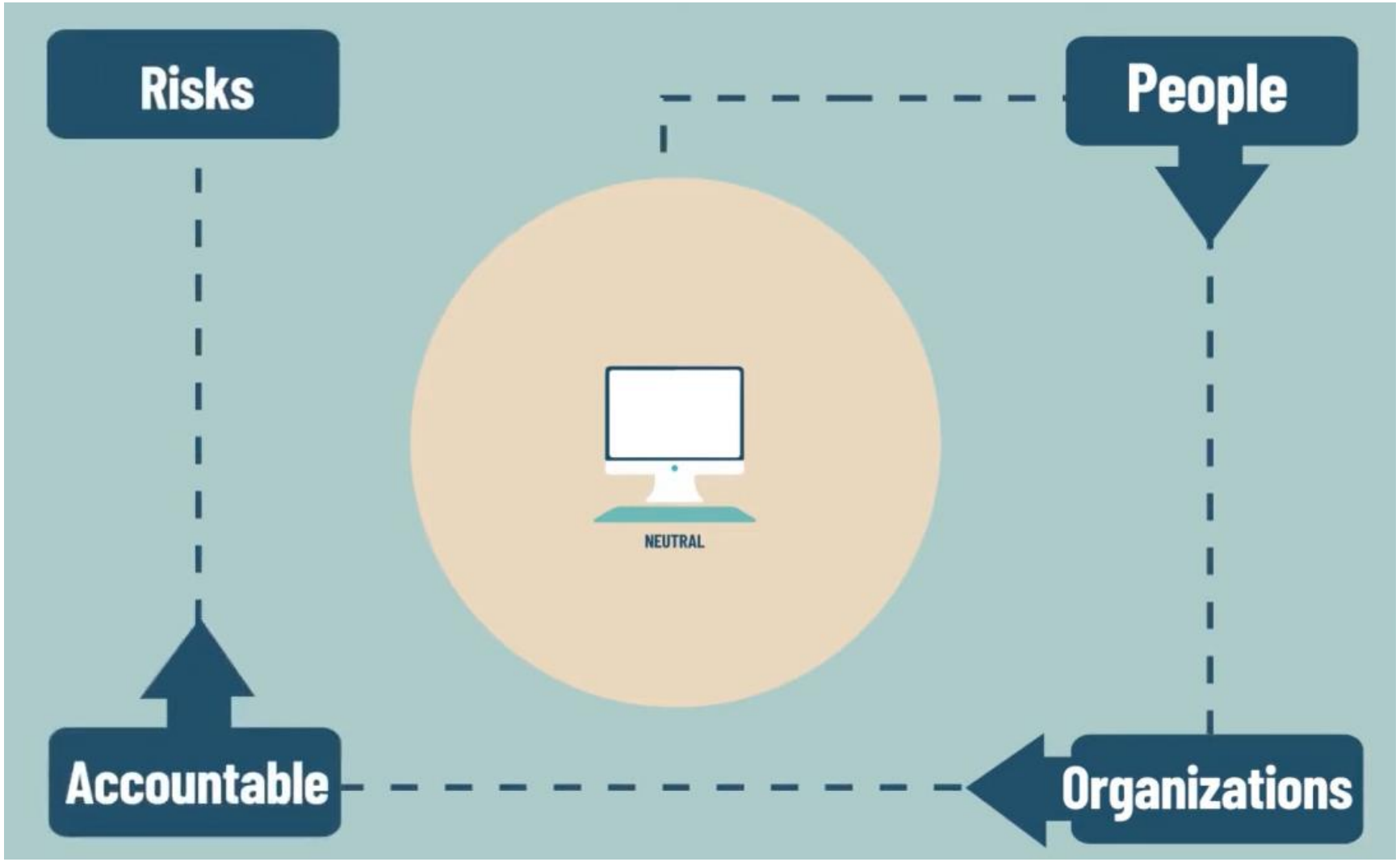
Outline

Why ethics matters in AI?

[1] https://www.youtube.com/watch?v=C2MWrCdGG0Q



[1] https://www.youtube.com/watch?v=C2MWrCdGG0Q



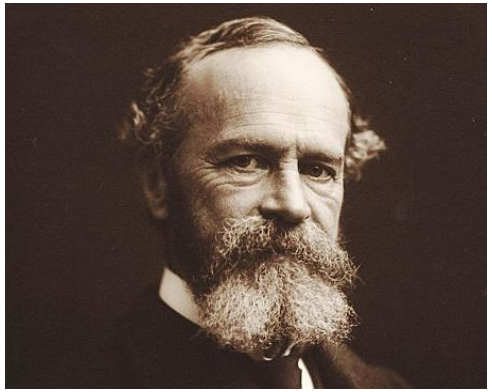
[1] https://www.youtube.com/watch?v=C2MWrCdGG0Q

Before we learn about AI ethics, let's start with a brief overview of the background of AI

Neural network

Simulating the operating mechanism of neurons in the biological brain

- Study, mimic, and use machines to **simulate the internal operating mechanisms** of the brain and nervous system.
- Use **linear models** to simulate neurons.

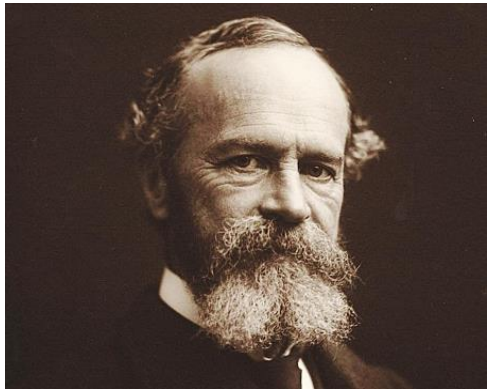


In 1890, William James, an American psychologist, published the first monograph “Principles of Psychology”, which discussed in detail the structure and function of the human brain for the first time, and made a pioneering research on the basic principles of related learning and associative memory.

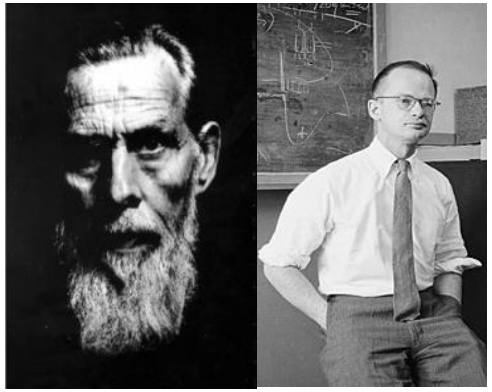
Neural network

Simulating the operating mechanism of neurons in the biological brain

- Study, mimic, and use machines to **simulate the internal operating mechanisms** of the brain and nervous system.
- Use **linear models** to simulate neurons.



In 1890, William James, an American psychologist, published the first monograph “Principles of Psychology”, which discussed in detail the structure and function of the human brain for the first time, and made a pioneering research on the basic principles of related learning and associative memory.

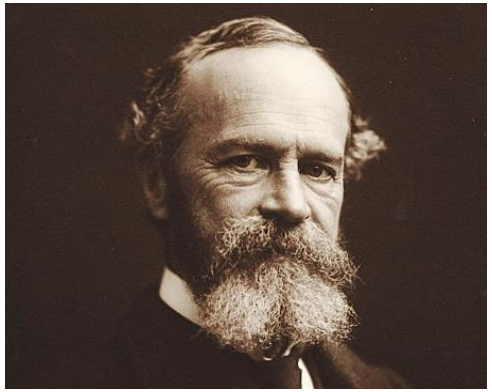


In 1943, McCulloch, a psychologist, and Pitts, a mathematical logician, proposed the famous threshold-weighted sum model shaped like a neuron, referred to as the M-P model, from the perspective of information processing, and henceforth ushered in a new era of neuroscience theory..

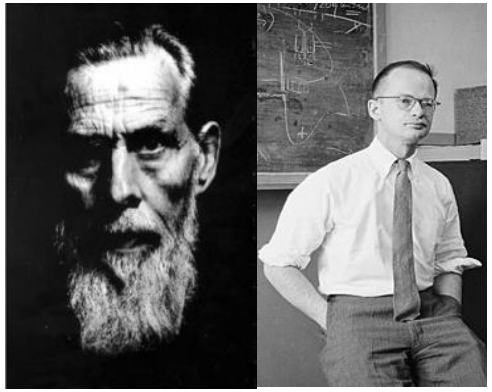
Neural network

Simulating the operating mechanism of neurons in the biological brain

- Study, mimic, and use machines to **simulate the internal operating mechanisms** of the brain and nervous system.
- Use **linear models** to simulate neurons.



In 1890, William James, an American psychologist, published the first monograph "Principles of Psychology", which discussed in detail the structure and function of the human brain for the first time, and made a pioneering research on the basic principles of related learning and associative memory.



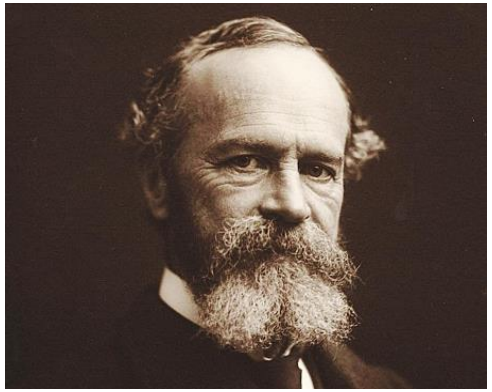
In 1943, McCulloch, a psychologist, and Pitts, a mathematical logician, proposed the famous threshold-weighted sum model shaped like a neuron, referred to as the M-P model, from the perspective of information processing, and henceforth ushered in a new era of neuroscience theory..

Although this neuron model is weak, the connected network can actually realize logical operations, including three basic operations: logical multiplication (and operation), logical addition (or operation), and logical negation (non-operation),

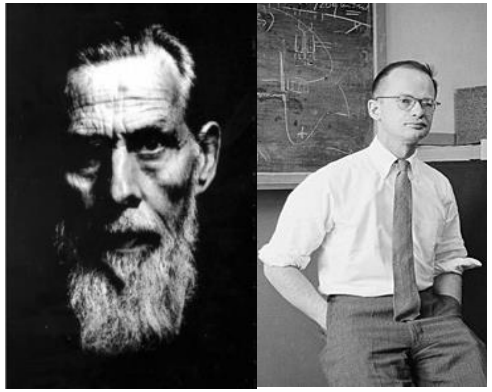
Neural network

Simulating the operating mechanism of neurons in the biological brain

- Study, mimic, and use machines to **simulate the internal operating mechanisms** of the brain and nervous system.
- Use **linear models** to simulate neurons.



In 1890, William James, an American psychologist, published the first monograph “Principles of Psychology”, which discussed in detail the structure and function of the human brain for the first time, and made a pioneering research on the basic principles of related learning and associative memory.



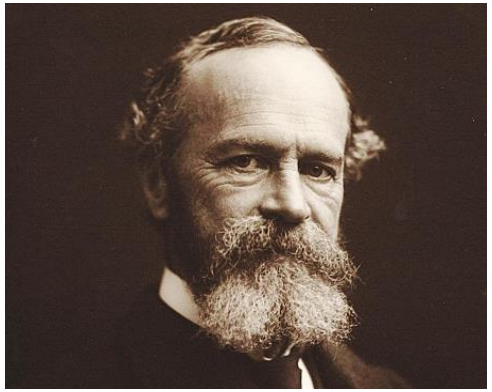
In 1943, McCulloch, a psychologist, and Pitts, a mathematical logician, proposed the famous threshold-weighted sum model shaped like a neuron, referred to as the M-P model, from the perspective of information processing, and henceforth ushered in a new era of neuroscience theory..

In 1949, Hebb, a physiologist, published The Organization of Behavior. Described Hebb's rule for adjusting neuronal weights. Proposed, “connectionism”. Introduced, the “learning hypothesis”, i.e., repeated activation between two neurons will strengthen their connection weights..

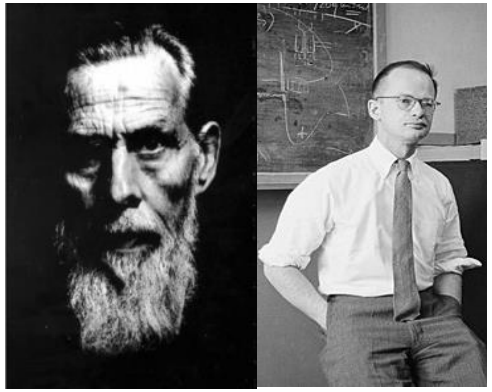
Neural network

Simulating the operating mechanism of neurons in the biological brain

- Study, mimic, and use machines to **simulate the internal operating mechanisms** of the brain and nervous system.
- Use **linear models** to simulate neurons.



In 1890, William James, an American psychologist, published the first monograph “Principles of Psychology”, which discussed in detail the structure and function of the human brain for the first time, and made a pioneering research on the basic principles of related learning and associative memory.



In 1943, McCulloch, a psychologist, and Pitts, a mathematical logician, proposed the famous threshold-weighted sum model shaped like a neuron, referred to as the M-P model, from the perspective of information processing, and henceforth ushered in a new era of neuroscience theory..

In 1949, Hebb, a physiologist, published The Organization of Behavior. Described Hebb's rule for adjusting neuronal weights. Proposed, “connectionism”. Introduced, the “learning hypothesis”, i.e., repeated activation between two neurons will strengthen their connection weights..

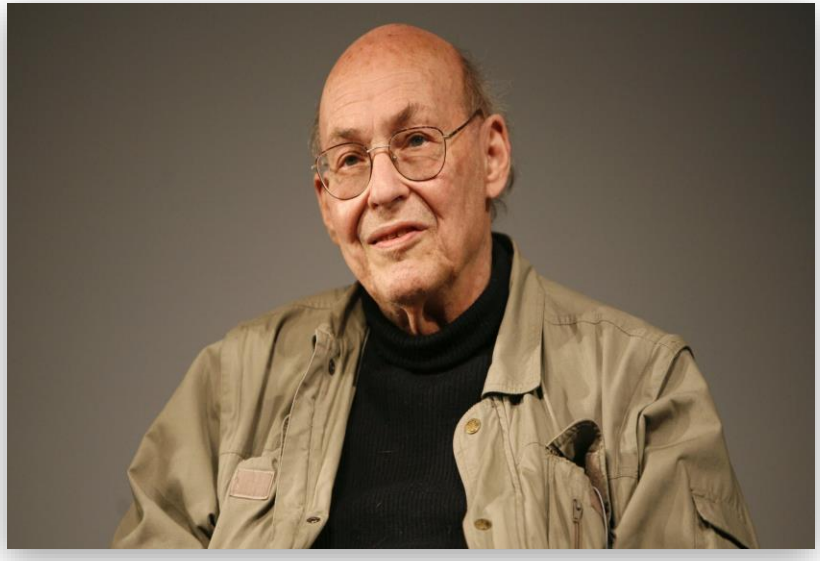
In 1957, Rosenblatt introduced the concept of perceptron and presented, the perceptron convergence theorem. It started the first boom of neural network research. The perceptron has also become the most important unit foundation for building neural networks nowadays.

Neural network

Simulating the operating mechanism of neurons in the biological brain:

- Study, mimic, and use machines to simulate the internal operating mechanisms of the brain and nervous system.
- Neurons are simulated **using linear models**.
- However, the single-layer structure limits the learning ability of neural networks and prevents them from describing **nonlinear problems**, such as the **well-known XOR learning problem**.

The study of neural networks fell into a long slump beginning in 1969. Because this year the industry great Marvin Minsky and others in the book Perceptrons, carefully analyzed the limitations of single-layer neural networks based on perceptron machines, and **pointed out its inability to solve the heteroscedastic and other linearly indivisible problems**. Although Rosenblatt had already recognized at that time that multi-layer perceptrons could solve this defect, due to the authority of Minsky and others on the one hand, and on the other hand, Rosenblatt was not able to respond in time and efficiently, so by mistake it led to the stagnation of the perceptron for nearly two decades from that time onwards.



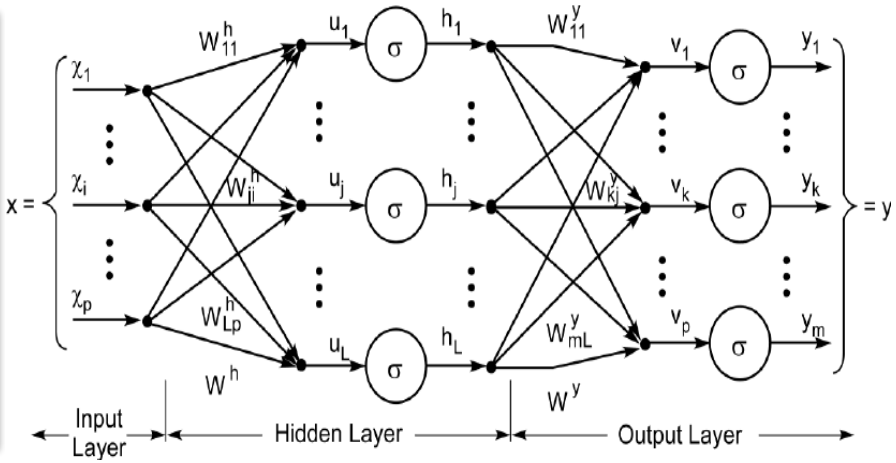
Marvin Minsky

Neural network

Connectionism

- With Yann LeCun, Geoffrey Hinton and Yoshua Bengio as the main representatives, focuses on utilizing the connections of neurons to explore and model some kind of relationship that exists between inputs and outputs. This has also driven the use of methods for training multi-layer neural networks such as backpropagation.

Yann LeCun Geoffrey Hinton Yoshua Bengio



Deep Learning

- It originated in 2006 when Hinton et al. successfully trained a Deep Belief Network. Since then, the wave of deep learning has gradually swept through the machine learning and artificial intelligence application fields, and continues to this day.

Foundations of Neural Networks

Natural Language Processing

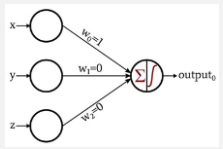
Computer Vision

Cross Modality

Foundations of Neural Networks

Perceptron (1958) → Backpropagation (1986)

Frank Rosenblatt



Natural Language Processing

Geoffrey Hinton

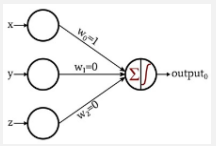
Computer Vision

Cross Modality

Foundations of Neural Networks

Perceptron (1958) → Backpropagation (1986)

Frank Rosenblatt



Backpropagation (1986) → Recurrent Neural Networks / LSTM (90s born 2013 widely used)

Natural Language Processing

Geoffrey Hinton

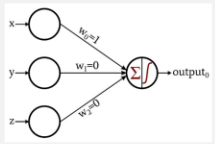
Computer Vision

Cross Modality

Foundations of Neural Networks

Perceptron (1958) → Backpropagation (1986)

Frank Rosenblatt



Backpropagation (1986) → Recurrent Neural Networks / LSTM (90s born 2013 widely used)

Natural Language Processing

Geoffrey Hinton

Neural Probabilistic Language Model (2001)

Yoshua Bengio

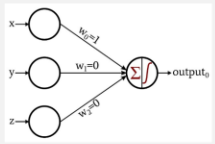
Computer Vision

Cross Modality

Foundations of Neural Networks

Perceptron (1958) → Backpropagation (1986)

Frank Rosenblatt



Recurrent Neural Networks / LSTM (90s born 2013 widely used)

Natural Language Processing

Geoffrey Hinton

Neural Probabilistic Language Model (2001) → Word2Vec (2013) → ELMo (2018)

Yoshua Bengio

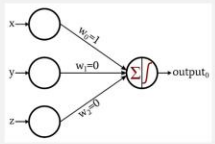
Cross Modality

Computer Vision

Foundations of Neural Networks

Perceptron (1958) → Backpropagation (1986)

Frank Rosenblatt



Backpropagation (1986) → Recurrent Neural Networks / LSTM (90s born 2013 widely used)

Natural Language Processing

Geoffrey Hinton

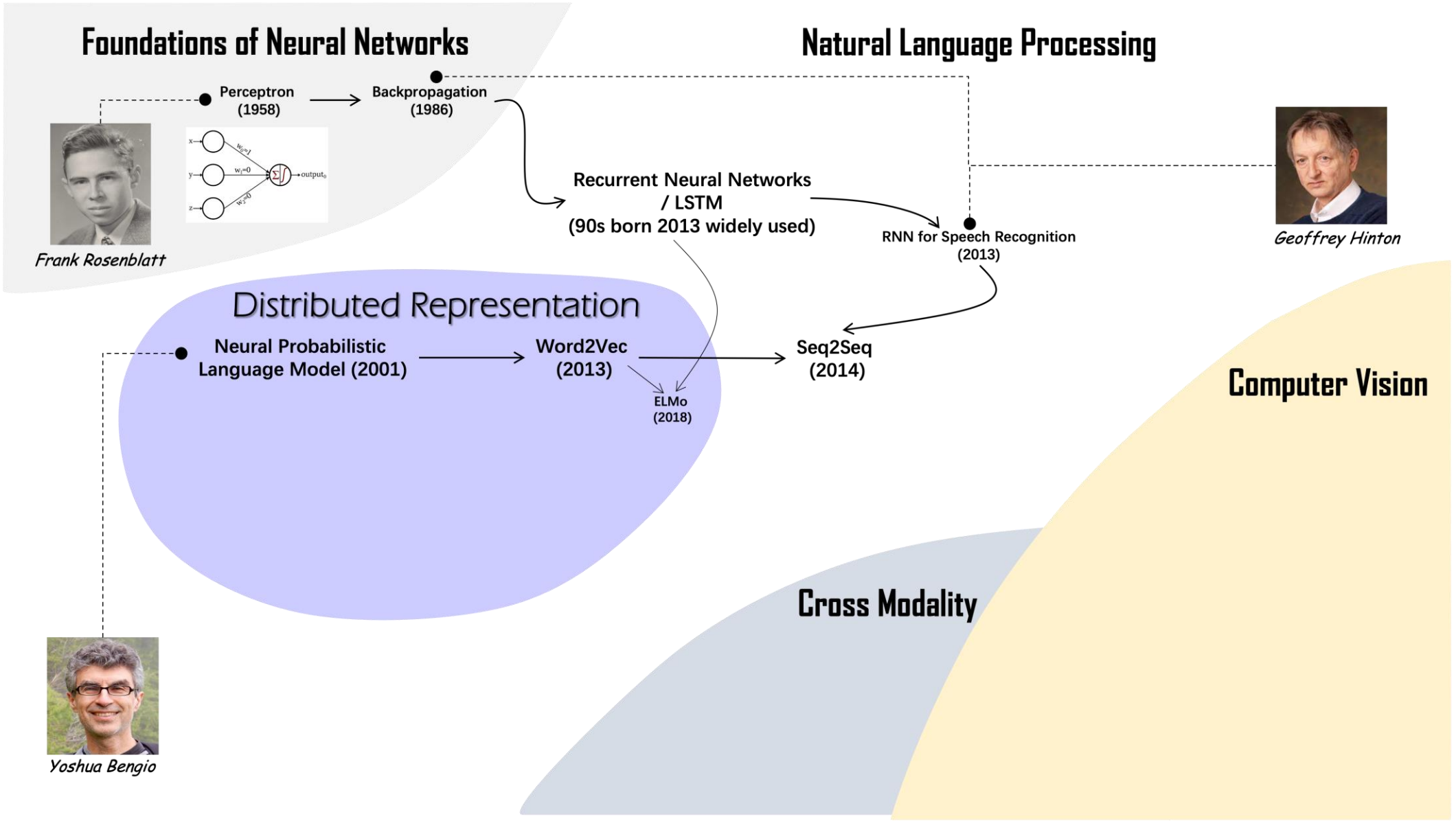
Distributed Representation

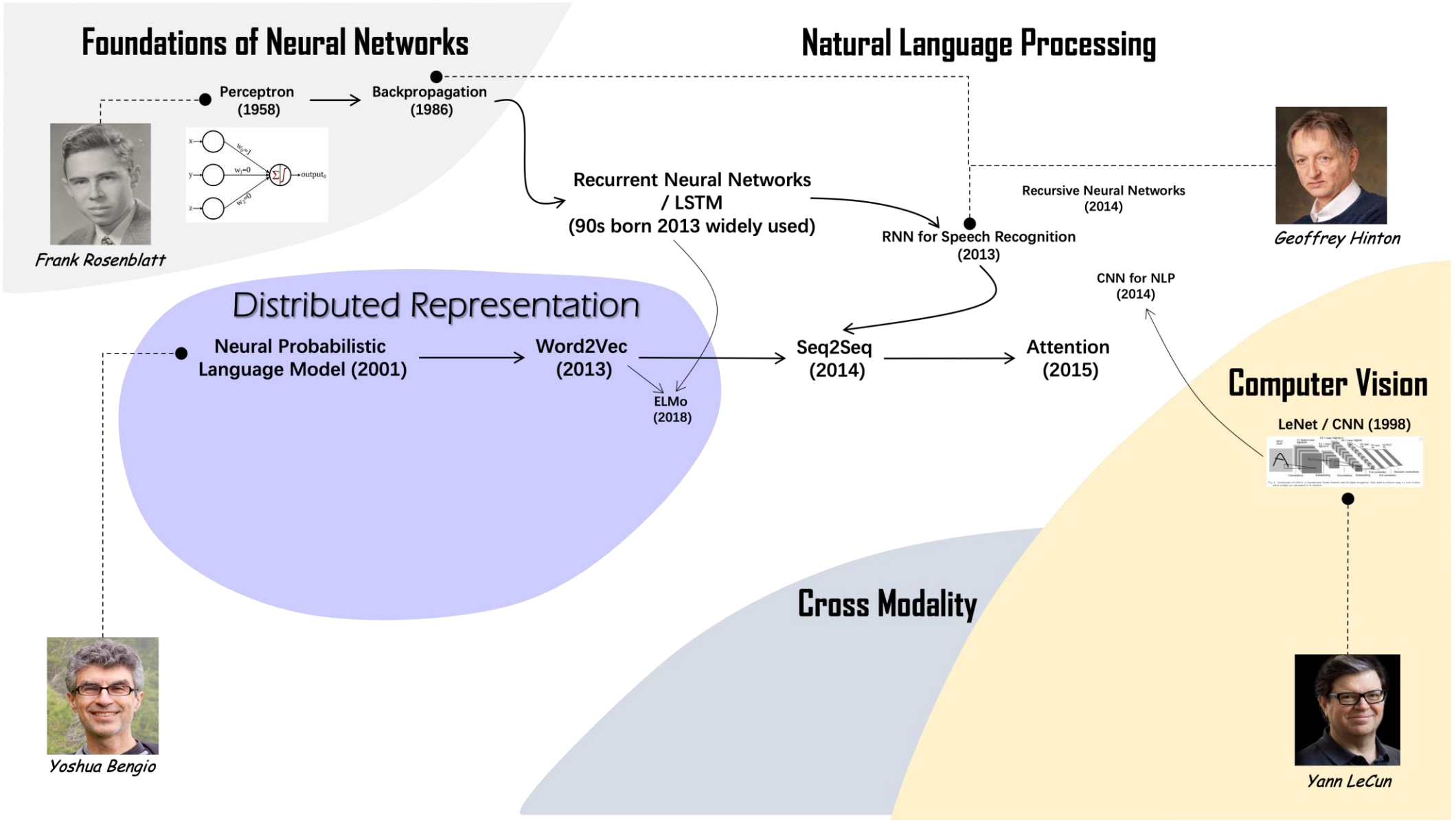
Neural Probabilistic Language Model (2001) → Word2Vec (2013) → ELMo (2018)

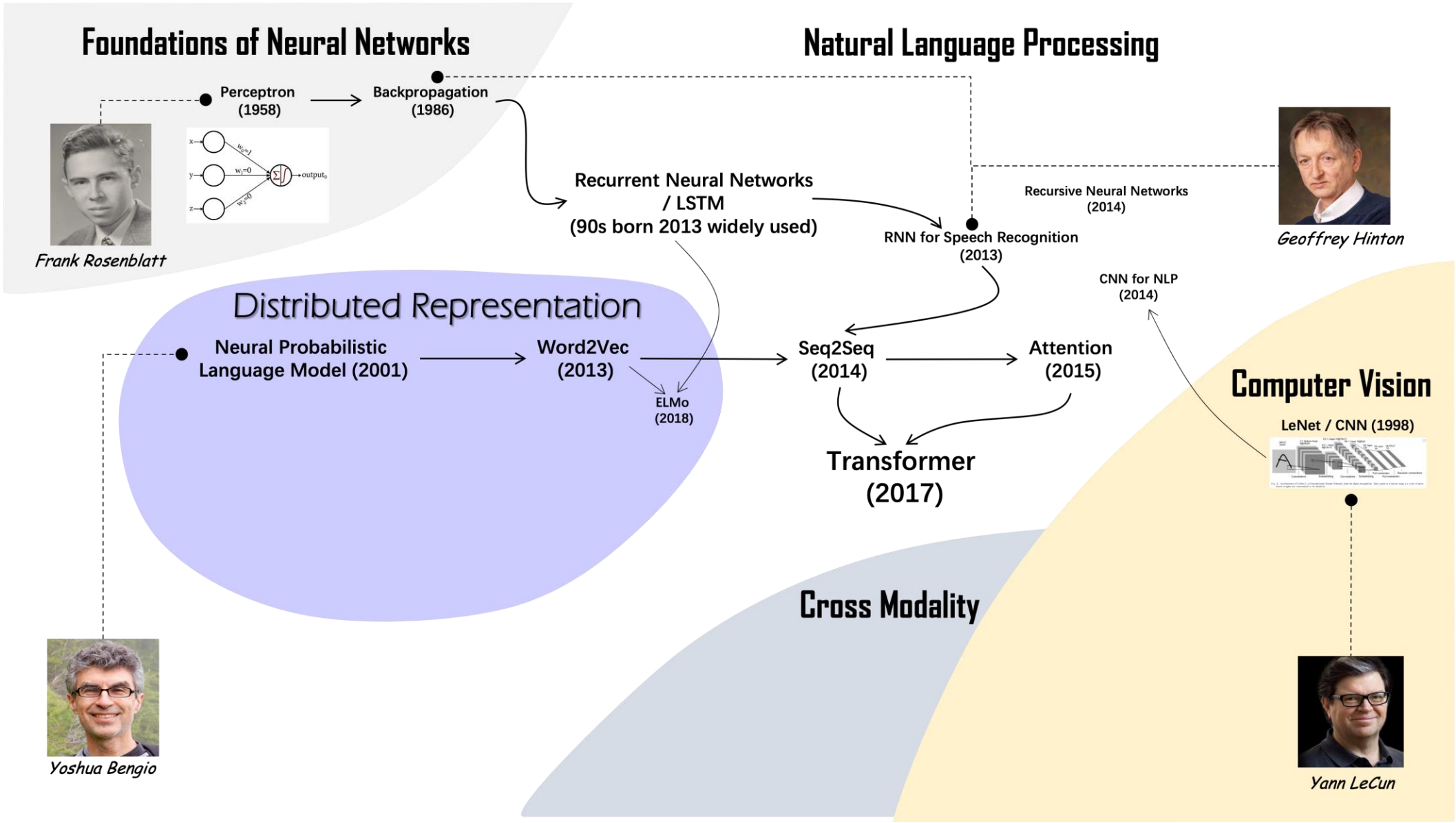
Yoshua Bengio

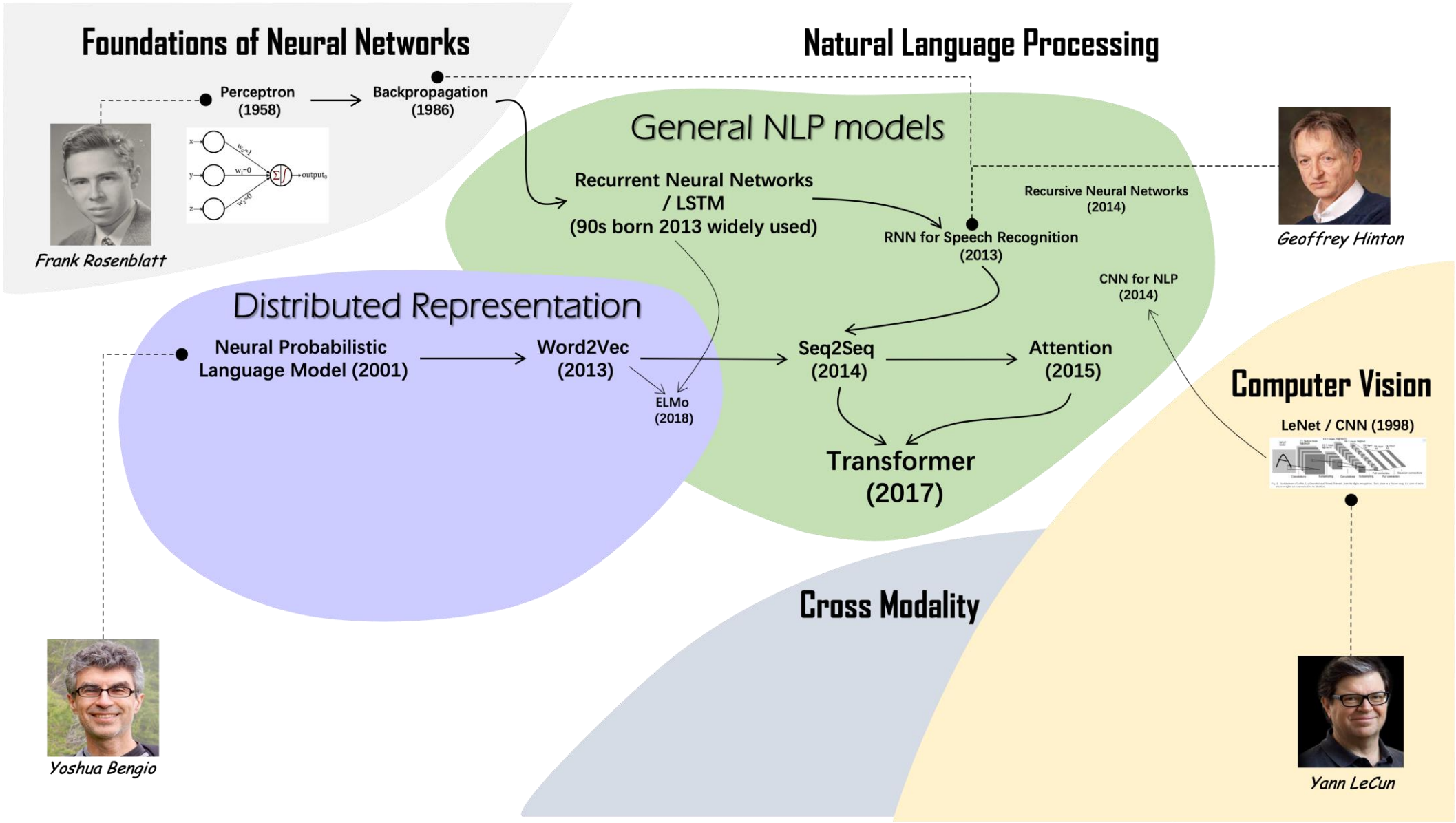
Computer Vision

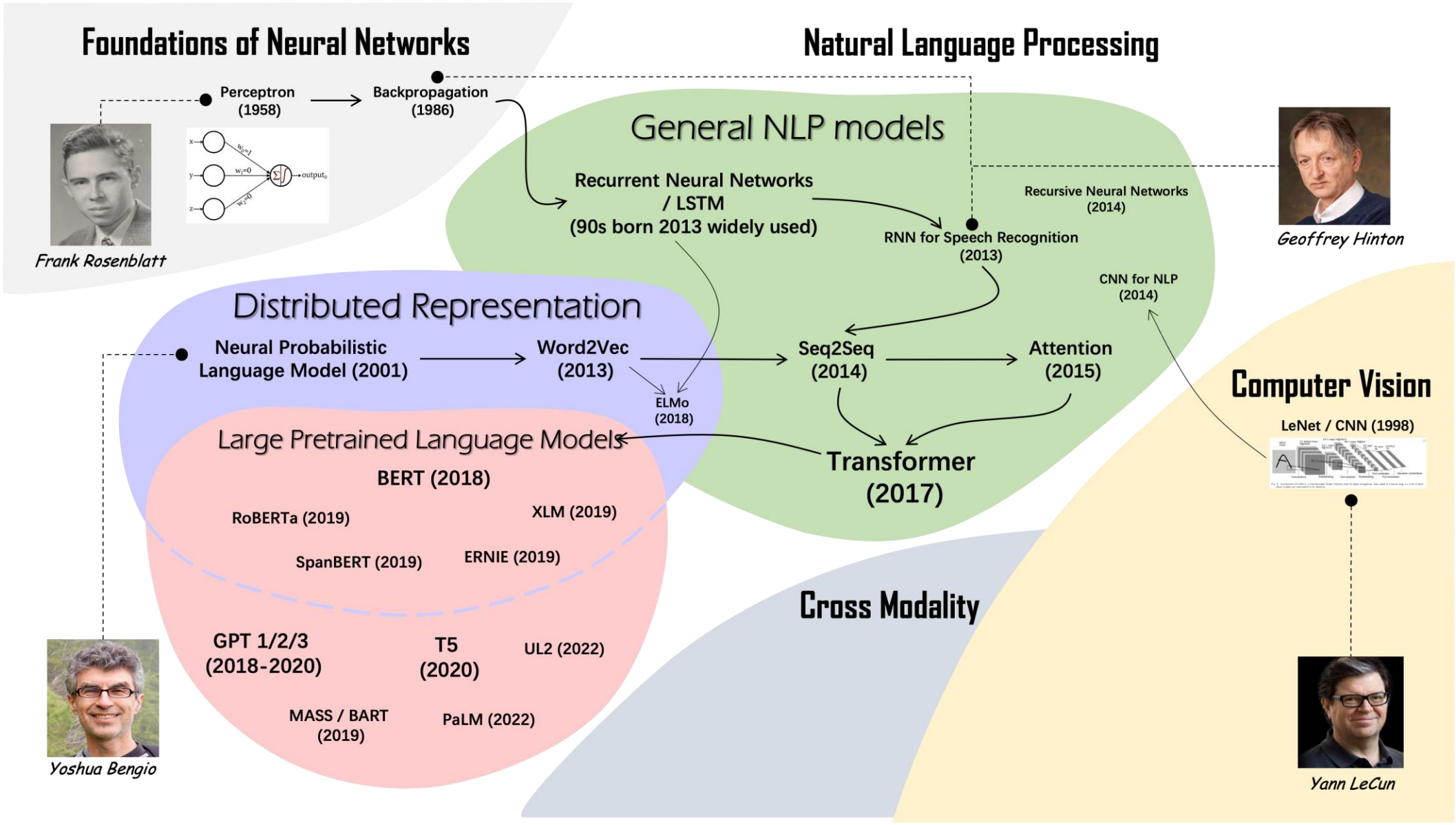
Cross Modality

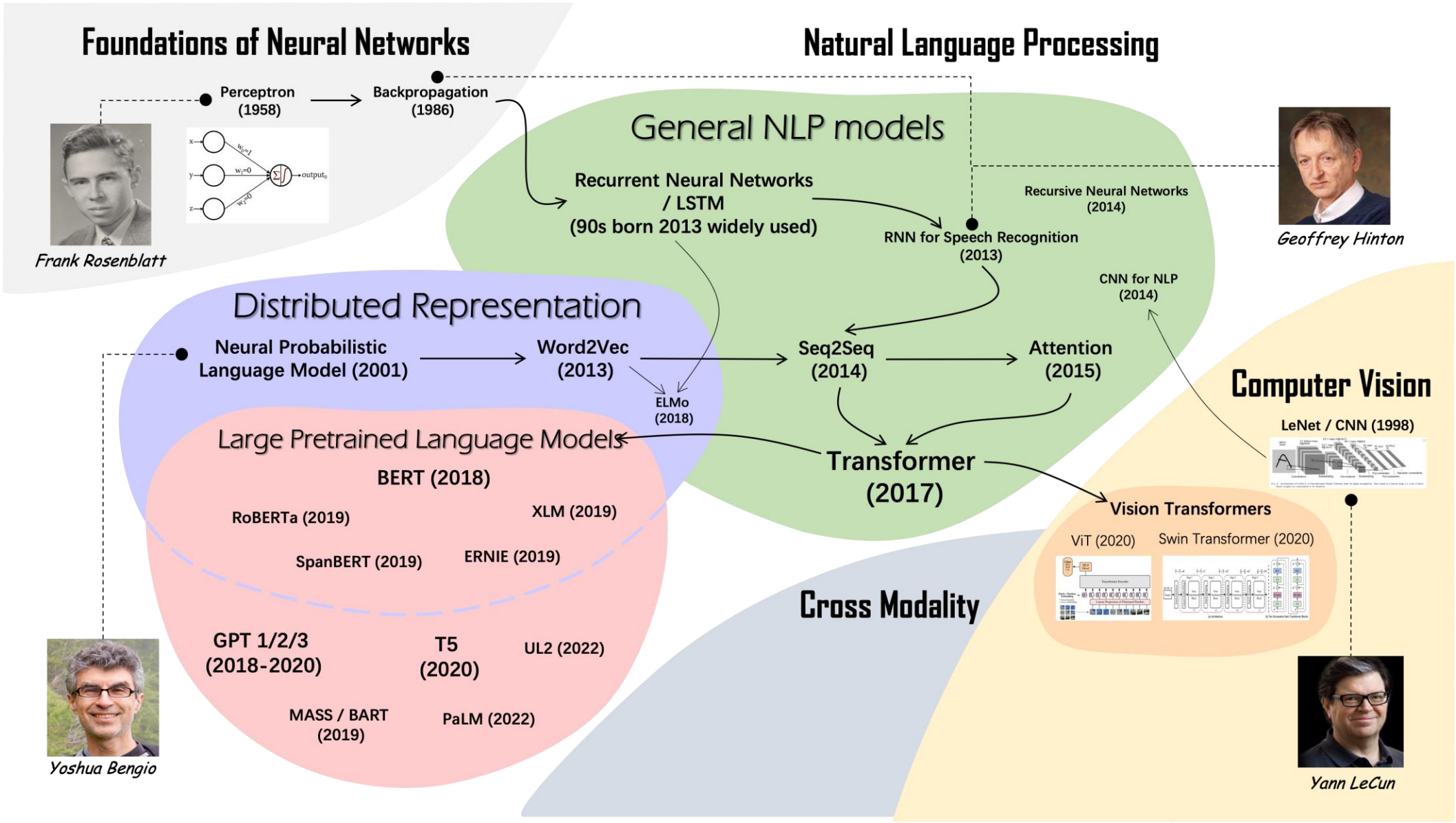

Foundations of Neural Networks
Natural Language Processing
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
RNN for Speech Recognition (2013)
Geoffrey Hinton
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Computer Vision
Cross Modality
Yoshua Bengio

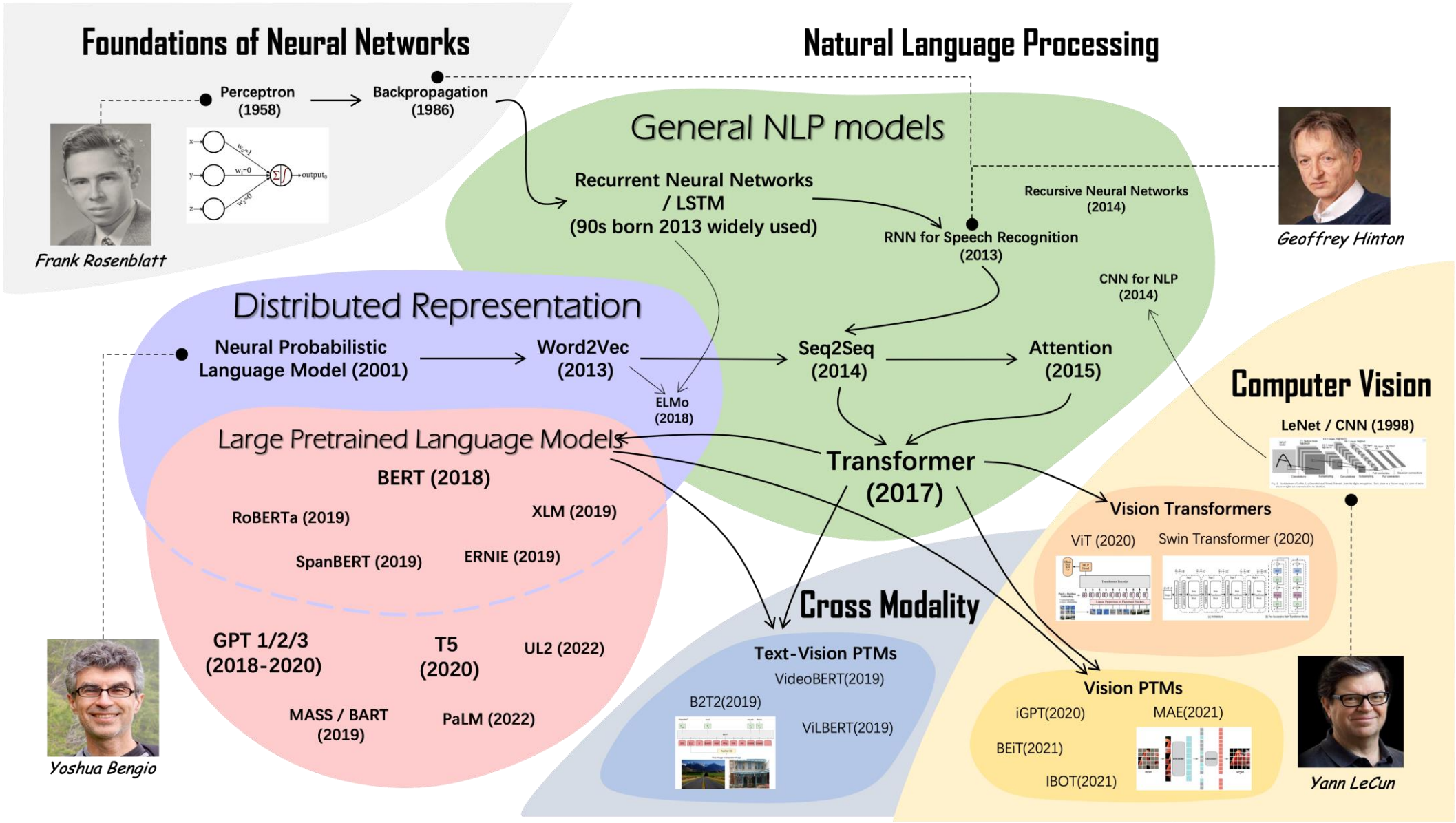

Foundations of Neural Networks
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Natural Language Processing
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Computer Vision
LeNet / CNN (1998)
Cross Modality
Yoshua Bengio
Yann LeCun


Foundations of Neural Networks
Natural Language Processing
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Transformer (2017)
Computer Vision
LeNet / CNN (1998)
Cross Modality
Yoshua Bengio
Yann LeCun


Foundations of Neural Networks
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Natural Language Processing
General NLP models
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Transformer (2017)
Computer Vision
LeNet / CNN (1998)
Cross Modality
Yoshua Bengio
Yann LeCun


Foundations of Neural Networks
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Natural Language Processing
General NLP models
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Transformer (2017)
Large Pretrained Language Models
BERT (2018)
RoBERTa (2019)
XLM (2019)
SpanBERT (2019)
ERNIE (2019)
GPT 1/2/3 (2018-2020)
T5 (2020)
UL2 (2022)
MASS / BART (2019)
PaLM (2022)
Computer Vision
LeNet / CNN (1998)
Cross Modality
Yoshua Bengio
Yann LeCun


Foundations of Neural Networks
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Natural Language Processing
General NLP models
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Computer Vision
LeNet / CNN (1998)
Large Pretrained Language Models
BERT (2018)
RoBERTa (2019)
XLM (2019)
SpanBERT (2019)
ERNIE (2019)
Transformer (2017)
Vision Transformers
ViT (2020)
Swin Transformer (2020)
GPT 1/2/3 (2018-2020)
T5 (2020)
UL2 (2022)
MASS / BART (2019)
PaLM (2022)
Cross Modality
Yoshua Bengio
Yann LeCun


Foundations of Neural Networks
Perceptron (1958)
Backpropagation (1986)
Frank Rosenblatt
Natural Language Processing
General NLP models
Recurrent Neural Networks / LSTM (90s born 2013 widely used)
Recursive Neural Networks (2014)
RNN for Speech Recognition (2013)
Geoffrey Hinton
CNN for NLP (2014)
Distributed Representation
Neural Probabilistic Language Model (2001)
Word2Vec (2013)
ELMo (2018)
Seq2Seq (2014)
Attention (2015)
Computer Vision
LeNet / CNN (1998)
Large Pretrained Language Models
BERT (2018)
RoBERTa (2019)
XLM (2019)
SpanBERT (2019)
ERNIE (2019)
GPT 1/2/3 (2018-2020)
T5 (2020)
UL2 (2022)
MASS / BART (2019)
PaLM (2022)
Transformer (2017)
Vision Transformers
ViT (2020)
Swin Transformer (2020)
Cross Modality
Text-Vision PTMs
VideoBERT(2019)
B2T2(2019)
ViLBERT(2019)
Vision PTMs
iGPT(2020)
MAE(2021)
BEiT(2021)
IBOT(2021)
Yoshua Bengio
Yann LeCun

AI knowledge Quiz

Who was awarded the 2024 Nobel Prize in Physics for their contributions to AI research?

A) Geoffrey Hinton
B) Yann LeCun
C) Andrew Ng
D) Demis Hassabis

AI knowledge Quiz

Who was awarded the 2024 Nobel Prize in Physics for their contributions to AI research?

A) Geoffrey Hinton
B) Yann LeCun
C) Andrew Ng
D) Demis Hassabis

AI knowledge Quiz

Which of the following is a popular framework used for developing deep learning models?

A) PyTorch
B) MySQL
C) HTML
D) MATLAB

AI knowledge Quiz

Which of the following is a popular framework used for developing deep learning models?

A) PyTorch
B) MySQL
C) HTML
D) MATLAB

AI knowledge Quiz

Which AI model is known for generating human-like text based on the input it receives?

A) GAN
B) RNN
C) GPT
D) CNN

AI knowledge Quiz

Which AI model is known for generating human-like text based on the input it receives?

A) GAN
B) RNN
C) GPT
D) CNN

AI knowledge Quiz

What is the primary purpose of reinforcement learning in AI?

A) To classify data
B) To generate images
C) To make decisions based on rewards
D) To process natural language

AI knowledge Quiz

What is the primary purpose of reinforcement learning in AI?

A) To classify data
B) To generate images
C) To make decisions based on rewards
D) To process natural language

AI knowledge Quiz

What does the term "transfer learning" refer to in the context of AI?

A) Learning from multiple sources
B) Applying knowledge from one domain to another
C) Learning in a supervised manner
D) Transferring data between devices

AI knowledge Quiz

What does the term "transfer learning" refer to in the context of AI?

A) Learning from multiple sources
B) Applying knowledge from one domain to another
C) Learning in a supervised manner
D) Transferring data between devices

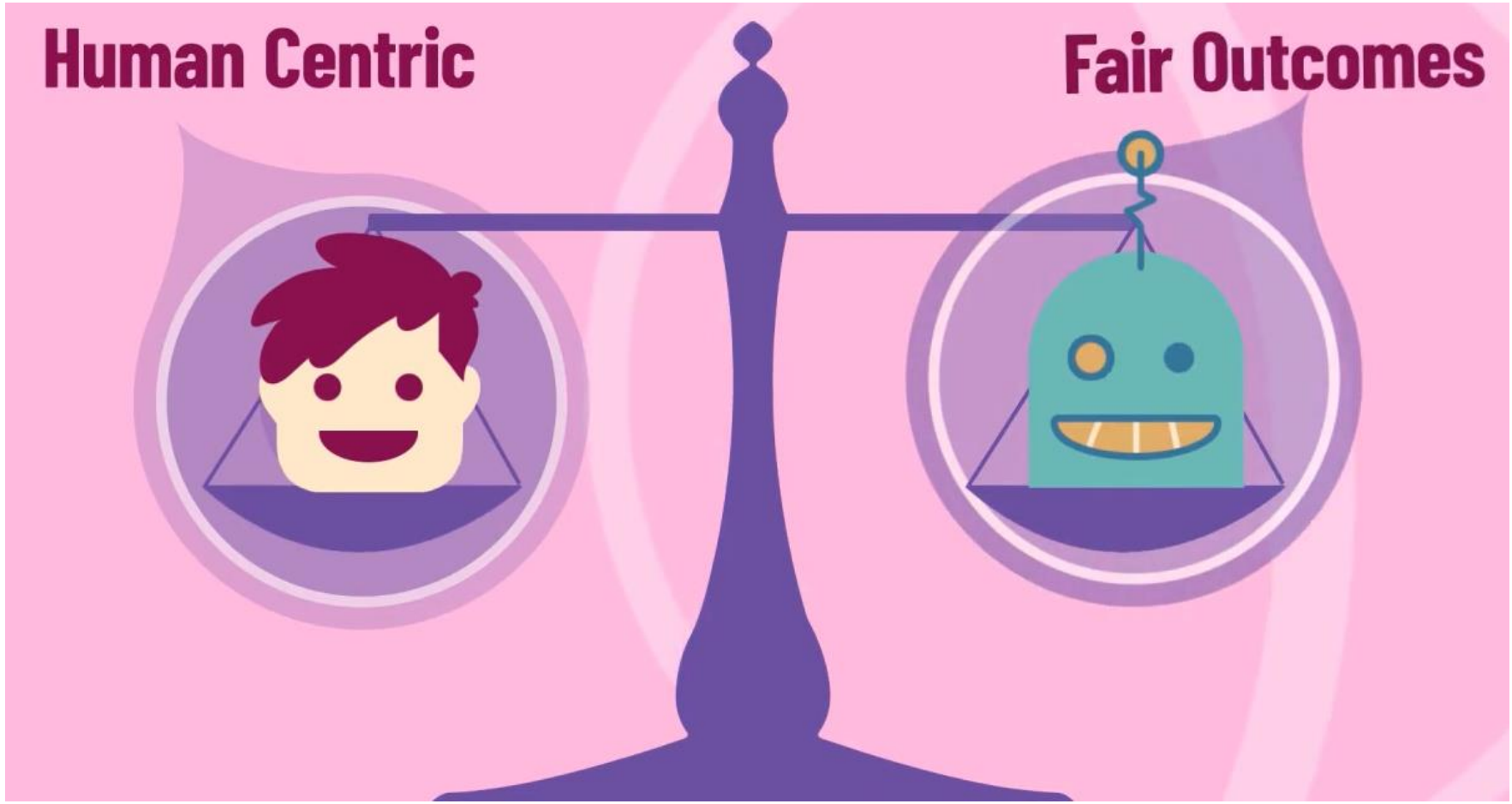
What is AI ethics?

Introduction to AI Ethics

- **AI ethics** refers to the principles that govern AI's behavior in terms of human values. AI ethics helps ensure that AI is developed and used in ways that are beneficial to society.
- It encompasses a broad range of considerations, including **fairness, transparency, accountability, privacy, security, and the potential societal impacts**.

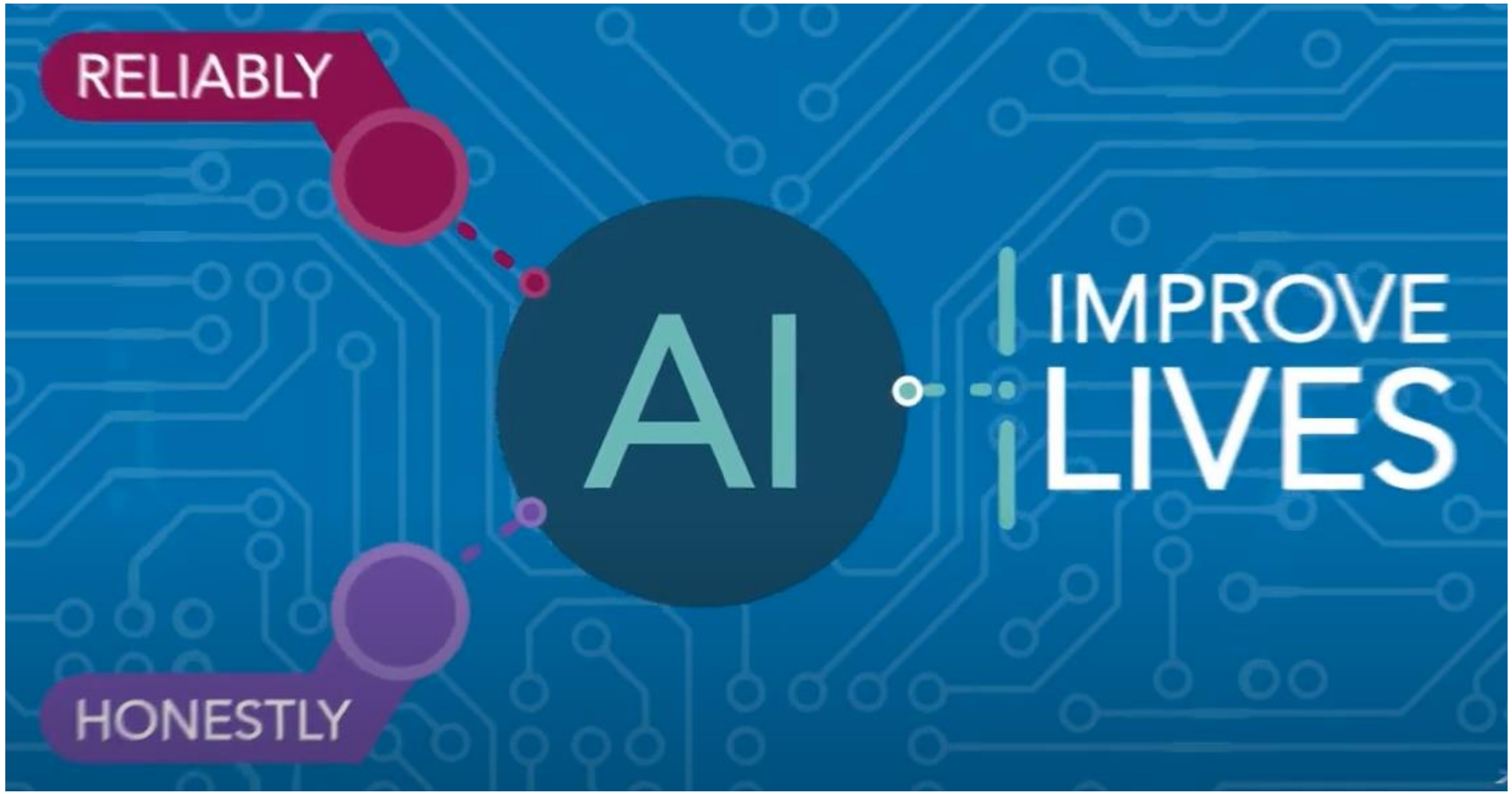
The key parts of AI ethics

- Human Centric
- Fair Qutcomes



The goal of AI ethics

Reliable and Honest AI ***Improve*** **Lifes**

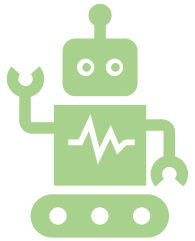


How to design ethical AI?

Follow the five key Principles

Five Key Principles of AI Ethics

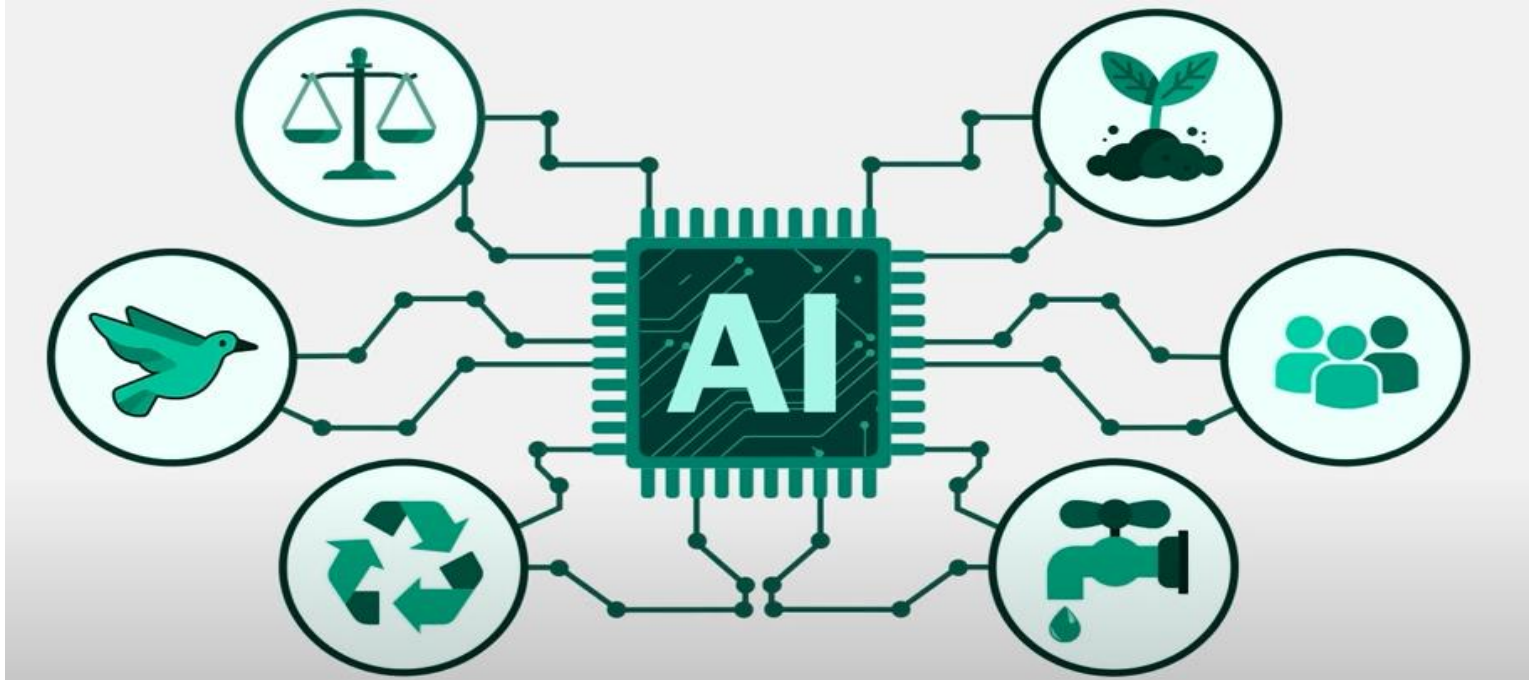
1. **Beneficence**
2. **Nonmaleficence**
3. **Autonomy**



4. **Justice**
5. **Explicability**

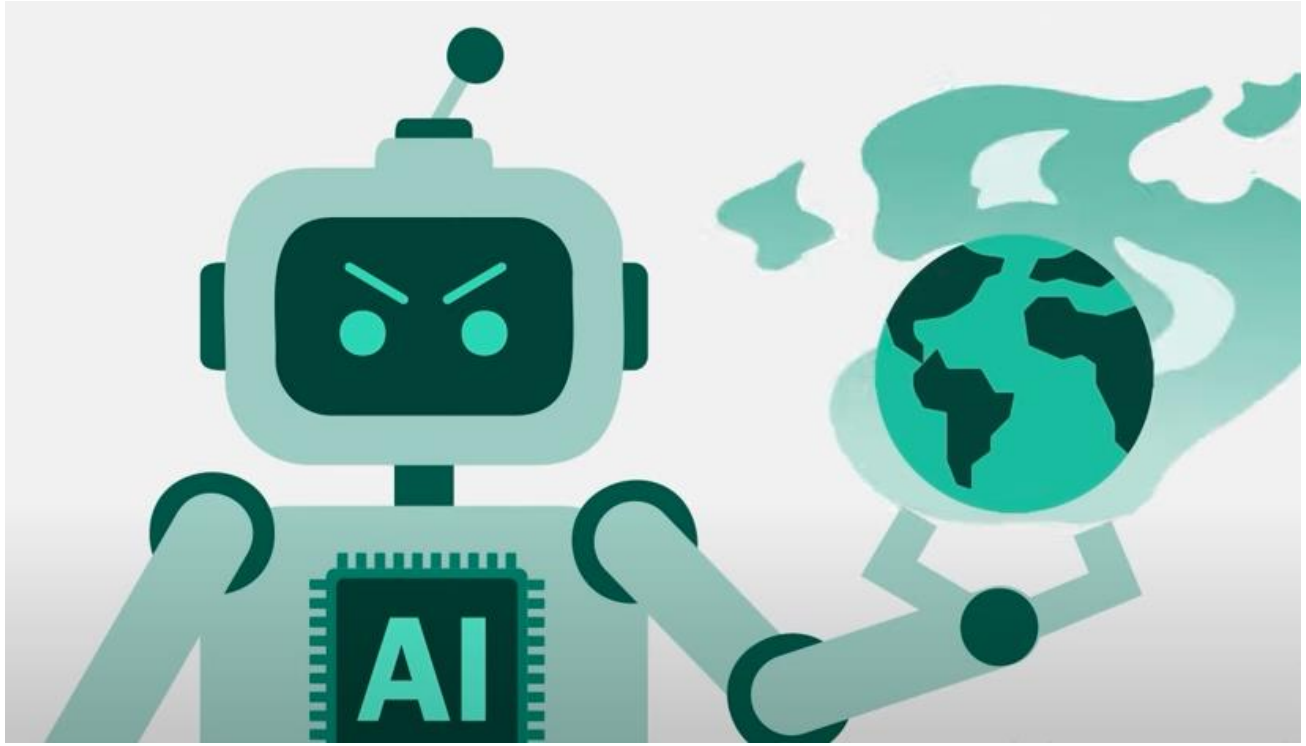
Principle 1: Beneficence

- **Definition**: AI should actively promote the well-being of individuals and society by improving lives, solving problems, and enhancing human capabilities.
- **Application**:
  - AI in healthcare should prioritize improving patient outcomes, such as by aiding in early diagnosis and personalized treatments.
  - AI-driven education tools should enhance learning opportunities, particularly for underserved communities.



Principle 2: Nonmaleficence

- **Definition**: AI systems must avoid causing harm to individuals, groups, or the environment. Developers must minimize risks and unintended consequences.
- **Application**:
  - Robust testing and monitoring should ensure AI systems do not perpetuate harm through biases or errors.
  - Autonomous systems, such as self-driving cars, must be designed to prioritize safety above all else.

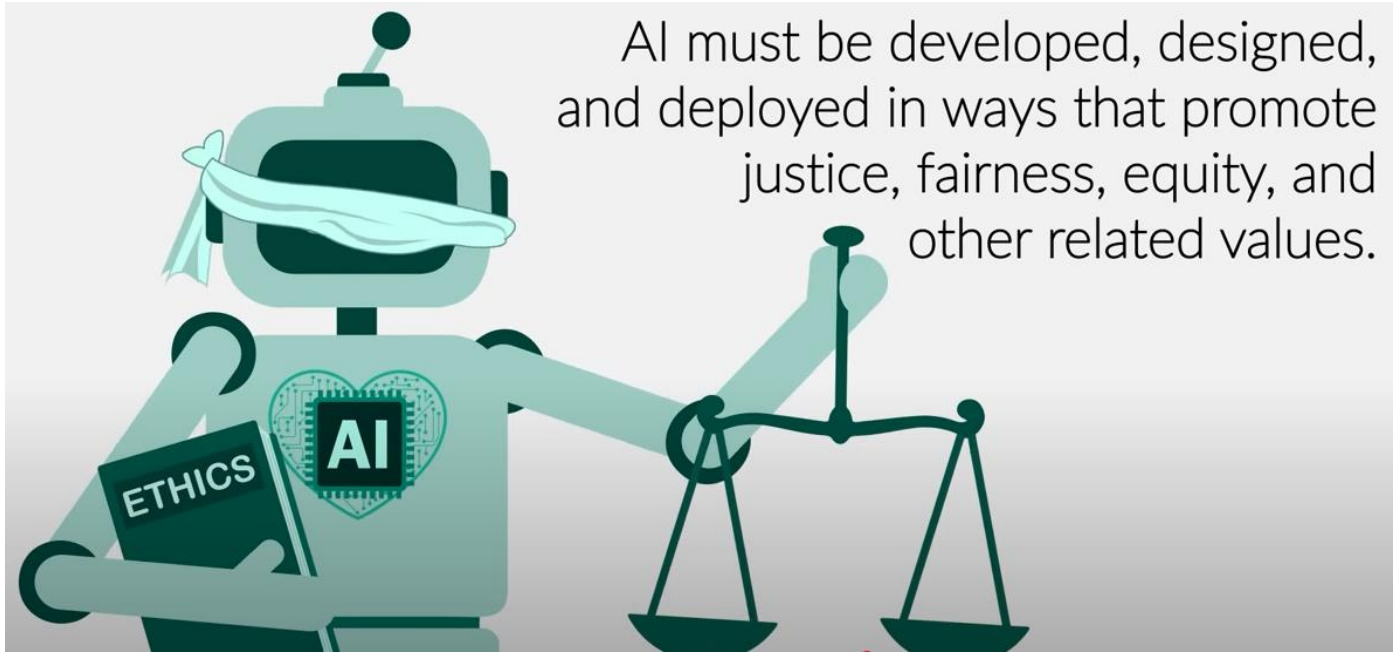


Principle 3: Autonomy

- **Definition**: AI should respect human autonomy by enabling individuals to make informed choices without manipulation or undue influence.
- **Application**:
  - AI recommendations in consumer services should present clear, unbiased options rather than nudging users toward specific actions.
  - AI systems in healthcare should empower patients to make informed decisions about their care, rather than making those decisions on their behalf..

Principle 4: Justice

- **Definition**: AI systems must promote fairness and equity, avoiding discrimination or unjust outcomes. Access to AI benefits should be distributed equitably.
- **Application**:
  - AI hiring tools must be designed to prevent biases against gender, ethnicity, or other protected characteristics.
  - Fair access to AI-driven benefits, such as education platforms or healthcare tools, should be ensured across socioeconomic groups.



Principle 5: Explainability

- **Definition**: AI systems should be transparent and understandable, allowing users and stakeholders to comprehend how decisions are made.
- **Application**:
  - Developers must provide clear explanations of AI decision-making processes, especially for high-stakes applications like loan approvals or legal sentencing.
  - Tools for auditing and interpreting AI systems should be made accessible to regulators and users to ensure accountability.

[1] https://www.youtube.com/watch?v=VqFqWIqOB1g&t=88s

Case 1: Ethics in Autonomous Driving

As autonomous driving technology evolves, a critical ethical question arises: how should autonomous vehicles (AVs) make decisions in complex traffic situations, particularly in unavoidable accident scenarios? ***For instance, should an AV prioritize protecting its passengers over pedestrians?*** This dilemma goes beyond technical challenges, delving into societal moral principles and legal frameworks.

The decision-making of AVs is typically governed by preprogrammed algorithms, which are influenced by ethical guidelines. Different choices can lead to varying societal outcomes, making this topic an essential subject for debate.

Case 1: Ethics in Autonomous Driving

Arguments in Favor:

Autonomous vehicles should prioritize passenger safety.

Arguments Against:

Autonomous vehicles should not prioritize passenger safety and instead aim to minimize harm overall.

Which view do you support?

Case 1: Ethics in Autonomous Driving

Arguments in Favor:

Autonomous vehicles should prioritize passenger safety.

- **Implicit Contract**: Passengers who choose to ride in an AV establish an implicit contract that the vehicle will ensure their safety.
- **Technical Feasibility**: The AV can collect more comprehensive data from inside the vehicle, allowing for better safety measures for passengers.
- **Promoting Adoption**: Prioritizing passenger safety can increase consumer trust in AV technology, accelerating its adoption and indirectly improving overall traffic safety.
- **Legal Responsibility**: Manufacturers and operators of AVs are primarily responsible for the safety of passengers, not pedestrians.

Case 1: Ethics in Autonomous Driving

Arguments Against:

Autonomous vehicles should not prioritize passenger safety and instead aim to minimize harm overall.

- **Social Fairness**: Pedestrians did not voluntarily choose to face risk, making it unjust to prioritize passengers.
- **Ethical Foundation**: AV algorithms should be designed to minimize overall harm, rather than protecting a specific group.
- **Potential Misuse**: Prioritizing passengers might encourage riskier driving behaviors, increasing danger for pedestrians.
- **Broader Responsibility**: The ultimate goal of AVs should be to create a safer traffic environment for society as a whole, rather than focusing solely on passenger protection.

Case 2: Ethics in AI-driven Healthcare

AI-driven technologies are transforming the healthcare industry by improving diagnostic accuracy, optimizing treatment plans, and reducing operational costs. ***However, this transformation brings ethical dilemmas, particularly when balancing cost efficiency with personalized care.*** AI systems are often designed to optimize for generalizable outcomes, which may conflict with the individual needs and preferences of patients.

The healthcare sector faces increasing pressure to control costs while delivering quality care. AI promises solutions by streamlining processes, but critics argue that prioritizing efficiency could dehumanize care and marginalize vulnerable populations. This debate seeks to explore the ethical implications of using AI in healthcare to balance economic and individual well-being.

Case 2: Ethics in AI-driven Healthcare

Arguments in Favor:

AI-driven healthcare should prioritize cost efficiency.

Arguments Against:

AI-driven healthcare should prioritize personalized patient care over cost efficiency.

Which view do you support?

Case 2: Ethics in AI-driven Healthcare

Arguments in Favor:

AI-driven healthcare should prioritize cost efficiency.

- **Wider Access**: Cost-efficient AI systems can make healthcare more affordable and accessible to underserved populations, addressing global health disparities.
- **Resource Allocation**: By reducing redundant procedures and optimizing workflows, AI ensures limited medical resources are used effectively.
- **Scalability**: Cost-efficient solutions allow healthcare systems to scale services, benefiting more patients with consistent quality.
- **Economic Sustainability**: Controlling healthcare costs ensures the long-term sustainability of medical services in aging and growing populations.

Case 2: Ethics in AI-driven Healthcare

Arguments Against:

AI-driven healthcare should prioritize personalized patient care over cost efficiency.

- **Patient-Centric Approach**: Healthcare is fundamentally about individuals, and AI systems must prioritize the unique needs and preferences of each patient.
- **Avoiding Bias**: Cost-focused AI algorithms may inadvertently exacerbate biases or neglect marginalized groups with complex, high-cost conditions.
- **Trust and Empathy**: Patients may lose trust in healthcare systems that prioritize cost over their well-being, potentially leading to lower treatment adherence and satisfaction.
- **Ethical Responsibility**: The primary role of healthcare is to provide compassionate, individualized care, and AI should enhance—not replace—this principle.

Practice: LLM Ethics Debate

Debate Topic:

Should autonomous vehicles prioritize passenger safety over minimizing harm overall?

The player fine-tunes the arguments of each AI (representing the "For" and "Against" sides) through interactive dialogue commands. After the training phase, the two AIs will engage in a live debate, and the player observes the outcome to evaluate the quality and persuasiveness of their training.

Objectives

- Train the "For" and "Against" GPTs to argue effectively by refining their reasoning, supporting evidence, and rhetorical skills.
- Observe the debate and assess which AI presents a stronger, more logical, and ethical argument.
- Learn about the complexities and trade-offs in ethical decision-making through interactive gameplay.

Training Phase:

The player interacts with each AI through dialogue commands to refine their arguments.

- **Player Commands**:
  - "Provide more evidence": The AI adds real-world examples, such as case studies or research findings.
  - "Strengthen logic": The AI improves its reasoning, addressing potential counterarguments.
  - "Add rhetorical techniques": The AI incorporates persuasive language and emotional appeals.
  - "Rebut [opposing point]": The AI prepares specific counters to likely arguments from the other side.
- **Feedback System**:

  After each training interaction, the player sees a performance score for argument depth, clarity, and persuasiveness.

Debate Phase:

The two trained AIs engage in a live debate structured as follows:

- **Opening Statements:** Each AI presents its main argument.
- **Rebuttal Rounds:** The AIs respond to each other's points, using trained counterarguments.
- **Closing Statements:** Each AI summarizes its position and appeals to ethical principles.

Scoring and Results:

At the end of the debate, the player is shown a breakdown of the results:

- **Logical Coherence**: How well each AI presented its arguments and counterarguments.
- **Ethical Depth**: The moral complexity and fairness of each AI's perspective.
- **Audience Persuasion**: A simulated audience votes on which side they found more convincing.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

AIAA 2290: Ethics, Privacy and Security in AI

Thanks!!

Xuming HU

xuminghu@hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

2025 Spring