人工智慧
(Artificial Intelligence)

人工智慧哲學與倫理，人工智慧的未來
(Philosophy and Ethics of AI, The Future of AI)

1092AI12
MBA, IM, NTPU (M5010) (Spring 2021)
Wed 2, 3, 4 (9:10-12:00) (B8F40)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Institute of Information Management, National Taipei University
國立臺北大學 資訊管理研究所
https://web.ntpu.edu.tw/~myday

2021-06-09

課程大綱 (Syllabus)

國立臺北大學 National Taipei University

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2021/02/24	人工智慧概論 (Introduction to Artificial Intelligence)
2	2021/03/03	人工智慧和智慧代理人 (Artificial Intelligence and Intelligent Agents)
3	2021/03/10	問題解決 (Problem Solving)
4	2021/03/17	知識推理和知識表達 (Knowledge, Reasoning and Knowledge Representation)
5	2021/03/24	不確定知識和推理 (Uncertain Knowledge and Reasoning)
6	2021/03/31	人工智慧個案研究 I (Case Study on Artificial Intelligence I)

課程大綱 (Syllabus)

國立臺北大學 National Taipei University

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2021/04/07	放假一天 (Day off)
8	2021/04/14	機器學習與監督式學習 (Machine Learning and Supervised Learning)
9	2021/04/21	期中報告 (Midterm Project Report)
10	2021/04/28	學習理論與綜合學習 (The Theory of Learning and Ensemble Learning)
11	2021/05/05	深度學習 (Deep Learning)
12	2021/05/12	人工智慧個案研究 II (Case Study on Artificial Intelligence II)

課程大綱 (Syllabus)

國立臺北大學 National Taipei University

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2021/05/19	強化學習 (Reinforcement Learning)
14	2021/05/26	深度學習自然語言處理 (Deep Learning for Natural Language Processing)
15	2021/06/02	機器人技術 (Robotics)
16	2021/06/09	人工智慧哲學與倫理，人工智慧的未來 (Philosophy and Ethics of AI, The Future of AI)
17	2021/06/16	期末報告 I (Final Project Report I)
18	2021/06/23	期末報告 II (Final Project Report II)

Philosophy and Ethics of AI, The Future of AI

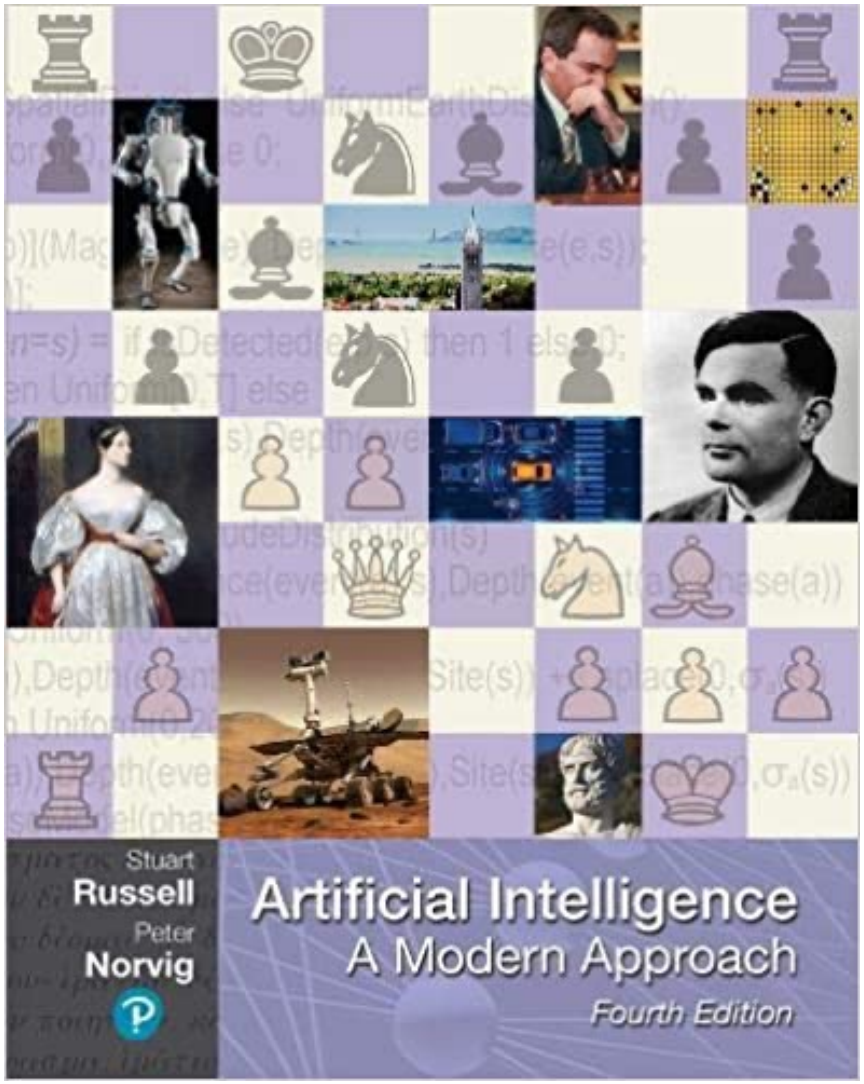
Outline

- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach

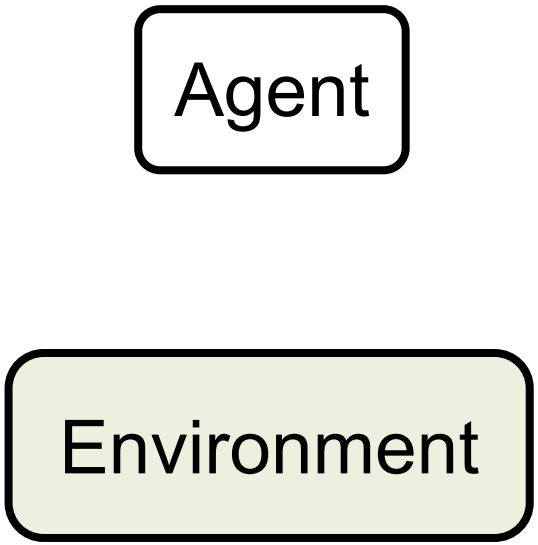
1. Artificial Intelligence
2. Problem Solving
3. Knowledge and Reasoning
4. Uncertain Knowledge and Reasoning
5. Machine Learning
6. Communicating, Perceiving, and Acting
7. Philosophy and Ethics of AI

Philosophy and Ethics of AI

Artificial Intelligence:
7. Philosophy and Ethics of AI

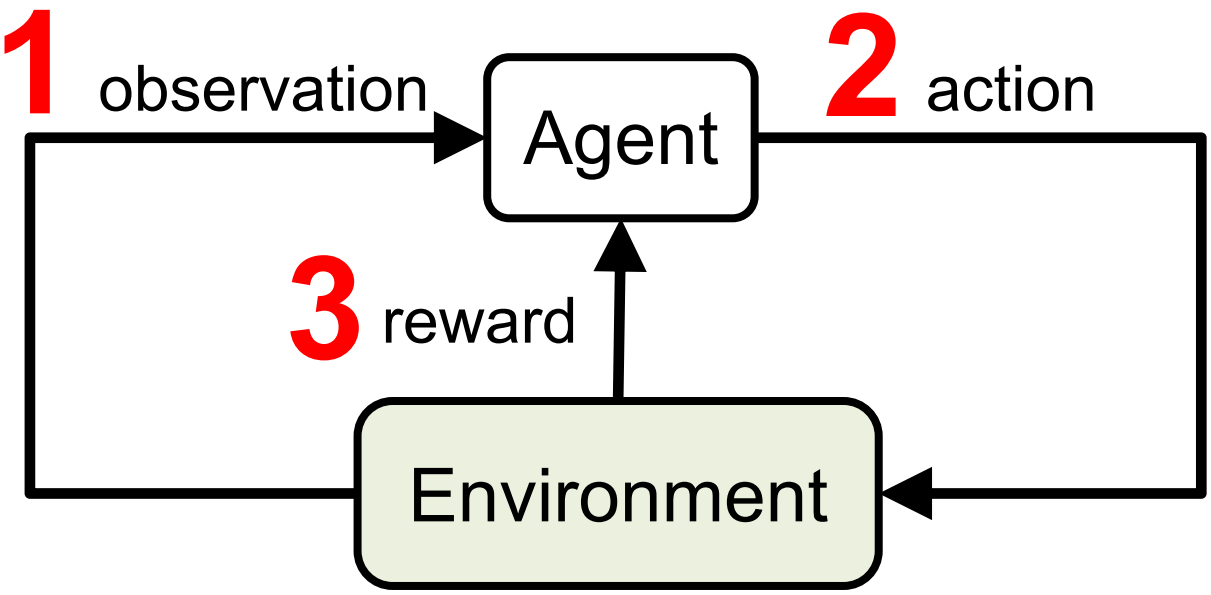
- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

Reinforcement Learning (DL)



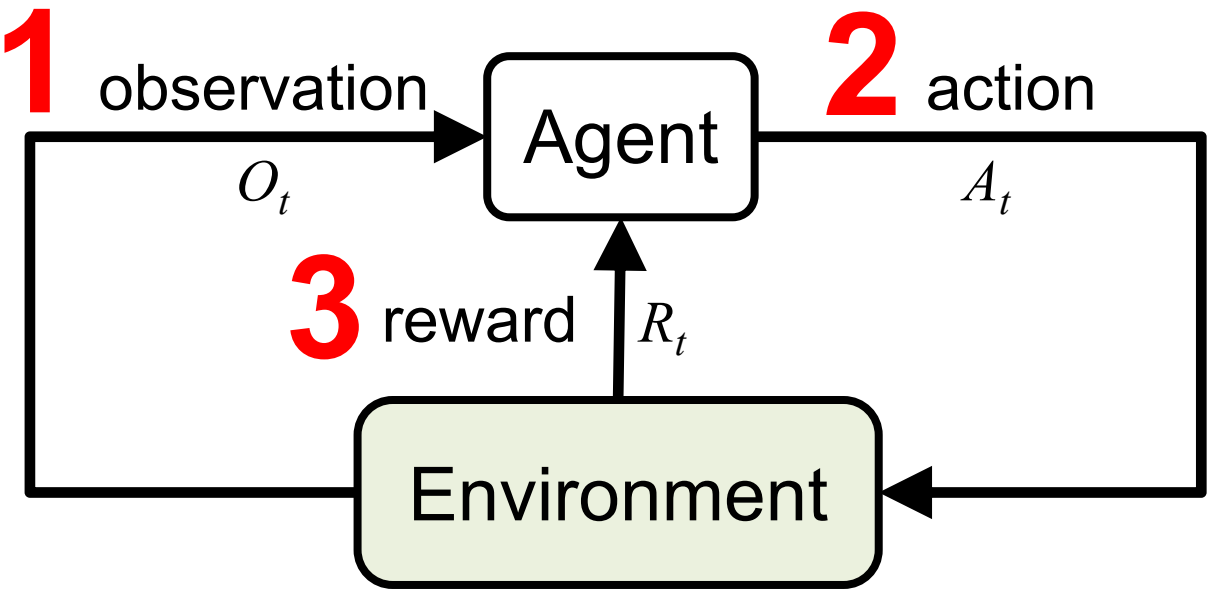
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



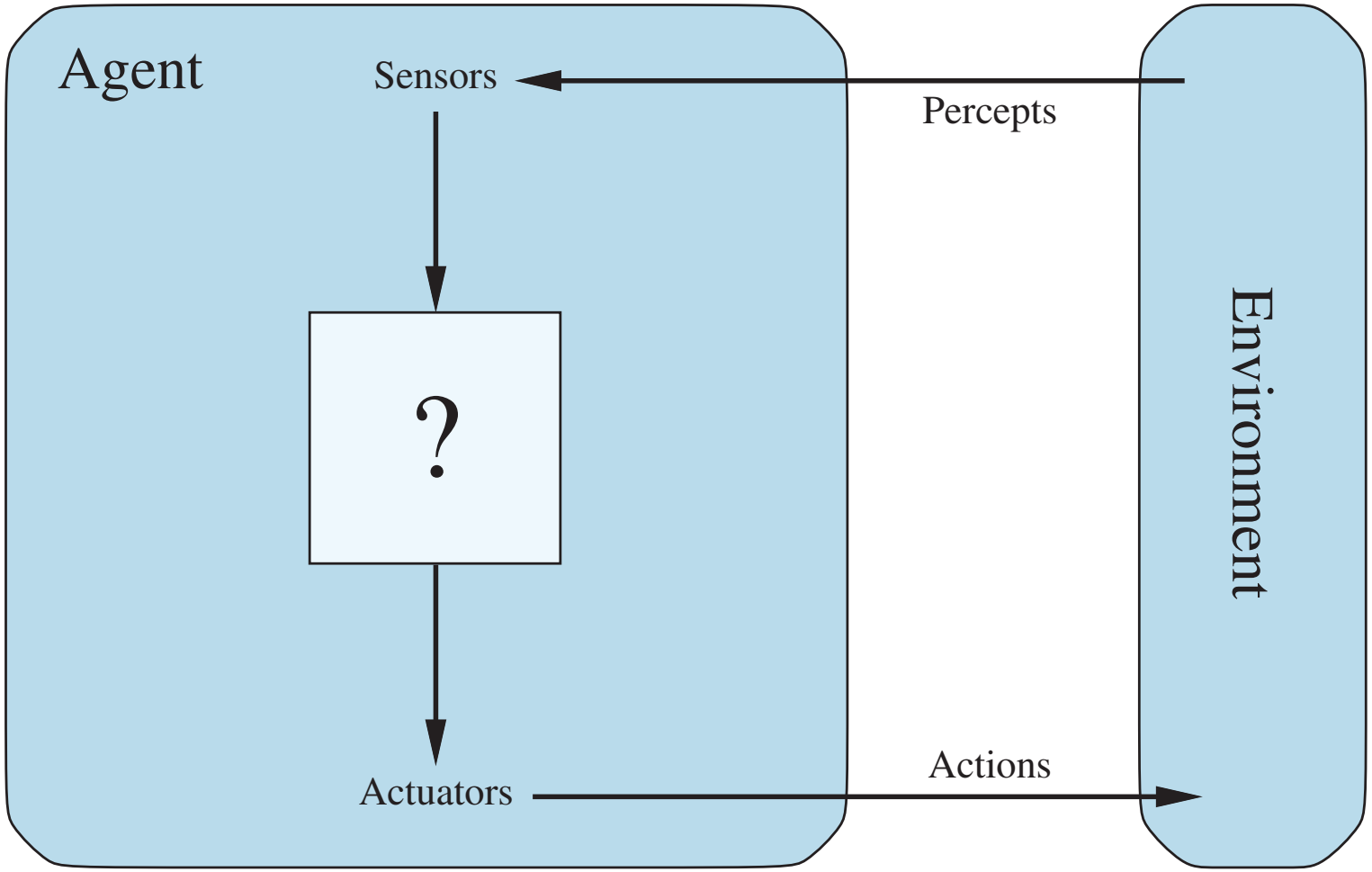
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



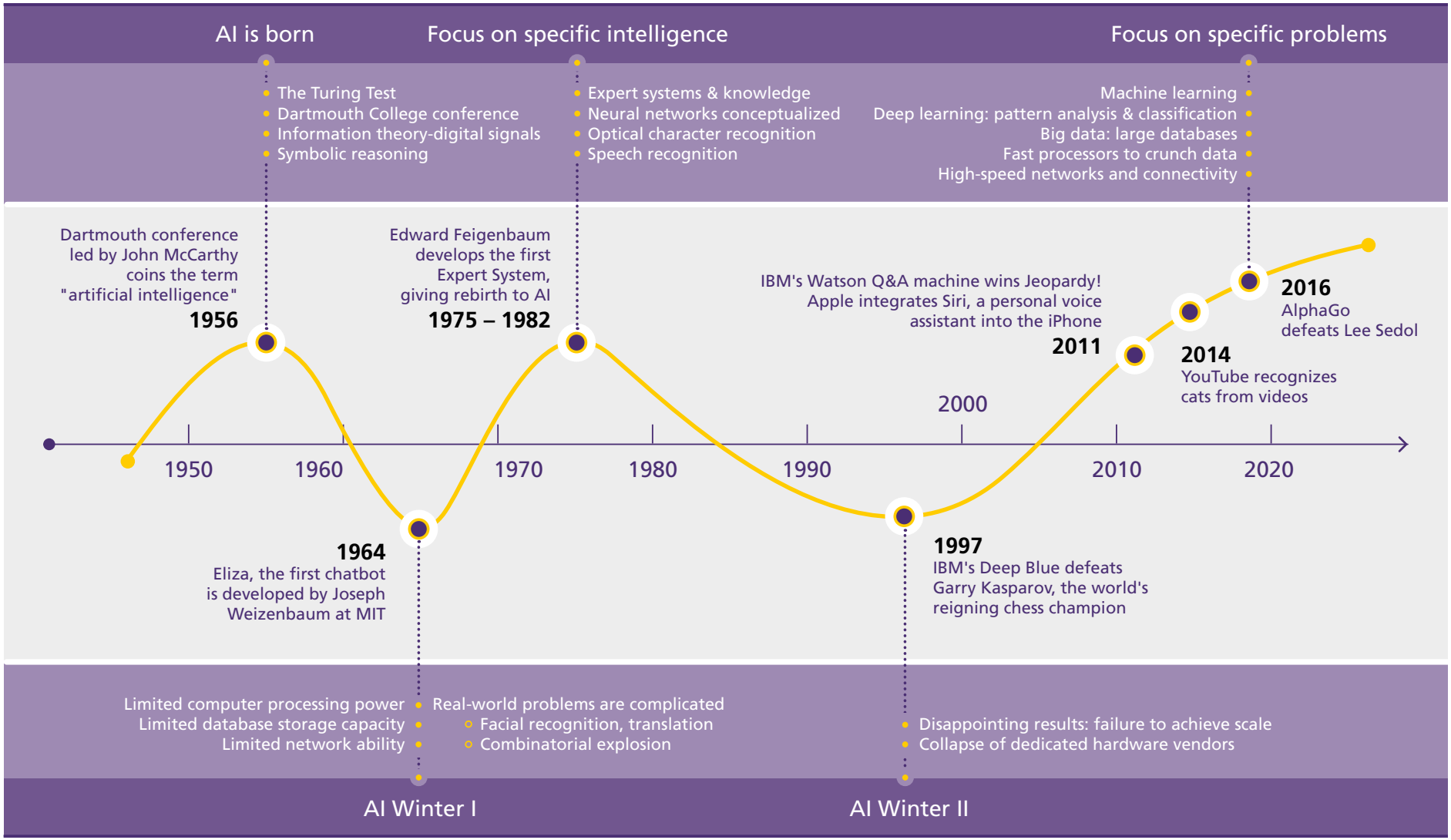
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agents interact with environments through sensors and actuators



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

The Rise of AI



Source: DHL (2018), Artificial Intelligence in Logistics, http://www.globalhha.com/doclib/data/upload/doc_con/5e50c53c5bf67.pdf/

Philosophy and Ethics of AI

Philosophy, Ethics, and Safety of AI

- The Limits of AI
- Can Machines Really Think?
- The Ethics of AI

Philosophy of AI

- Philosophers use the term
  - weak AI for the hypothesis that machines could possibly behave intelligently
  - strong AI for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds)

4 Approaches of AI

Thinking Humanly	**Thinking Rationally**
Acting Humanly	**Acting Rationally**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

4 Approaches of AI

2. **Thinking Humanly:** **The Cognitive Modeling Approach**	**3.** **Thinking Rationally:** **The "Laws of Thought" Approach**
1. **Acting Humanly:** **The Turing Test Approach (1950)**	**4.** **Acting Rationally:** **The Rational Agent Approach**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Can machines think?

- Alan Turing rejected the question “Can machines think?” and replaced it with a behavioral test.
  - Alan Turing anticipated many objections to the possibility of thinking machines.
- Concentrate on their systems’ performance on practical tasks
  - rather than the ability to imitate humans.
- Consciousness remains a mystery.

AI Acting Humanly: The Turing Test Approach (Alan Turing, 1950)

- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
  - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

The Ethics of AI

- Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones.

Principles of Robotics and AI

- Ensure safety
- Ensure fairness
- Respect privacy
- Promote collaboration
- Provide transparency
- Limit harmful uses of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Principles of Robotics and AI

- Establish accountability
- Uphold human rights and values
- Reflect diversity/inclusion
- Avoid concentration of power
- Acknowledge legal/policy implications
- Contemplate implications for employment

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Safety of AI

- AI is a powerful technology, and as such it poses potential dangers, through lethal autonomous weapons, security and privacy breaches, unintended side effects, unintentional errors, and malignant misuse.
- Those who work with AI technology have an ethical imperative to responsibly reduce those dangers.

Robot Ethics

Laws of Robotics (Isaac Asimov, 1942, 1950)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Fair, trustworthy, and transparent of AI

- AI systems must be able to demonstrate they are fair, trustworthy, and transparent.
- There are multiple aspects of fairness, and it is impossible to maximize all of them at once.
- So a first step is to decide what counts as fair.

Explainable AI (XAI)

- When an AI system turns you down for a loan, you deserve an explanation.
  - In Europe, the GDPR enforces this for you.
  - An AI system that can explain itself is called explainable AI (XAI).

Explainable AI (XAI)

- A good explanation properties of XAI
  - it should be understandable and convincing to the user
  - it should accurately reflect the reasoning of the system
  - it should be complete
  - it should be specific in that different users with different conditions or different outcomes should get different explanations

Automation

- Automation is already changing the way people work.
- As a society, we will have to deal with these changes.

The Future of AI

- AI Components
- AI Architectures

AI Components

- Sensors and actuators
- Representing the state of the world
- Selecting actions
- Deciding what we want
- Learning
- Resources
  - Shared data
  - Shared model

Learning

- Deep learning
- Data science
- Big data
- Transfer learning
- Apprenticeship learning
- Differentiable programming
- Weakly supervised learning
- Predictive learning

AI Architectures

- Which of the agent architectures should an agent use?
  - All of them!
- Real-time AI
- Anytime algorithm
- Decision-theoretic metareasoning
- Reflective architecture
- Agent = Architecture + Program
- Bounded optimality

Real-time AI

- As AI systems move into more complex domains, all problems will become real-time, because the agent will never have long enough to solve the decision problem exactly.

General AI

- Narrow tasks AI
  - DARPA Grand Challenge for autonomous cars
  - ImageNet object recognition competition
  - For each separate task, we build a separate AI system
  - A separate machine learning model trained from scratch with data collected specifically for this task.
- Human-level AI (HLAI)

AI Engineering

- Powerful tools and frameworks
  - TensorFlow, Keras, PyTorch, CAFFE, Scikit-Learn and SCIPY.
- Promising approaches
  - GANs
  - Deep reinforcement learning
  - Train properly in a new domain

The Future of AI

- Past: Build each new system from scratch
- Future: Start with a single huge system
  - For each new task, extract from it the parts that are relevant to the task.
- Transformer language models (e.g., BERT, GPT-2) with billions of parameters
  - An "outrageously large" ensemble neural network architecture that scales up to 68 billion parameters in one experiment.

The Future of AI

- AI has made great progress in its short history.
- We can see only a short distance ahead,
  but we can see that much remains to be done.
  (Alan Turing, 1950)
  [Computing Machinery and Intelligence]

Summary

- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

References

- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson.
- Aurélien Géron (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O'Reilly Media.
- Min-Yuh Day (2021), Python 101, https://tinyurl.com/aintpupython101