Artificial Intelligence 1 (Künstliche Intelligenz 1)

Frank Puppe

based on: **S. Russell, P. Norwig: Artificial Intelligence, A modern Approach, 4th edition, 2020**

chapter 1-11 (introduction, problem solving & search, logic, knowledge representation, ethics)

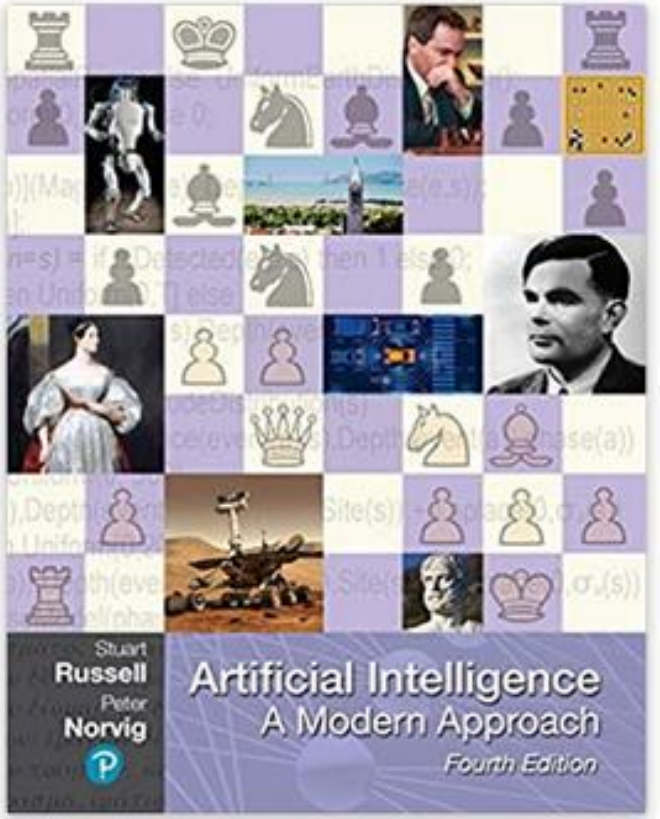
Continuation:

Artificial Intelligence 2 (Künstliche Intelligenz 2)

chapter 12-26 (uncertain knowledge and reasoning, machine learning, communication, vision)

Russell, Norwig: Artificial Intelligence (2020)

- **Part I Artificial Intelligence**
  - Introduction, Intelligent Agents
- **Part II Problem Solving**
  - Searching, Games, Constraint Solving
- **Part III Knowledge, Reasoning and Planning**
  - Logic, Knowledge Representation, Planning
- **Part IV Uncertain Knowledge and Reasoning**
  - Probabilistic Reasoning, Bayesian Networks, Utility, Multiagents
- **Part V Machine Learning**
  - Symbolic, Probabilistic, Deep, and Reinforcement Learning,
- **Part VI Communicating, Perceiving, and Acting**
  - Natural Language Processing, Computer Vision, Robotics
- **Part VII Conclusions**
  - Philosophy, Ethics, and Safety of AI, Future of AI



M. Ford: Rise of the Robots (2015)

1. The Automation Wave
2. Is This Time Different?
3. Information Technology: An Unprecedented Force for Disruption
4. White-Collar Jobs at Risk
5. Transforming Higher Education
6. The Health Care Challenge
7. Technologies and Industries of the Future
8. Consumers, Limits to Growth … and Crisis?
9. Super-Intelligence and the Singularity
10. Toward a New Economic Paradigm

Wikipedia https://en.wikipedia.org/wiki/Artificial_intelligence (06.09.2020)

1 History, 2 Basics, 3 Challenges of AI
- 3.1 Reasoning, problem solving
- 3.2 Knowledge representation
- 3.3 Planning
- 3.4 Learning
- 3.5 Natural language processing
- 3.6 Perception
- 3.7 Motion and manipulation
- 3.8 Social intelligence
- 3.9 General intelligence

4 Approaches
- 4.1 Cybernetics and brain simulation
- 4.2 Symbolic
- 4.3 Sub-symbolic
- 4.4 Statistical learning
- 4.5 Integrating the approaches

5 Applications https://en.wikipedia.org/wiki/Applications_of_artificial_intelligence#Tools_for_computer_science
- 5.3 Agriculture
- 5.4 Cybersecurity
- 5.5 Education
- 5.6 Finance
- 5.7 Government and Military
- 5.8 Health
- 5.9 Law
- 5.10 Service Sector
- 5.11 Media and e-commerce
- 5.12 Utilities (e.g. energy technologies)
- 5.13 Manufacturing
- 5.14 Transportation (Automotive, Aviation)

5a Tools https://en.wikipedia.org/wiki/Artificial_intelligence_tools_for_computer_science
- 5a.1 Search and optimization
- 5a.2 Logic
- 5a.3 Probabilistic methods for uncertain reasoning
- 5a.4 Classifiers & statistical learning meth.
- 5a.5 Artificial neural networks
- 5a.6 Evaluating progress

6 Philosophy and ethics
- 6.1 Limits of artificial general intelligence
- 6.2 Ethical machines
- 6.3 Machine consciousness, sentience & mind
- 6.4 Superintelligence

7 Impact
- 7.1 Risks of narrow AI
- 7.2 Risks of general AI

Definition of AI

- Any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals [Poole et al. 98, Russell & Norvig 03].
- A system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation [Kaplan & Haenlein 19].
- …

What is AI?

- Different approaches (dimensions):
  - Comparison to human performance vs. abstract definition like „rationality“
  - Concentration on thought processes vs. behaviour (acting)

- Four possible combinations:
  - **Acting humanly:** The Turing test approach
  - **Thinking humanly:** The cognitive modelling approach
  - **Thinking rationally:** The „laws of thought“ approach
  - **Acting rationally:** The rational agent approach
    - *A rational agent acts to achieve the best expected outcome*

The Turing test approach

- A computer passes the Turing test, if a human interrogator cannot tell, whether his or her communication partner is Human or Computer (originally in written dialogue).
- Necessary capabilities:
  - **Natural language processing**
  - **Knowledge representation**
  - **Automated reasoning**
  - **Machine learning**

- Extension: Total Turing test including interaction with objects and people in the real world
  - **Speech recognition and speech synthesis**
  - **Computer vision**
  - **Robotics**

Thinking humanly: The cognitive modelling approach

- How do humans think -> **Cognitive Science** with approaches
  - Introspection
  - Psycholocical experiments
  - (brain imaging)

- Often claims supported by intution
- But no gold standard
- Often inspiring

Thinking rationally: The „laws of thought“ approach

- Goes back to the Greek philosopher Aristotle
  - Syllogisms to yield correct conclusions from correct premises
  - e.g. From premises (1) „Socrates is a man“ and (2) „All men are mortal“ conclude „Sokrates is mortal“.
- Formalisation within the field of **logic**
- Extension by **theory of probability**

Acting rationally: The rational agent approach

- **A rational agent acts to achieve the best expected outcome**
  - dependent on **performance measure**, **environment**, **actors** and **sensors** of the agent
  - All skills from total Turing test necessary
  - Trade-off between time and optimality of outcome („limited rationality“)
  - Difficulty to define a good performance measure in complex environments
  - it is easy, to act intelligent in simple environments

- „Standard model“

Beneficial Machines

- Difficulty to describe an objective in the real world
  - E.g.: The objective of a self-driving car is to reach the destination safely.
    - But driving has inherent risks resulting from other drivers or technical failures.
    - Strict goal for safety would require staying at home
    - How to specify the necessary trade-off?
  - E.g.: Defensive or offensive driver style of self driving car?
- **Value Alignment Problem:** How to achieve agreement between our (implicit) values and the objective
  - Chess computer in the real world: It might disturb the opponent by noise or distracting movies.
- New formulation of objectives: The machine should pursue our objectives, but is necessarily uncertain as to what they are
  - Incentive to act cautiously, ask for permission, etc.

Foundations of AI

- **Philosophy:** Formal rules to draw valid conclusions
  - Mind-brain problem, origin of knowledge, how does knowledge lead to action?
- **Mathematics:** Formal rules to draw valid conclusions
  - What can be computed? Reasoning with uncertain information.
- **Economics:** Optimal decisions
  - Depending on others and expectations about the future
- **Neuroscience:** How do brains process information?
- **Psychology:** How do humans and animals think and act?
- **Computer engineering:** Building efficient hardware, software engineering
- **Control theory and cybernetics:** How can artefacts operate under their own control?
- **Linguistics:** How does language relate to thought?

History of AI

1. Surge 1956-74: Reasoning as Search, Language Translation, Microworlds

- 1.AI-Winter 1974-1980: exaggerated expectations (Lighthill-Report in England, Darpa in USA): Solutions to toy problems, unusable translations, Theoretical limitations of perceptrons.

2. Surge: 1980-1987: Expert systems, e.g. R1/XCON, LISP-Maschines, Fifth Generation Project (Japan), connektionism

- 2. AI-Winter 1987-1993: Market collapse for LISP-Maschines, expense knowledge acquisition, “Brittleness”

3. Surge: 1993-2001: Moore’s law, AI-Agents, theory (Logic, Bayesian nets, HMM, …), success in subsections (search engines, TTS, STT, machine translation, Robotics, …)

4. Surge: (current): Big Data, Deep Neural Networks,

State for the Art in AI 2018/2019

- **Publications:** 20 000 per year (2019) from 2 000 per year (2010)
  - most popular categories: Machine Learning, Computer Vision, Natural Language Proc.
- **Sentiment:** 70% neutral, rather positive 30%
- **Students:** 5 times (USA) to 16 times (internationally) more than 2010
- **Gender Distribution:** 80% male, 20% female (Professors, PhD-Students, Industry Hires)
- **Industry:** AI startups more than 800 in the US (20 times more than 2010)
- **Internationalization:** similar frequent publications from China, Europe, USA
- **Vision:** Error rates in object detection (LSVRC) improved from 28% (2010) to 2% (2017)
- **Speed:** Training time for image recognition task dropped by a factor of 100 in past 2 years
- **Language:** F1 in question answering In SQUAD increased from 60% (2015) to 95% (2019)
- **Human Benchmarks:** AI ahead: in chess, Go, poker, Joepardy!, object detection, speech recognition and translation in limited domains, skin, prostate, cancer detection etc.

Applications (1)

- **Robot Vehicles:** Autonomous Cars (Waymo achieved 10 million miles without serious accident and human driver taking control once every 6 000 miles); Quadcopters
- **Legged Locomotion:** BigDog, quadruped robot resembles animal movement
- **Autonomous Planning and Scheduling:** Great success in spacecraft and military
- **Machine Translation:** Online systems for over 100 languages in understandable quality, for closely related language (e.g. French and English) close to human level; Skype provides real time speech-to-speech translation in 10 languages.
- **Speech Recognition:** Alexa, Siri, Google …
- **Recommendation:** Very successful based on past experiences and others like you

Applications (2)

- **Game Playing** (champion in many games, strategic and multi-agent)
  - Phase 1: Building systems by learning from expert games and expert knowledge (AlphaGo)
  - Phase 2: Building systems by learning from self-play (AlphaZero)
- **Image Understanding:**
  - Phase 1: Recognition of the content of images (very good with enough training data)
  - Phase 2: Recognition of image captioning (e.g. „a person riding a motorcycle on a dirt road“ or „a group of young people playing frisbee“): not as good as humans
- **Medicine:**
  - diagnosis based on images often as good as experts
  - Current emphasis on human-machine partnerchips

Risks and Benefits

- **Benefits:**
  - Free humanity from simple repetive work
  - Dramatically increase the production of goods and services
  - Better medical therapies and preventions
  - Solutions for climate change and ressources shortage etc.
  - D. Hassabis, CEO of DeepMind: „First solve AI, then use AI to solve everything else“.
- **Risks:**
  - Lethal autonomous weapons; problem of scalability (small group with many weapons)
  - Mass surveillance (partially already in use in China)
  - Persuasion and manipulation (e.g. tailored recommenders in social media)
  - Biased decision making (e.g. by bias in learning data)
  - Impact on employment (also on information workers)
  - Safety-critical applications (e.g. driving cars, managing water supply of cities)
  - Cybersecurity (can be used for both defending and contributing to cyberattacks)

Regulation

- Voluntary self governance principles?
- Regulation by governments or international organizations?
  - Advisory boards

Superhuman AI

- Currently: Specialized (weak) AI
- Future: General (strong) AI
  - Can learn everything
  - Has no predefined objectives
  - Inbuild ethical rules? (e.g Asimov‘s laws)

- Impact of **strong AI**?
  - **Gorilla problem**: 7 million years ago, apes evolved in gorillas and humans (a.o.). Now, gorillas have no control about their future.
  - **King Midas problem**: Humans often do not know, what they really want (e.g. Midas wanted, that everything he touches turns into gold).