Australian National University

Artificial Intelligence COMP3620/6320

--

Introduction

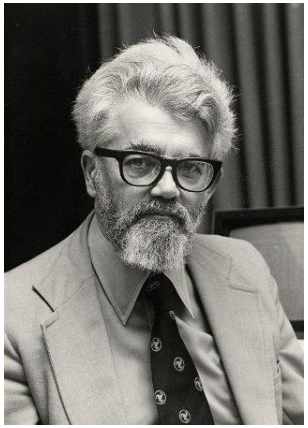
Sylvie Thiebaux

Topics

- What is AI?
- Foundational and Current Disciplines
- Brief History
- Ethics

What is AI?

Artificial Intelligence



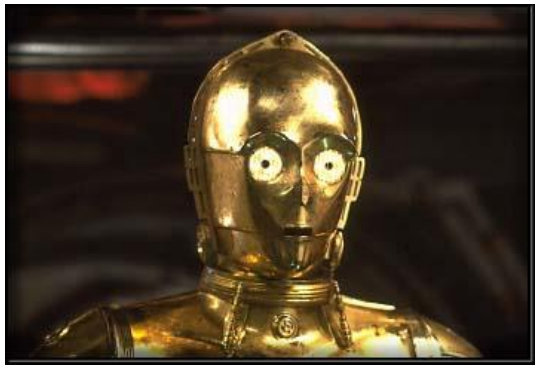
John McCarthy
1927-2011

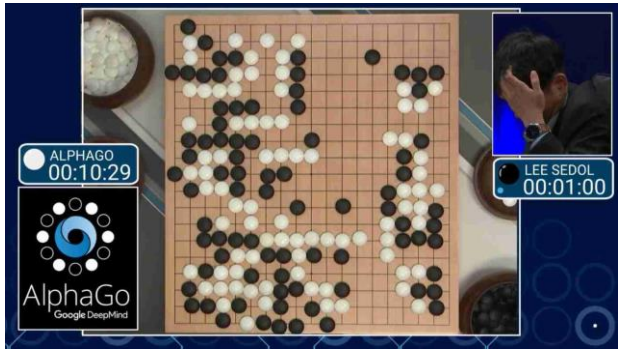
"The science and engineering of making intelligent machines"

- **Official birth:** Dartmouth College Meeting, 1956
- **Ambitious goals:**
  1. Understand "intelligence"
  2. Build "intelligent" machines

But … what counts as an intelligent machine?

What is an Intelligent Machine?





What is an Intelligent Machine?

think = reason action = behavior

- 4 different perspectives corresponding to two dimensions
  1. **thinking vs acting** (thought/reasoning vs actions/behavior)
  2. **human vs rational** (close to a human vs close to optimal)

Humans are not optimal sometimes.

Systems that think like humans	Systems that think rationally
Systems that act like humans	Systems that act rationally

- **Rationale:**
  1. The ability to think is what distinguishes humans from animals. But a machine could behave intelligently without thinking.
  2. Human exhibit intelligence. But why not aim at super-human, optimal behavior, especially for specific tasks?

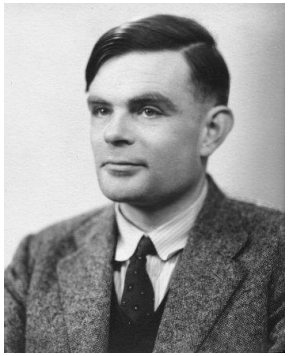
What is an Intelligent Machine?

Systems that think like humans	Systems that think rationally
Systems that act like humans	Systems that act rationally

- Making machines with minds in the full and literal sense [Haugeland 85]
- Making computers do things at which, for now, people are better [Kurzweil 90]
- Studying the computations that enable perception, reasoning, and actions [Winston 92]
- Concerned with the automation of intelligent behavior [Luger & Subblefield 93]

Whether machines can think is as relevant as whether submarines can swim [Dijkstra 84]

Acting Humanly: The Turing Test



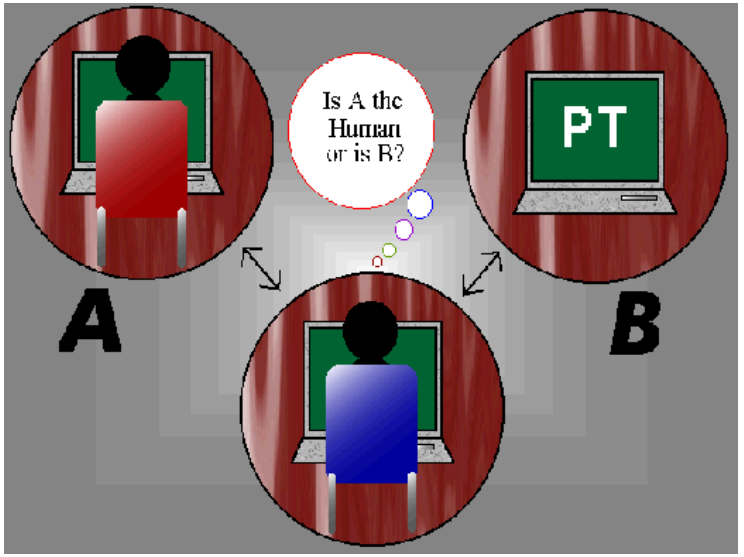
Alan Turing
1912-1954

- "Computing Machinery and Intelligence" [Turing 50]

- Operational test for intelligent behavior
- Imitation game: computer passes the test if interrogator cannot tell which is the machine (no physical presence)

- Focused on major AI components:
  - Natural language processing
  - Knowledge representation & reasoning
  - Learning

- Prediction: by 2000, 30% chance to fool a human in 5mn conversation



Acting Humanly: The Turing Test

- Eliza (Weizenbaum 66) rule-based program replicating a psychotherapist.
  Primitive natural language processing.

Patient: I feel sad.
Eliza: Do you often feel sad?
Patient: Not very often.
Eliza: Please, go on.

- Loebner Prize ($100,000) since 1991
  - A.L.I.C.E 3 bronze medals
  - Mitsuku 3 bronze medal in 2013/16/17 www.mitsuku.com

Sylvie: Does god exist?

Mitsuku: The bible is the most significant piece of evidence that God exists, but it is not a scientific proof.

Sylvie: Okay, but do you believe in god?

Mitsuku: Yeah, okay, I believe there is probably something out there but I doubt it is a man with a long beard sat on a throne.

Thinking Humanly: Modelling Cognition

- Objective: develop scientific theories of the activities of the brain

- Two approaches:

1. Cognitive Science (top down):
   uses computer models and experimental psychology techniques to predict and test behavior of human subjects

2. Cognitive Neuroscience (bottom up):
   uses computer imaging & other neurological data to observe the brain in action
   - Project to simulate the brain www.humanbrainproject.eu
   - Related to the AI field of neural networks (see deep learning)

- These days, both disciplines are distinct from AI

Thinking Rationally: Laws of Thought

- Objective: formalise and mechanise valid reasoning
- Direct line through maths and philosophy to modern AI

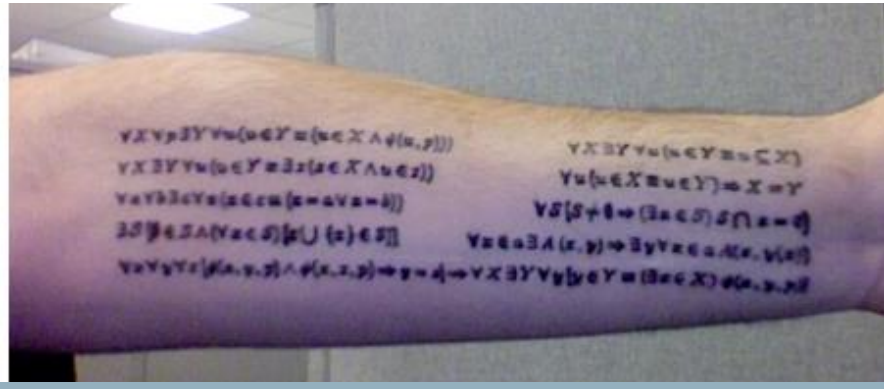
- Logic: notation and rules to derive valid conclusions
  - Aristotle's syllogism
  - Mathematical development of classical logic
    - Propositional & first-order logic (Boole, Frege, 1850s)
    - Most of mathematics can be derived from axioms of set theory
  - Non-classical logic to formalise commonsense reasoning
    - Default logic (by default, birds fly)

Tweety is a bird
Birds fly

Tweety flies

P(a)
∀x P(x) → Q(x)

Q(a)



Thinking Rationally: Laws of Thought

- **Limit 1: Undecidability**
  - Goedl's Theorem: every axiomatisable consistent theory extending arithmetic has formulas that are true but not provable within the theory.
- **Limit 2: Complexity**
  - Non-trivial to formalise a real-world problem in logic
  - Most problems are NP-complete or harder
- **Limit 3: Scope**
  - Not all intelligent behavior requires reasoning (much doesn't)
- **Limit 4: Purpose**
  - Reasoning to prove what? Notion of "goal" is missing

Acting Rationally: Rational Agents

- An agent is an entity that perceives and acts in its environment (driverless car, electronic trading system, energy management system)

- Rationality is about doing the right thing:
- Decision which achieves the best (expected) outcome, given the information available and time available (limited rationality)

- This course (and much of today's AI) is about designing rational agents: for any given class of environment and task, we seek the agent with the best performance.

Artificial Intelligence

"The science and engineering of making intelligent machines"

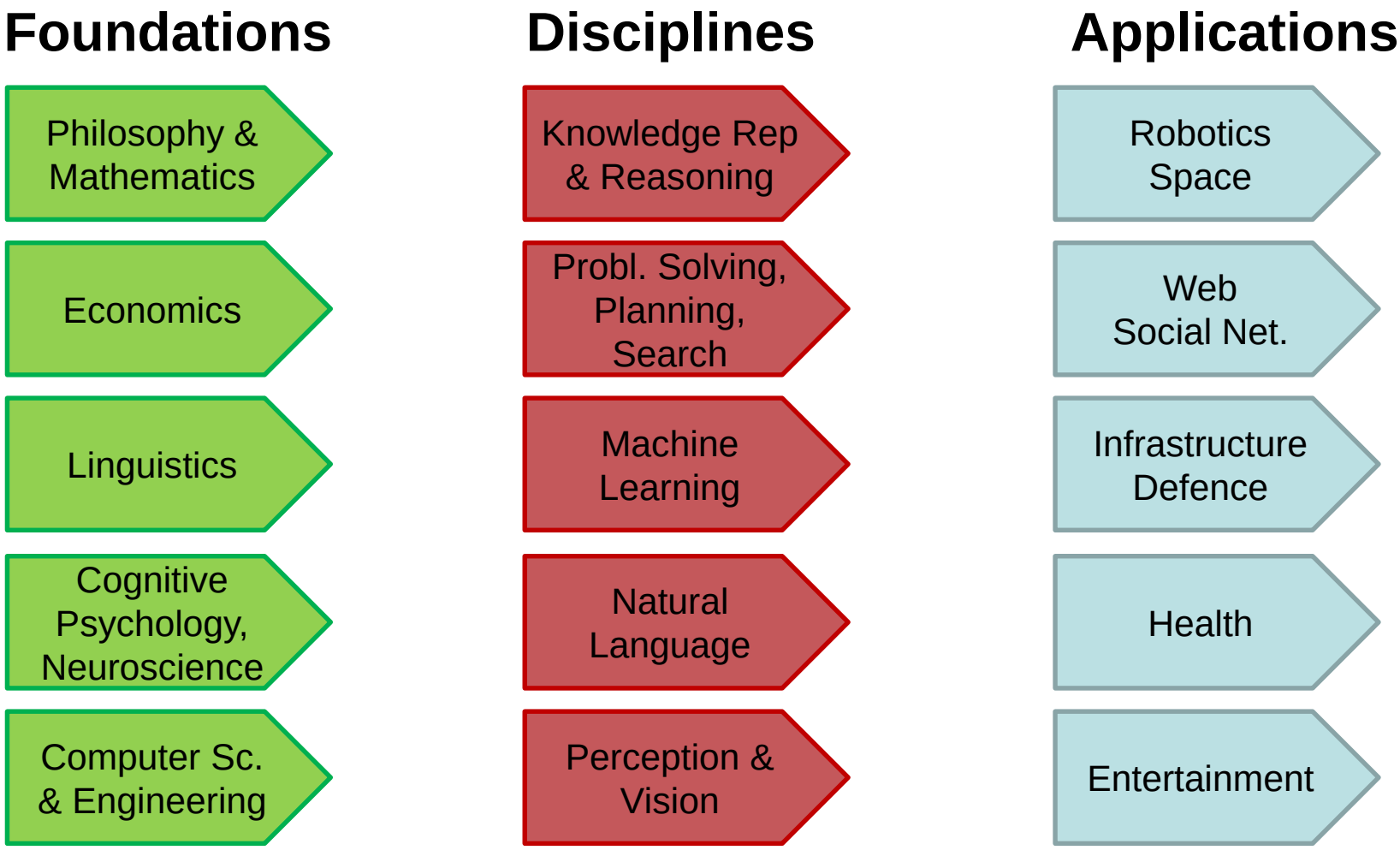
- **Ambitious goals:**

1. Understand "intelligence"
   - ➢ Accurate models of cognition are now the focus of cognitive science, neuroscience and psychology

2. Build "intelligent" machines
   - ➢ Focus on developing methods that match or exceed human performance in certain domains, possibly by different means.

Foundational and Current Disciplines

The Artificial Intelligence Field

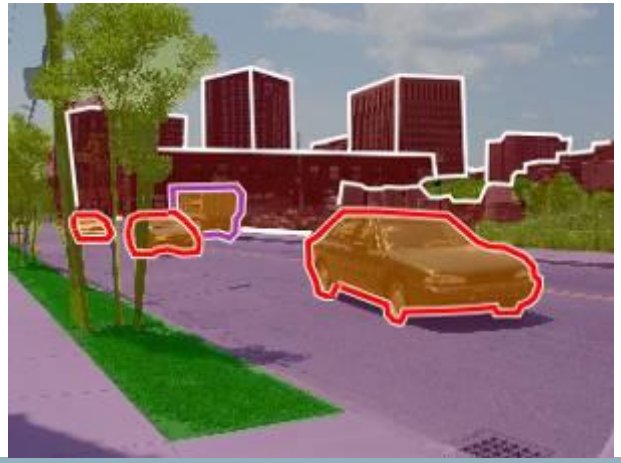


Foundational Disciplines

- **Philosophy:** logic, reasoning methods, foundations of learning, language, and rationality
- **Mathematics:** proofs, decidability, complexity, probability
- **Economics:** theory of rational decisions, game theory
- **Computer Sc. & Engineering:** algorithms, efficient computer design, control theory concepts (e.g. stability)
- **Cognitive Science:** behaviorism, adaptation, perception, experimental methods
- **Neuroscience:** information processing by the brain
- **Linguistics:** language representation, language & thought

Current Disciplines

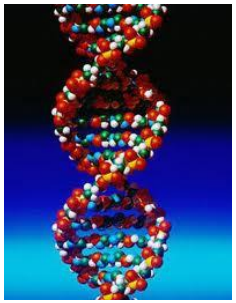
- **Knowledge representation & reasoning:** how to efficiently represent information and use this to answer questions and draw conclusions
- **Problem solving, planning, and search:** how to constructively solve problems and make decisions.
- **Machine learning:** inference from data to extrapolate patterns and adapt to new situations.
- **Natural language processing:** verbal communication with humans.
- **Computer vision:** processing and making sense of visual information about the environment.



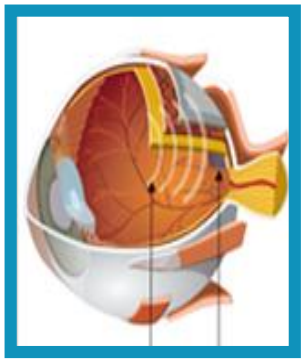
Applications

Financial markets

Space

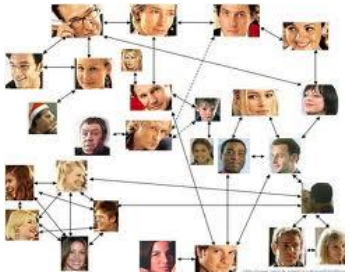


Health

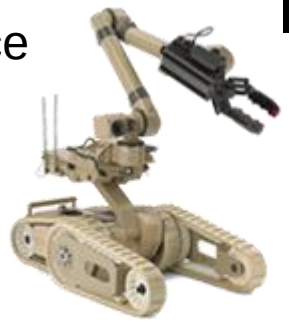


Defence

Energy



Web mining and applications



Transport

Brief History

Brief History

1950: Turing test
1950s: Early programs including checkers, theorist, neural nets
1956: Dartmouth meeting, “Artificial Intelligence” adopted
1965: Robison’s complete algorithm for logical reasoning
1966-74: AI discovers complexity, neural nets research disappears
1969-79: Early knowledge-based systems
1980-88: Expert systems industry booms
1988-93: Expert systems industry “busts”, AI Winter
1988-00: Greater technical depth, Resurgence of probabilities,
1985-95: Neural nets return, lead to, and replaced by modern SML
2003- : Human-level AI back on the agenda
2010-: Deep learning: neural nets research is favour again
2013-: Ethical issues make the headlines

Birth
Optimism
Realism

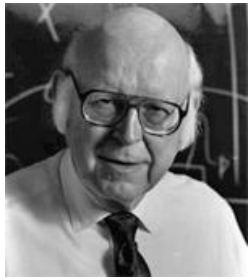
Expert Syst.

Winter
Foundations
NN returns

Data, multicore
NN again!
Maturity?

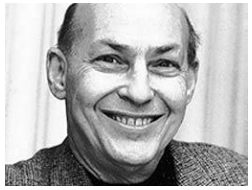
AI Achievements – Predictions

- **1958**: "within ten years a digital computer will be the world's chess champion" [Allen Newell]
- **1965:** "machines will be capable, within twenty years, of doing any work a man can do." [Herb Simon]
- **1970:** "In from three to eight years we will have a machine with the general intelligence of an average human being." [Marvin Minsky]



Allen Newell
1927-1992

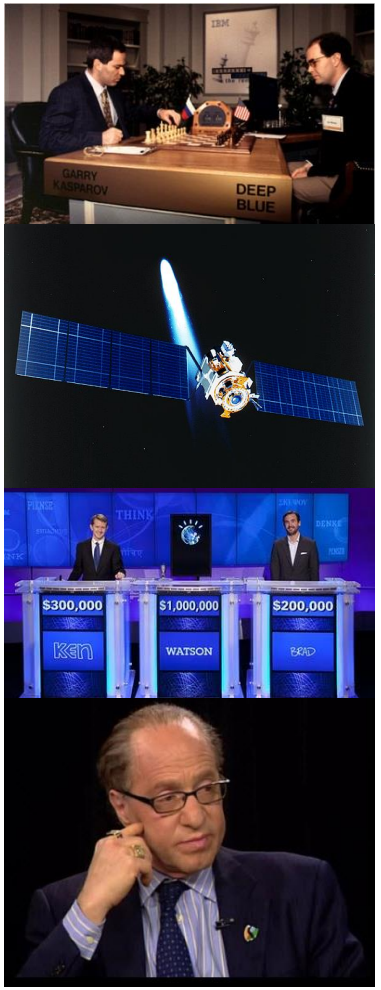
Herb Simon
1916-2001



Marvin Minsky
1927-2016

AI Achievements – The Reality

- **1991:** Proverb solves crosswords better than human
- **1991:** AI solves Gulf-war logistics planning problems
- **1997**: IBM Deep Blue beats chess champion Kasparov
- **1999:** AI agent controls NASA deep space 1 probe
- **2001:** autonomous military drones unveiled
- **2005:** Driverless vehicles complete the 212km DARPA Grand Challenge through the Mojave desert
- **2007:** Checkers game completely solved.
- **2009:** Google autonomous car drives in traffic
- **2011:** IBM Watson wins Jeopardy!
- **2016:** Google alphago beats go champion Lee Sedol
- **Today:** AI is everywhere, injects billions into economy
- **2060+:** 50% probability of human-level intelligence



AI Achievements – New Predictions

- **2030**: “an AI system with an ongoing existence at the level of a mouse" [Rodney Brooks]
- **2050:** “Germany will loose to a robot soccer team.” [Toby Walsh]
- **Not in his lifetime:** “a robot that has any real idea about his own existence, or the existence of humans in a way a 6 years old child would” [Rodney Brooks]

Rod Brooks
1954-

Toby Walsh
1964-

AI Ethics

AI Ethics and Risks

- People might lose their jobs
+ AI creates wealth and does dangerous and boring jobs for us

- Accountability loss: who is responsible, AI, owner, creator?
+ Similar issues elsewhere (medicine, software, plane crash)

- AI reproducing our negative biases and attitudes (e.g. racism)
+ AI should share our *positive* values

- Use of AI as weapon (e.g. drones)
+ Can also save lives? Every beneficial invention can be misused

AI Ethics and Risks

- AI Success might end of the human era
  - Kurtzweil, Musk, Hawking!
  - Once machine surpasses human intelligence it can design smarter machines.
  - Intelligence explosion and **singularity** at which human era ends

- Many counter arguments
  - nothing special about human intelligence
  - limits to intelligence
  - computational complexity
  - “intelligence to do a task” ≠ “ability to improve intelligence to do a task”

Robotics Laws

The Three Laws of Robotics [Azimov 1942]

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 the orders given 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

0. A robot may not injure humanity, or, through inaction, allow humanity to come to harm

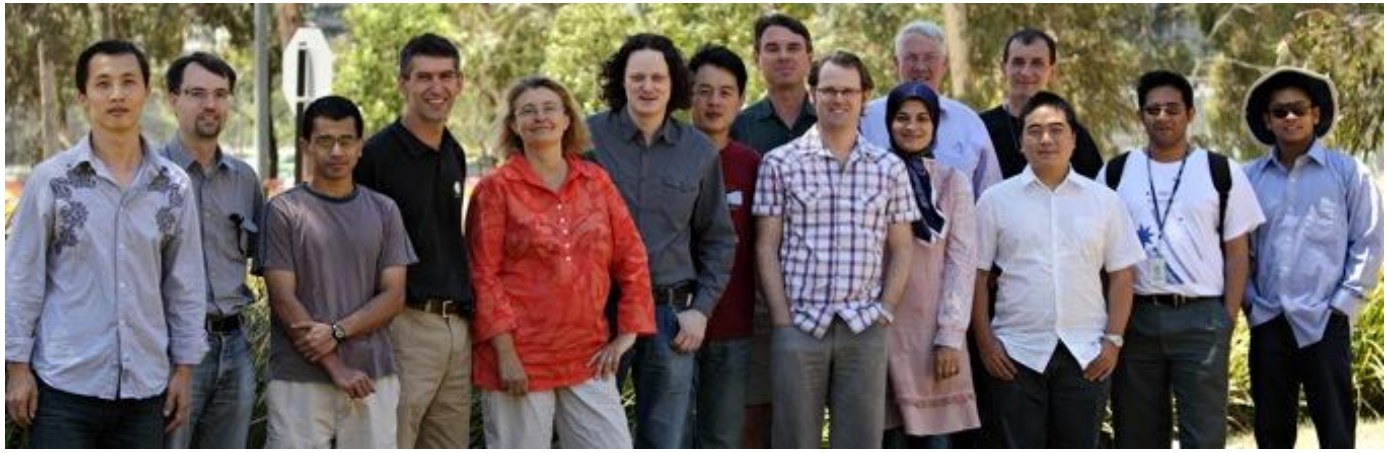
UK Principles of Robotics [EPSRC 2011]

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

AI@ANU

AI Group

- 70 Staff and PhD students, working on:
  - Theoretical foundations of AI
  - Core AI disciplines (learning, vision, planning, reasoning, nat. lang …)
  - Building high-performance AI tools
  - Addressing important AI applications (health, energy, social networks, …)
- ERA 5 “outstanding, well, above world standards”



Relevant Courses

ANU courses

- COMP3620 Artificial Intelligence
- COMP4620 Advanced Topics in AI
- COMP4670 Introduction to SML
- COMP4680 Advanced Topics in SML
- COMP4660 Bio-Inspired Computing
- COMP4650 Document Analysis
- ENG4528 Computer Vision

Massive on-line courses

- http://www.coursera.org

Background courses

- COMP2620 Logic
- COMP2610 Information Theory
- COMP3600 Algorithms
- COMP3630 Theory of Computation

Project courses

- COMP3006, 3130, 3710, 3740 …

Research opportunities

- ANU Summer scholarships
- Honours projects

Contacts: first.last@anu.edu.au

Computer Vision

- Nick Barnes (Bionic Eye)
- Stephen Gould (object recognition)
- Richard Hartley (theory)
- Hongdong Li (object recognition)
- Rob Mahony, Lars Petersson (robotics)
- Antonio Robles-Kelly (hyperspectral)

Planning, Scheduling, Diagnosis

- Alban Grastien (diagnosis)
- Charles Gretton (planning, routing)
- Patrik Haslum (planning, search, cybersec.)
- Phil Kilby (routing, scheduling, transport)
- Sylvie Thiebaux (planning & sched., energy)

Natural Language

- Hanna Suominen, Gabriella Ferraro

Learning

- Marcus Hutter (universal AI, RL)
- Mark Reid (theory, game theory)
- Bob Williamson (theory)
- Lexing Xie (social nets, deep learning)

Optimisation

- Sid Chau (opt. algorithms, energy)
- Guanglei Wang(MIP, MINLP,energy)
- Paul Scott (distributed opt., energy)

Knowledge Representation

- Jochen Renz (spatial/temporal)

Summary

- How to think or how to behave? Being like humans or being rational?
- This course about acting rationally

- AI related to many fields including philosophy, mathematics, economics, neuroscience, psychology, computer sci. and control theory
- 50+ years of progress along many different paradigms: logic, expert systems, neural nets, learning, probabilities
- Increasingly scientific: focus on experimental comparisons and theoretical foundations

- AI is a high-risk high gain area with major ethical implications
- Great opportunities to specialise in AI at the ANU