RUSSELL & NORVIG, CHAPTERS 1–2: INTRODUCTION TO AI

DIT411/TIN175, Artificial Intelligence

Peter Ljunglöf

16 January, 2018

TABLE OF CONTENTS

What is AI? (R&N 1.1–1.2)

- What is intelligence?
- Strong and Weak AI

A brief history of AI (R&N 1.3)

- Notable AI moments, 1940–2018
- “The three waves of AI”

Interlude: What is this course, anyway?

- People, contents and deadlines

Agents (R&N chapter 2)

- Rationality
- Enviroment types

Philosophy of AI

- Is AI possible?
- Turing’s objections to AI

WHAT IS AI? (R&N 1.1–1.2)

WHAT IS INTELLIGENCE?

STRONG AND WEAK AI

WHAT IS INTELLIGENCE?

> *”It is not my aim to surprise or shock you – but the simplest way I can summarize is to say that there are now in the world machines that can think, that learn, and that create.*
>
> *Moreover, their ability to do these things is going to increase rapidly until — in a visible future — the range of problems they can handle will be coextensive with the range to which human mind has been applied.”*
>
> *by Herbert A Simon **(1957)***

STRONG AND WEAK AI

Weak AI — acting intelligently

- the belief that machines can be made to act as if they are intelligent

Strong AI — being intelligent

- the belief that those machines are actually thinking

Most AI researchers don't care

- *"the question of whether machines can think…*
  …is about as relevant as whether submarines can swim."
  (Edsger W Dijkstra, 1984)

WEAK AI

Weak AI is a category that is flexible

- as soon as we understand how an AI-program works,
  it appears less “intelligent”.

And as soon as AI is successful, it becomes an own research area!

- e.g., search algorithms, natural language processing,
  optimization, theorem proving, machine learning etc.

And AI is left with the remaining hard-to-solve problems!

WHAT IS AN AI SYSTEM?

Do we want a system that…

- thinks like a human?
    - cognitive neuroscience / cognitive modelling
    - AGI = artificial general intelligence
- acts like a human?
    - the Turing test
- thinks rationally?
    - “laws of thought”
    - from Aristotle’s syllogism to modern day theorem provers
- acts rationally?
    - “rational agents”
    - maximise goal achievement, given available information

A BRIEF HISTORY OF AI (R&N 1.3)

NOTABLE AI MOMENTS, 1940–2018

“THE THREE WAVES OF AI”

NOTABLE AI MOMENTS (1940–1970)

1943	McCulloch & Pitts: Boolean circuit model of brain
1950	Alan Turing’s “Computing Machinery and Intelligence”
1951	Marvin Minsky develops a neural network machine
1950s	Early AI programs: e.g., Samuel’s checkers program, Gelernter’s Geometry Engine, Newell & Simon’s Logic Theorist and General Problem Solver
1956	Dartmouth meeting: “Artificial Intelligence” adopted
1965	Robinson’s complete algorithm for logical reasoning
1966	Joseph Weizenbaum creates Eliza
1969	Minsky & Papert show limitations of the perceptron Neural network research almost disappears

NOTABLE AI MOMENTS (1970–2000)

1971	Terry Winograd’s Shrdlu dialogue system
1972	Alain Colmerauer invents Prolog programming language
1976	MYCIN, an expert system for disease diagnosis
1980s	Era of expert systems
1990s	Neural networks, probability theory, AI agents
1993	RoboCup initiative to build soccer-playing robots
1997	IBM Deep Blue beats the World Chess Champion

NOTABLE AI MOMENTS (2000–2018)

2003	Very large datasets: genomic sequences
2007	Very large datasets: WAC (web as corpus)
2011	IBM Watson wins Jeopardy
2012	US state of Nevada permits driverless cars
2010s	Deep learning takes over: recommendation systems, image analysis, board games, machine translation, pattern recognition
2017	Google AlphaGo beats the world’s best Go player, Ke Jie AlphaZero learns boardgames by itself and beats the best programs
2018	Volvo will test-drive 100 driverless cars in Gothenburg

“THE THREE WAVES OF AI”

“To summarize, we see at DARPA that there have been three waves of AI,

- the first of which was handcrafted knowledge. It’s still hot, it’s still relevant, it’s still important.
- The second wave, which is now very much in the mainstream for things like face recognition, is about statistical learning where we build systems that get trained on data.
- But those two waves by themselves are not going to be sufficient. We see the need to bring them together. And so we’re seeing the advent of a third wave of AI technology built around the concept of contextual adaption.”

(by John Launchbury, March 2017: Youtube video, written article)

In this course, we focus on first wave AI!

INTERLUDE: WHAT IS THIS COURSE, ANYWAY?

PEOPLE, CONTENTS AND DEADLINES

PEOPLE AND LITERATURE

Course website	http://chalmersgu-ai-course.github.io/
Teachers	Peter Ljunglöf, Divya Grover, Herbert Lange, Inari Listenmaa, Claes Strannegård
Student representatives (randomly assigned)	Rasmus Andersson (MPIDE), Joel Sanderöd Roxell (MPSOF), Naichen Wang (MPALG), Widjaja Damarputra (MPALG), Lisanu Tebikew Yallew (MPCSN), Philip Tibom (GU), Johanna Torbjörnsson (GU)
Course book	Russell & Norvig (2002/10/14) Read it online at Chalmers library: http://goo.gl/6EMRZr

Note for GU students: *Don’t forget to register, today!*

COURSE CONTENTS

This is what you (hopefully) will learn during this course:

- Introduction to AI history, philosophy and ethics.
- Basic algorithms for searching and solving AI problems:
    - heuristic search,
    - local search,
    - nondeterministic search,
    - games and adversarial search,
    - constraint satisfaction problems.
- Group collaboration:
    - write an essay,
    - complete a programming project.

WHAT IS *NOT* IN THIS COURSE?

This course is an introduction to AI, giving a broad overview of the area and some basic algorithms.

- We do not have the time to dig into the most recent algorithms and techniques that are so hyped in current media.
- Therefore, you will not learn how these things work:
    - machine learning,
    - deep neural networks,
    - self-driving cars,
    - beating the world champion in Go,
    - etc.

DEADLINES FOR COURSE MOMENTS

Group work:

- Form a group (19 Jan)

Group work: Shrdlite programming project

- Submissions: A* search (31 Jan) + interpreter (7 Feb) + planner (28 Feb)
- Complete the final project (13 Mar)

Group work: Write an essay

- Write a 6-page essay about AI (27 Feb)
- (Individually) review one essay each (6 Mar)
- Revise your essay according to the reviews you got (16 Mar)

Written and oral examination

- *Peer-corrected* exam (13 Feb) + normal re-exams (5 Jun, 24 Aug)
- Oral review of the project (14–16 Mar)
- Individual self- and peer evaluation (16 Mar)

RECURRING COURSE MOMENTS

Lectures

- Tuesday and Friday, 10:00–11:45, during weeks 3–6

Obligatory group supervision

- Wednesdays and Thursdays (mostly) during weeks 4–10
- Supervision is compulsory for all group members!

Drop-in supervision

- Wednesday and Thursday week 3 (this week!)
- Mondays and Tuesdays (mostly) during weeks 4–10

Practice sessions

- Tuesday and Friday, 8:00–9:45, weeks 5–6

GRADING

All 3 subcourses are graded (U/345 resp. U/G/VG), and the final grade is:

- **GU**: To get final grade VG, you need a VG grade on at least two subcourses.
- **Chalmers**: The final grade is the average of the subcourse grades, weighted by the size of the subcourse, rounded like this:

Weighted average	Final grade
< 3.65	3
3.65–4.50	4
> 4.50	5

Note that the final grades on all subcourses are individual!

- This means that you can get a higher or lower grade than what your other group members will get, depending on your personal contributions to the group work.

THE LECTURES

There are 8 lectures:

Tue 16 Jan	Introduction
Fri 19 Jan	Search I, Classic and heuristic search
Tue 23 Jan	Search II, Heuristic search
Fri 26 Jan	NLP, Natural language interpretation
Tue 30 Jan	CSP I, Backtracking, consistency and heuristics
Fri 2 Feb	Search III, Non-classical and adversarial search
Tue 6 Feb	CSP II, Local search and problem structure
Fri 9 Feb	Repetition

Followed by the written exam, **Tue 13 Feb**

THE WRITTEN EXAMINATION

The exam is 13th February (in the middle of the course)

- *Why?* So that you can focus on Shrdlite and the essay in the end

The exam is peer-corrected

- *Why?* It's not only an exam, it's also a learning experience.
- *How?* First you write your exam. We collect all theses, shuffle and hand them out again, so that you will get someone else's exam to correct. We go through the answers on the blackboard and you correct the exam in front of you. Finally, we check all corrections.
- And don't worry – everything will be anonymous!

THE ESSAY

Your project group will write a 6-page essay about the historical, ethical and/or philosophical aspects of an AI topic.

After submitting your essay, each one of you will get another essay to review.
- the reviewing should be done individually!

Your group will get 4–5 reviews on your essay.
You update it and submit a final version.

Claes Strannegård is responsible for the essay.
He will organise supervision sessions for all of you, regarding the essay.

SHRDLITE, THE PROGRAMMING PROJECT

Your group will implement a dialogue system for controlling a robot that lives in a virtual block world and whose purpose in life is to move around objects of different forms, colors and sizes.

You will program in TypeScript

- *Why?* It’s a type-safe version of Javascript (runs in the browser), and it’s a new language for almost all of you!

Every group will get a personal supervisor, which you meet once every week.

There are three intermediate labs, which you submit by showing them to your supervisor.

Note: the Shrdlite webpage is quite long, and not everything makes sense when you start the project. Make sure to visit the webpage regularly when you are developing your project — there is a lot of important information there.

CHANGES SINCE LAST YEAR

Changes to the course structure and grading

- The previous big project subcourse is now divided into two subcourses.
- The grading calculation has been simplified.
- The written exam is graded (i.e., the final grade depends on all subcourses).

Changes to the theoretical content

- I have dropped some advanced content.

Changes to the Shrdlite project

- The template code is improved, and the interpreter has a better skeleton.
- There is a third intermediate submission, the planner.

Changes to the essay

- The main essay work will be in the week directly after the written examination.
- The essay reviews are now individual (i.e., every essay will get more reviews).

LET’S HAVE A LOOK AT THE WEB PAGES!

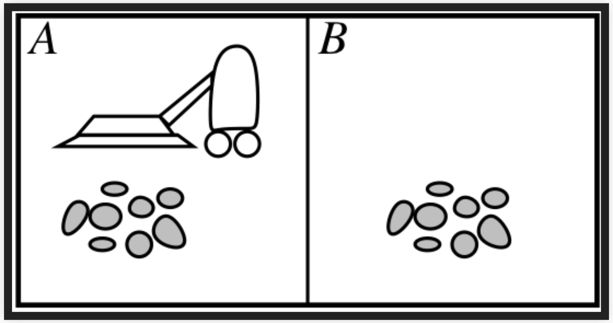
http://chalmersgu-ai-course.github.io/

AGENTS (R&N CHAPTER 2)

RATIONALITY

ENVIROMENT TYPES

EXAMPLE: A VACUUM-CLEANER AGENT



Percepts: location and contents, e.g. $(A, Dirty)$
Actions: *Left*, *Right*, *Suck*, *NoOp*

A simple agent function is:

- If the current square is dirty, then suck; otherwise, move to the other square.

How do we know if this is a good agent function?

- What is the best function? — Is there one?
- Who decides this?

RATIONALITY

A *performance measure* is an objective criterion for success:

- one point per square cleaned up in time T?
- one point per clean square per time step, minus one per move?
- penalize for $> k$ dirty squares?

A *rational agent* chooses any action that

- maximizes the expected value of the performance measure
- given the history of percepts, and builtin knowledge

Rationality and success

- Rational ≠ omniscient — percepts may not supply all relevant information
- Rational ≠ clairvoyant — action outcomes may not be as expected
- Hence, rational ≠ successful

PEAS

To design a rational agent, we must specify the *task environment*, which consists of the following four things:

***P*erformance measure**
: the agent’s criterion for success

***E*nvironment**
: the outside world interacting with the agent

***A*ctuators**
: how the agent controls its actions

***S*ensors**
: how the agent percieves the outside world

EXAMPLE PEAS: AUTONOMOUS CAR

The task environment for an autonomous car:

***P*erformance measure**
getting to the right place, following traffic laws,
minimising fuel consumption/time, maximising safety, …

***E*nvironment**
roads, other traffic, pedestrians, road signs, passengers, …

***A*ctuators**
steering, accelerator, brake, signals, loudspeaker, …

***S*ensors**
cameras, sonar, speedometer, GPS, odometer, microphone, …

ENVIROMENT TYPES: DIMENSIONS OF COMPLEXITY

Dimension	Possible values
Observable?	*full vs. partial*
Deterministic?	*deterministic vs. stochastic*
Episodic?	*episodic vs. sequential*
Static?	*static vs. dynamic (semidynamic)*
Discrete?	*discrete vs. continuous*
Number of agents	*single vs. multiple (competetive/cooperative)*

The environment type largely determines the agent design

ENVIRONMENT TYPES, EXAMPLES

	Chess (w. clock)	Poker	Driving	Image recognition
Observable?	*fully*	*partially*	*partially*	*fully*
Deterministic?	*determ.*	*stochastic*	*stochastic*	*determ.*
Episodic?	*sequential*	*sequential*	*sequential*	*episodic*
Static?	*semidyn.*	*static*	*dynamic*	*static*
Discrete?	*discrete*	*discrete*	*continuous*	*disc./cont.*
N:o agents	*multiple* *(compet.)*	*multiple* *(compet.)*	*multiple* *(cooper.)*	*single*

The real world is (of course):

partially observable, stochastic, sequential, dynamic, continuous, multi-agent

DEFINING A SOLUTION

Given an informal description of a problem, what is a solution?

- Typically, much is left unspecified, but the unspecified parts cannot be filled in arbitrarily.
- Much work in AI is motivated by *common-sense reasoning*. The computer needs to make common-sense conclusions about the unstated assumptions.

QUALITY OF SOLUTIONS

Does it matter if the answer is wrong or answers are missing?

Classes of solutions:

- An *optimal solution* is a best solution according to some measure of solution quality.
- A *satisficing solution* is one that is good enough, according to some description of which solutions are adequate.
- An *approximately optimal solution* is one whose measure of quality is close to the best theoretically possible.
- A *probable solution* is one that is likely to be a solution.

TYPES OF AGENTS

Simple reflex agent	selects actions based on *current percept* — ignores history
Model-based reflex agent	maintains an *internal state* that depends on the percept history
Goal-based agent	has a *goal* that describes situations that are desirable
Utility-based agent	has a *utility function* that measures the performance
Learning agent	any of the above agents can be a learning agent — learning can be *online* or *offline*

PHILOSOPHY OF AI

IS AI POSSIBLE?

TURING’S OBJECTIONS TO AI

IS AI POSSIBLE?

There are different opinions…

- …some are slightly positive:
    - “every […] feature of intelligence can be so precisely described that a machine can be made to simulate it” (McCarthy et al, 1955)
- …and some lean towards the negative:
    - “AI […] stands not even a ghost of a chance of producing durable results” (Sayre, 1993)

It’s all in the definitions:

- what do we mean by “thinking” and “intelligence”?

“COMPUTING MACHINERY AND INTELLIGENCE”

The most important paper in AI, of all times:

- (and I’m not the only one who thinks that…)
- “Computing Machinery and Intelligence” (Turing, 1950)
    - introduced the “imitation game” (Turing test)
    - discussed objections against intelligent machines, including almost every objection that has been raised since then
    - it’s also easy to read… so you really have to read it!

TURING’S (DISCUSSION OF) OBJECTIONS TO AI [1–3]

(1) The Theological Objection

- “Thinking is a function of man’s immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.”

(2) The “Heads in the Sand” Objection

- “The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so.”

(3) The Mathematical Objection

- Based on Gödel’s incompleteness theorem.

TURING’S (DISCUSSION OF) OBJECTIONS TO AI [4–5]

(4) The Argument from Consciousness

- “No mechanism could feel […] pleasure at its successes, grief when its valves fuse, […], be angry or depressed when it cannot get what it wants.”

(5) Arguments from Various Disabilities

- “you can make machines do all the things you have mentioned but you will never be able to make one to do X.”
- where X can… “be kind, resourceful, beautiful, friendly, […], have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, […], use words properly, be the subject of its own thought, […], do something really new.”

TURING’S (DISCUSSION OF) OBJECTIONS TO AI [6–8]

(6) Lady Lovelace’s Objection

- “The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform.”

(7) Argument from Continuity in the Nervous System

- “one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system.”

(8) The Argument from Informality of Behaviour

- “if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.”

THE FINAL OBJECTION [9]

(9) The Argument from Extrasensory Perception

- “Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as ‘What suit does the card in my right hand belong to?’ The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps gets 104 right, so the interrogator makes the right identification.”
- (this was the strongest argument according to Turing…
  “the statistical evidence […] is overwhelming”)

STRONG AI: BRAIN REPLACEMENT

The brain replacement experiment

- by Searle (1980) and Moravec (1988)
- suppose we gradually replace each neuron in your head with an electronic copy…
  - …what will happen to your mind, your consciousness?
  - Searle argues that you will gradually feel dislocated from your body
  - Moravec argues you won’t notice anything

STRONG AI: THE CHINESE ROOM

The Chinese room experiment (Searle, 1980)

- an English-speaking person takes input and generates answers in Chinese
    - he/she has a rule book, and stacks of paper
    - the person gets input, follows the rules and produces output
- i.e., the person is the CPU, the rule book is the program and the papers is the storage device

Does the system understand Chinese?

THE TECHNOLOGICAL SINGULARITY

Will AI lead to superintelligence?

- “…ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue” (von Neumann, mid-1950s)
- “We will successfully reverse-engineer the human brain by the mid-2020s. By the end of that decade, computers will be capable of human-level intelligence.” (Kurzweil, 2011)
- “There is not the slightest reason to believe in a coming singularity.” (Pinker, 2008)

ETHICAL ISSUES OF AI

What are the possible risks of using AI technology?

- AI might be used towards undesirable ends
  - e.g., surveillance by speech recognition, detection of “terrorist phrases”

- AI might result in a loss of accountability
  - what’s the legal status of a self-driving car?
  - or a medical expert system?
  - or autonomous military attack drones?

- AI might mean the end of the human race
  - can a military AI start a neuclear war? (accidentally or not)
  - if we get superintelligent robots, will they care about humans?