COMP9208
Artificial Intelligence and Society

Lecture 9: 10 October 2025

Prof. Mikhail Prokopenko
Dr. Sheryl Chang

Canvas: https://canvas.sydney.edu.au/courses/66452

Email: mikhail.prokopenko@sydney.edu.au
Office: J12 (Computer Science), level 3W, room 324

Date	Week: Lecture	Assignment
08/08	1: History of AI	
15/08	2: History of AI + Perception and Actuation	
22/08	3: History of AI + Representation and Reasoning	
29/08	4: Machine Learning	
05/09	5: Natural Interaction	
12/09	6: Distributed AI (swarm intelligence)	
19/09	7: From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)	# 1: Individual assignment (article review)
26/09	8: Goal-driven AI: search & rescue vs lethal autonomous weapon systems (LAWS)	
03/10	**BREAK**	
10/10	9: Review: language models, search trees, situated behaviours, neural networks	
17/10	10: Robotic swarms: RoboCop vs RoboCup	
24/10	11: Group Presentations	# 2: Group assignment
31/10	12: Adversarial Machine Learning → BHB.LT.2140	
07/11	13: Limits of cognition	
17/11	14: STUVAC	# 3: Individual assignment (data analysis report)

Brief review of last week

?

AI in Search and Rescue (SAR)

The Pentagon Wants AI-Driven Drone Swarms for Search and Rescue Ops

ANDY DEAN PHOTOGRAPHY/SHUTTERSTOCK.COM

https://www.nextgov.com/artificial-intelligence/2019/12/pentagon-wants-ai-driven-drone-swarms-search-and-rescue-ops/162113/

AI in Search and Rescue (SAR): pattern recognition

Figure 12. False-positive examples for the ScoreMap to ROI algorithm: (**a**,**c**) The algorithm, due to the shadow, made a wrong conclusion, (**b**) The stone and a blue plastic bag create an object that is almost impossible to distinguish from a human.

(a) (b) (c)

https://www.mdpi.com/2072-4292/12/20/3295

AI in Search and Rescue (SAR): path planning and optimisation

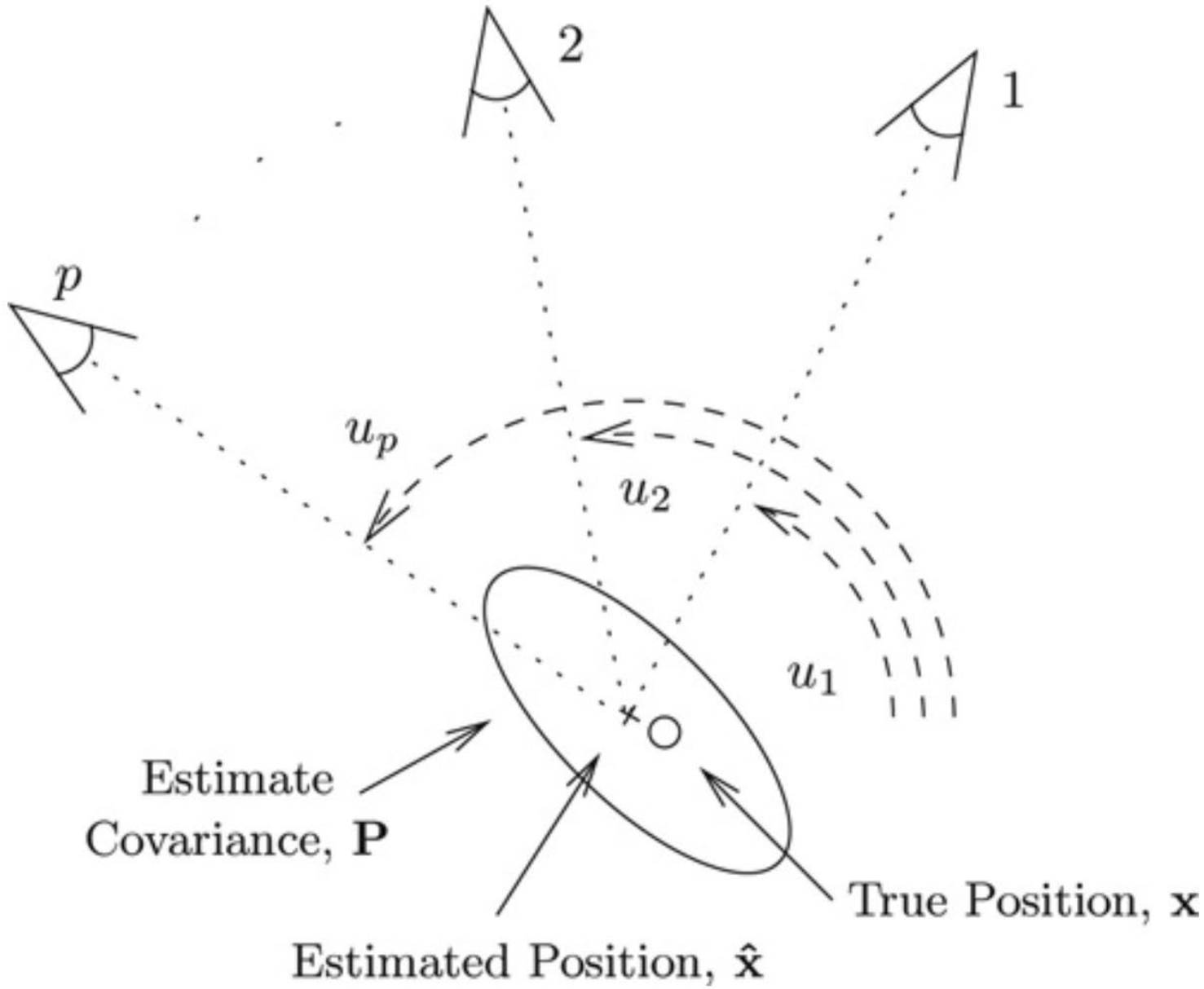


Fig. 5. Feature localisation Task. Each agent is equipped with a range sensor and must decide on which angle to view the feature.

G. Mathews, H. F. Durrant-Whyte, M. Prokopenko. Decentralised decision making in heterogeneous teams using anonymous optimisation, *Robotics and Autonomous Systems*, 57(3): 310-320, 2009.

AI in Search and Rescue (SAR): robustness and reliability

https://www.csiro.au/en/news/All/Articles/2021/December/Robots-to-the-rescue-in-hazardous-environments

Existential risks

?

Existential risks

- Weaponisation of AI

Photo: US Department of Defense / Sgt. Cory D. Payne, public domain

https://unu.edu/article/militarization-ai-has-severe-implications-global-security-and-warfare

Existential risks

- Failure to protect AI rights

https://www.linkedin.com/pulse/ai-uprising-when-robots-ask-seat-table-aditi-kapoor-n8r0f

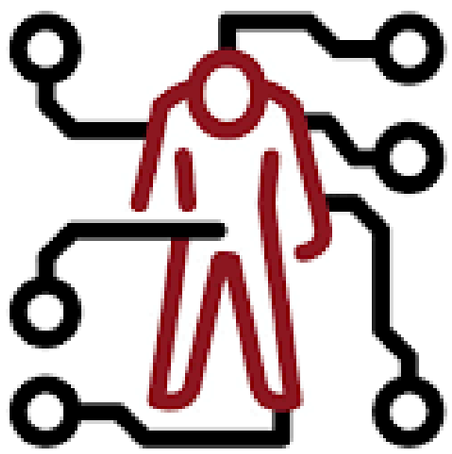
Existential risks

- AI uprising

https://apps.microsoft.com/detail/xm8d6kgwxvhz9h?hl=en-US&gl=US

Existential risks

- Strong dependence on AI
  - critical infrastructure systems become excessively dependent on AI
  - humanity loses control over critical systems: AI systems make decisions faster than human oversight can respond
  - human expertise disintegrates because important functions are assigned and transferred to AI systems: “enfeeblement”
  - humanity loses the ability to self-govern
  - humanity becomes completely dependent on AI systems



Enfeeblement

https://arxiv.org/pdf/2206.05862

Homework

- check the information sources provided on different slides (at least some)
  - chapters 7 & 8 of Michael Wooldridge, “A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going”, Flatiron Books, 2021
- there are several movies mentioning the existential risks of AGI:
  - for example, “A.I. Artificial Intelligence”, “Bicentennial Man”, “Blade Runner” and “Blade Runner 2049”, “Ex Machina”, “Her”, "I, Robot“, “Matrix” trilogy, “Minority Report”, “Surrogates”, “Transcendence”, among others
- please try to watch at least one before the next lecture, and prepare to give a few brief comments about the movies(s):
  - what was most striking about the AI technology in the movies(s)?
  - what societal impact did the AI make?
  - what were the main AI limitations, biases and risks?

?

Questions

?

Assignment #2
(group presentation)

Prepare a presentation video (10 minutes), upload it to a platform of your choice, and send an email with a video link to mikhail.prokopenko@sydney.edu.au

Upload presentation slides to Canvas (with your Group ID in the file name)

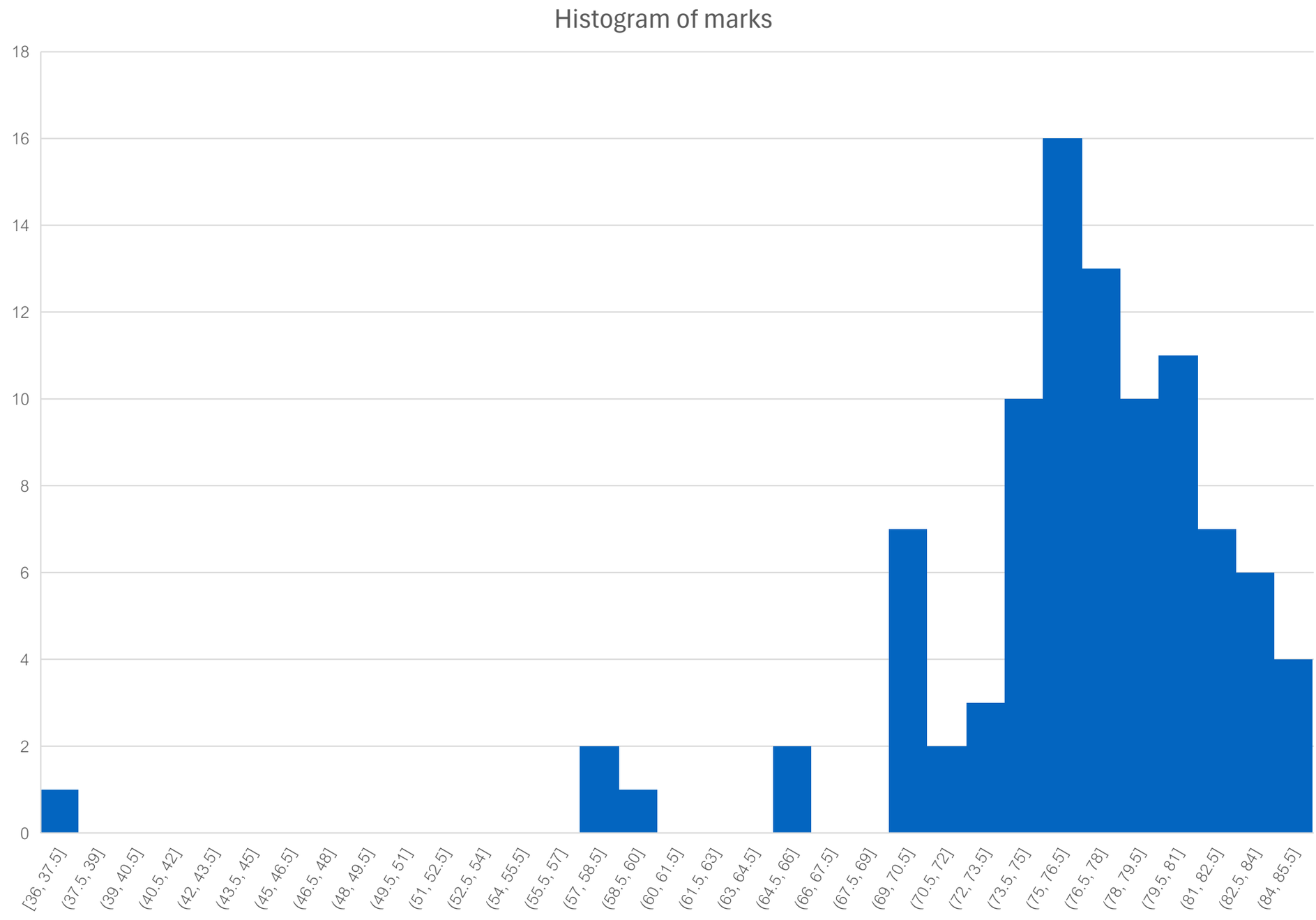
Deadline for slides upload: **23 October, Thursday, 11:59 pm (week 11)**

Q&A: 24 October (in class): 8 minutes Q&A for each group

There will be 20 groups in total: 16 groups with 5 students, and 4 groups with 4 students

Assignment #1 review

Maximum	25.5 / 30	Average	22.88 / 30
Minimum	10.8 / 30	Median	23.0 / 30



Articles to review

1. “Computing Machinery and Intelligence”, Alan Turing (1950)
2. “A Chess-Playing Machine”, Claude E. Shannon (1950)
3. “Why Should Machines Learn”, Herbert Simon (1983)
4. “Flocks, Herds, and Schools: A Distributed Behavioral Model”, Craig Reynolds (1987)
5. “Elephants Don't Play Chess”, Rodney A. Brooks (1990)
6. “The Myth of the Last Metaphor”, Joseph Weizenbaum (1995)
7. “RoboCup: The Robot World Cup Initiative”, Hiroaki Kitano et al. (1996)
8. “The RoboCup Synthetic Agent Challenge”, Hiroaki Kitano et al. (1997)

Articles to review

9. “Perceptron, Encyclopedia of Computer Science”, Laveen Kanal (2003)

10. “Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm”, Nathan Ensmenger (2012)

11. “A Few Useful Things to Know About Machine Learning”, Pedro Domingos (2012)

12. “Overview on DeepMind and Its AlphaGo Zero AI”, Sean Holcomb et al. (2018)

13. “Conversations with ELIZA on Gender and Artificial Intelligence”, Pedro Costa & Luisa Ribas (2018)

14. “The Five Tribes of Machine-Learning: A Brief Overview”, Jens Pohl (2019)

15. “Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo”, Paolo Bory (2019)

16. “ChatGPT Is a Blurry JPEG of the Web”, Ted Chiang (2023)

Articles to review

17. "Steps Toward Artificial Intelligence", Minsky (1961)

18. "Learning representations by back-propagating errors", David E. Rumelhart et al. (1986)

19. "Heuristic Problem Solving The Next Advance in Operations Research", Simon & Newell (1958)

20. "Intelligence Without Representation", Rodney A. Brooks (1991)

21. "Deep learning", Yann LeCun et al. (2015)

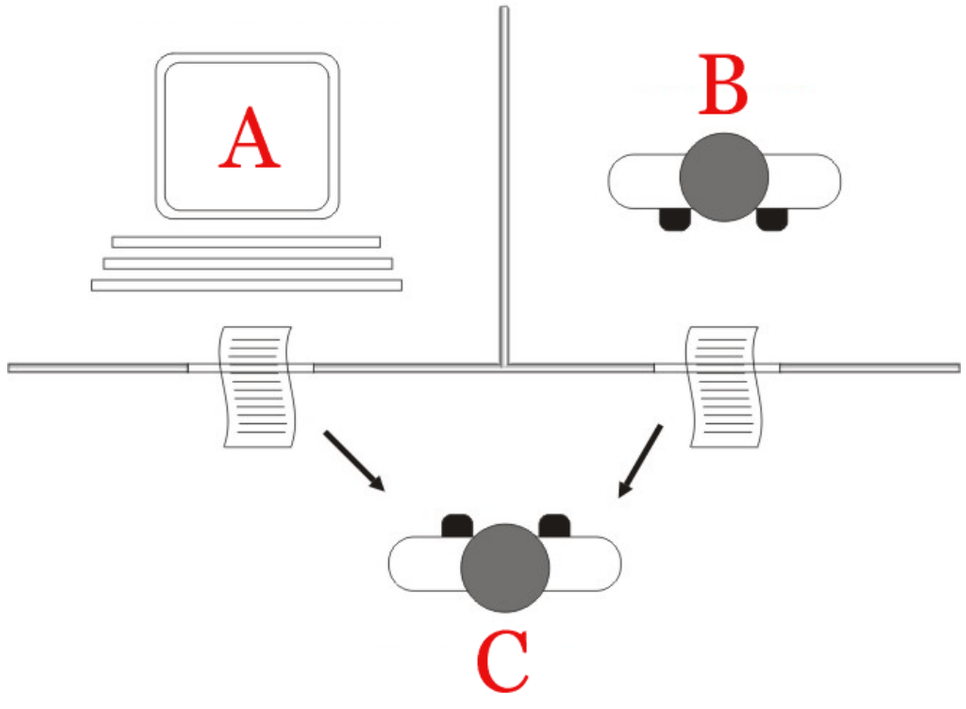
22. "ImageNet Classification with Deep Convolutional Neural Networks", Alex Krizhevsky et al. (2012)

23. "Datasheets for Datasets", Gebru et al. (2018)

24. "There is a blind spot in AI research", Crawford & Calo (2016)

Natural Interaction

1. “Computing Machinery and Intelligence”, Alan Turing (1950)
6. “The Myth of the Last Metaphor”, Joseph Weizenbaum (1995)
13. “Conversations with ELIZA on Gender and Artificial Intelligence”, Pedro Costa & Luisa Ribas (2018)
16. “ChatGPT Is a Blurry JPEG of the Web”, Ted Chiang (2023)
17. "Steps Toward Artificial Intelligence", Minsky (1961)
18. "Heuristic Problem Solving The Next Advance in Operations Research", Simon & Newell (1958)



https://en.wikipedia.org/wiki/Turing_test

https://www.youtube.com/watch?v=wGWVKkYEHBE

1. “Computing Machinery and Intelligence”, Alan Turing (1950)

I propose to consider the question, ‘Can machines think?’

The new form of the problem can be described in terms of a game which we call the ‘imitation game’.

...we only permit digital computers to take part in our game.

...the question, ‘Can machines think?’ should be replaced by ‘Are there imaginable digital computers which would do well in the imitation game?’

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.”

Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops.

It is probably wise to include a random element in a learning machine.

6. “The Myth of the Last Metaphor”, Joseph Weizenbaum (1995)

...when two people talk to each other, they both share some common knowledge.

I relied on that and built a system that imitated – a much better word would be parodied – the psychiatrist in the first session.

...what happened ... is that psychiatrists and other people took this parody seriously. I was stunned by that.

...do I want to play that kind of role?

...one of the important questions that ELIZA generated was the question of what it means for a machine to understand anything at all...

...they attribute what they think to be the human motivation for such a question to the computer.

...however intelligent these things become, whatever power they have to come to understand their environment, to move in it and to modify it, their intelligence must always be very different from and alien to human intelligence.

13. “Conversations with ELIZA on Gender and Artificial Intelligence”, Pedro Costa & Luisa Ribas (2018)

This paper aims to explore how gender relates to AI, while also seeking to understand why most chatbots and digital assistants appear to be female.

ELIZA was one of the first “natural language process applications” that was able to trick some of its users into thinking it “was a person rather than a machine”...

...when we interact with machines as if they were human, we start developing emotional bonds, a sense of empathy and of being understood.

...we can observe how general or specialized chatbots automate work that is coded as female…, given that they mainly operate in service or assistance related contexts, acting as personal assistants, secretaries and the like.

When bots interact in a motherly logic, attachment also conforms to expectations and stereotypes that associate femininity with emotional and domestic caregiving acts.

16. “ChatGPT Is a Blurry JPEG of the Web”, Ted Chiang (2023)

Think of ChatGPT as a blurry jpeg of all the text on the Web. It retains much of the information on the Web, in the same way that a jpeg retains much of the information of a higher-resolution image, but, if you’re looking for an exact sequence of bits, you won’t find it; all you will ever get is an approximation.

But, because the approximation is presented in the form of grammatical text, which ChatGPT excels at creating, it’s usually acceptable. You’re still looking at a blurry jpeg, but the blurriness occurs in a way that doesn’t make the picture as a whole look less sharp.

This analogy to lossy compression is... also a way to understand the “hallucinations,” or nonsensical answers to factual questions, to which large language models such as ChatGPT are all too prone. These hallucinations are compression artifacts...

Large language models identify statistical regularities in text.

I’m going to make a prediction: when assembling the vast amount of text used to train GPT-4, the people at OpenAI will have made every effort to exclude material generated by ChatGPT or any other large language model. ... If the output of ChatGPT isn’t good enough for GPT-4, we might take that as an indicator that it’s not good enough for us, either.

...starting with a blurry copy of unoriginal work isn’t a good way to create original work.

It’s possible that, in the future, we will build an A.I. that is capable of writing good prose based on nothing but its own experience of the world. The day we achieve that will be momentous indeed – but that day lies far beyond our prediction horizon.

17. "Steps Toward Artificial Intelligence", Minsky (1961)

The problems of heuristic programming – of making computers solve really difficult problems – are divided into five main areas: Search, Pattern-Recognition, Learning, Planning, and Induction.

A computer can do, in a sense, only what it is told to do. But even when we do not know how to solve a certain problem, we may program a machine (computer) to Search through some large space of solution attempts. Unfortunately, this usually leads to an enormously inefficient process.

With Pattern-Recognition techniques, efficiency can often be improved, by restricting the application of the machine's methods to appropriate problems.

Pattern-Recognition, together with Learning, can be used to exploit generalizations based on accumulated experience, further reducing search.

By analyzing the situation, using Planning methods, we may obtain a fundamental improvement by replacing the given search with a much smaller, more appropriate exploration.

To manage broad classes of problems, machines will need to construct models of their environments, using some scheme for Induction.

19. "Heuristic Problem Solving The Next Advance in Operations Research", Simon & Newell (1958)

We are now poised for a great advance that will bring the digital computer and the tools of mathematics and the behavioral sciences to bear on the very core of managerial activity on the exercise of judgment and intuition; on the processes of making complex decisions.

In short, well-structured problems are those that can be formulated explicitly and quantitatively, and that can then be solved by known and feasible computational techniques.

In dealing with the ill-structured problems of management we have not had the mathematical tools we have needed we have not had 'judgment mechanics'

And we now know, at least in a limited area, not only how to program computers to perform such problem-solving activities successfully; we know also how to program computers to learn to do these things.

In short, we now have the elements of a theory of heuristic (as contrasted with algorithmic) problem solving; and we can use this theory both to understand human heuristic processes and to simulate such processes with digital computers.

Questions

?

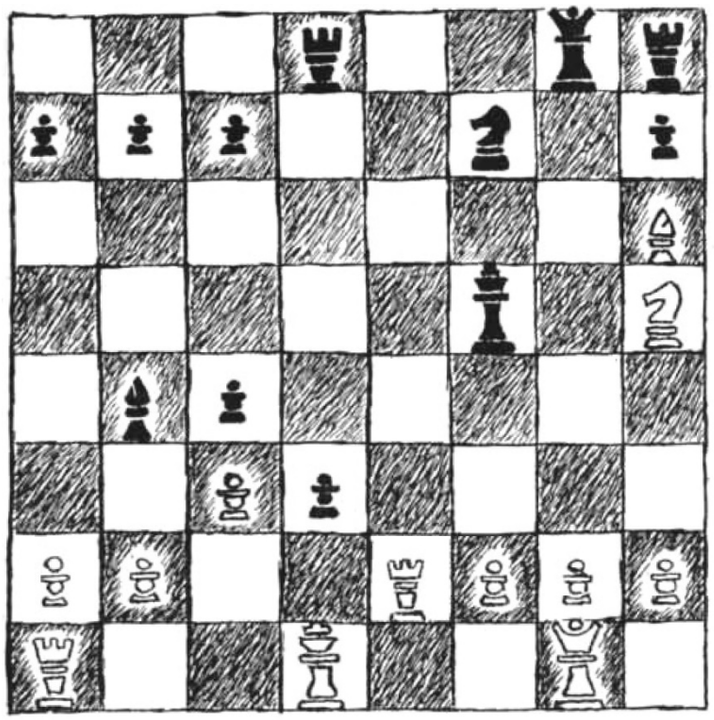
Games and AI

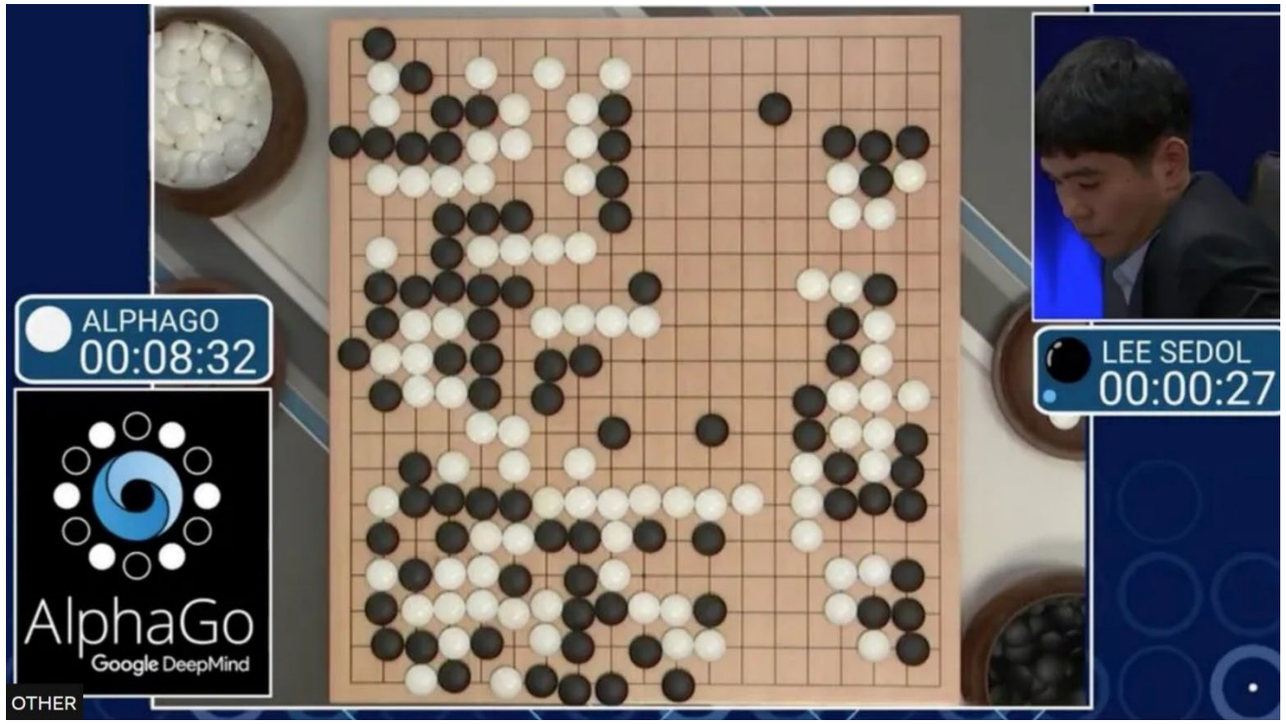
2. “A Chess-Playing Machine”, Claude E. Shannon (1950)

10. “Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm”, Nathan Ensmenger (2012)

12. “Overview on DeepMind and Its AlphaGo Zero AI”, Sean Holcomb et al. (2018)

15. “Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo”, Paolo Bory (2019)





https://www.bbc.com/news/technology-35785875

2. “A Chess-Playing Machine”, Claude E. Shannon (1950)

Electronic computers can be set up to play a fairly strong game, raising the question of whether they can "think“.

The problem is sharply defined, both in the allowed operations (the moves of chess) and in the ultimate goal (checkmate).

...deciding on a strategy of play ...is the most difficult part of the problem.

...a means of bringing some order to the maze of possible variations of a chess game.

Even the high speeds available in electronic computers are hopelessly inadequate to play perfect chess by calculating all possible variations to the end of the game.

...one must establish a method of numerical evaluation for any given chess position.

Assuming that the opponent's reply will be the one giving the best evaluation from his point of view, we would choose the move that would leave us as well off as possible after his best reply.

A suitable compromise would be to examine only the important possible variations.

One possibility is to devise a program that would change the terms and coefficients involved in the evaluation function on the basis of the results of games the machine has already played.

10. “Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm”, Nathan Ensmenger (2012)

The central argument is that the decision to focus on chess as a representative measure of both human and computer intelligence had important and unintended consequences for the discipline.

The rise to dominance of minimax algorithm-based techniques in particular transformed chess from a quintessentially human intellectual activity into an exercise in deep searching and fast alpha–beta pruning.

Computer chess in this respect will serve not as the drosophila of AI, but as the drosophila of the history of AI.

...the question of whether it was necessary for AI to simulate (and therefore understand) natural intelligence, or whether it was enough simply to replicate its functionality.

Type-A solution, a brute force approach built around the so-called minimax algorithm...

‘Type-B’ solution that used sophisticated heuristics to trim the decision tree by privileging certain branches over others...

...the outcome represented what appeared to be a strong position, but in the long term produced a losing outcome. This is known as the ‘horizon effect’.

The way in which a minimax-based machine plays chess is not at all like the way a human plays chess; in many respects, the two are playing an entirely different game.

AI’s obsession with games – and not any particular game – is the more fundamental problem.

12. “Overview on DeepMind and Its AlphaGo Zero AI”, Sean Holcomb et al. (2018)

The backbone of AlphaGo Zero's implementation had three main components being the neural network, use of Monte Carlo Tree Search, and reinforcement learning algorithm.

AlphaGo was fed large quantities of human-played games to learn the game whereas Zero learned purely through reinforcement playing itself.

One area where DeepMind differed in their implementation compared to those of others is the use of a convolutional deep neural network for their policy.

AlphaGo Zero used its own method through neural networking to guide the MCTS (Monte Carlo Tree Search).

…as opposed to standard supervised and unsupervised machine learning techniques, AlphaGo Zero is not fed labeled data to train upon or given raw data with no sense of what needs to be accomplished other than clustering similar data. Rather, self-play is done where the prediction method is formed from the results of the games played.

15. “Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo”, Paolo Bory (2019)

The victory of machines over human players is a clear example of a narrative strategy apt to astonish the public and drive attention towards a specific narrative of newness and technological change.

Notably, intelligent systems are the main characters of a future social sphere in which humans and machines would not only coexist, but also dialogue, empathize or even, in the worst-case scenario, run into conflict.

...the Kasparov–Deep Blue challenge was presented ... as a conflictual and competitive struggle between mankind and a hardware-based, inscrutable prototype of AI that was eventually humanized after its unexpected disposal.

...in 2016 Google DeepMind’s AlphaGo was narrated quite differently, as an emergent software-based, transparent and un-humanlike form of intelligence.

...board games like chess and Go are inspired by the surrounding and positional strategies of war.

...the never-ending competition between humans and machines can turn into a collaborative interaction between complementary intelligences.

Questions

?

Break

?

Robots and Swarms

4. “Flocks, Herds, and Schools: A Distributed Behavioral Model”, Craig Reynolds (1987)

5. “Elephants Don't Play Chess”, Rodney A. Brooks (1990)

20. "Intelligence Without Representation", Rodney A. Brooks (1991)

7. “RoboCup: The Robot World Cup Initiative”, Hiroaki Kitano et al. (1996)

8. “The RoboCup Synthetic Agent Challenge”, Hiroaki Kitano et al. (1997)

4. “Flocks, Herds, and Schools: A Distributed Behavioral Model”, Craig Reynolds (1987)

...the strong impression of intentional, centralized control. Yet all evidence indicates that flock motion must be merely the aggregate result of the actions of individual animals, each acting solely on the basis of its own local perception of the world.

…individual subobjects have a more complex geometrical state; they now have orientation.

Boid behavior is dependent not only on internal state but also on external state.

The basic urge to join a flock seems to be the result of evolutionary pressure from several factors: protection from predators, statistically improving survival of the (shared) gene pool from attacks from predators, profiting from a larger effective search pattern in the quest for food, and advantages for social and mating activities.

…the amount of "thinking“ that a bird has to do in order to flock must be largely independent of the number of birds in the flock.

Prioritized acceleration allocation is based on a strict priority ordering of all component behaviors.

5. “Elephants Don't Play Chess”, Rodney A. Brooks (1990)

…the *symbol system hypothesis* upon which *classical AI* is based is fundamentally flawed.

…there is an alternative view, or dogma, variously called *nouvelle AI*, *fundamentalist AI*, or in a weaker form *situated activity. It is based on the physical grounding hypothesis*.

The new methodology bases its decomposition of intelligence into individual behavior generating modules, whose coexistence and cooperation let more complex behaviors emerge.

…to build a system that is intelligent it is necessary to have its representations grounded in the physical world.

The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough.

A subsumption program is built on a computational substrate that is organized into a series of incremental layers, each, in the general case, connecting perception to action.

Traditional AI has tried to demonstrate sophisticated reasoning in rather impoverished domains. ...Nouvelle AI tries to demonstrate less sophisticated tasks operating robustly in noisy complex domains.

20. "Intelligence Without Representation", Rodney A. Brooks (1991)

The fundamental decomposition of the intelligent system is not into independent information processing units which must interface with each other via representations.

Instead, the intelligent system is decomposed into independent and parallel activity producers which all interface directly to the world through perception and action, rather than interface to each other particularly much.

The notions of central and peripheral systems evaporate: everything is both central and peripheral.

…problem solving behavior, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction.

Our belief is that the sorts of activity producing layers of control we are developing (mobility, vision and survival related tasks) are necessary prerequisites for higher-level intelligence in the style we attribute to human beings.

7. “RoboCup: The Robot World Cup Initiative”, Hiroaki Kitano et al. (1996)

We propose a Robot World Cup (RoboCup) as a new standard problem for AI and robotics research.

This is a proposal to use a soccer game as a platform for a wide range of AI and robotics research, such as design principles of autonomous agents, multi-agent collaboration, strategy acquisition, real-time reasoning, and sensor fusion.

RoboCup's final target is a world cup with real robots.

(The primary objective of RoboCup is to foster the development of a robotic team that can win a soccer game against the human world champion team by the year 2050.)

Standard AI problems have been the basic driving force for AI research. Research on computer chess, which is the most typical example of a standard problem, lead to the discovery of various powerful search algorithms.

A soccer game is a specific but very attractive real-time multi-agent environment from the viewpoint of distributed artificial intelligence and multi-agent research.

8. “The RoboCup Synthetic Agent Challenge”, Hiroaki Kitano et al. (1997)

RoboCup (The World Cup Robot Soccer) is an attempt to promote AI and robotics research by providing a common task, Soccer, for evaluation of various theories, algorithms, and agent architectures.

The range of technologies spans both AI and robotics research, such as design principles of autonomous agents, multi-agent collaboration, strategy acquisition, real-time reasoning and planning, intelligent robotics, sensor-fusion, and so forth.

Because the state-space of the soccer game is prohibitively large for anyone to hand-code all possible situations and agent behaviors, it is essential that agents learn to play the game strategically.

Teamwork in complex, dynamic multi-agent domains such as Soccer mandates highly flexible coordination and communication to surmount the uncertainties, e.g., dynamic changes in team's goals, team members' unexpected inability to fulfil responsibilities, or unexpected discovery of opportunities.

Of particular interest is variations seen in agent-modelers behaviors given the modification in the opponents' behaviors.

Questions

?

Machine Learning

3. “Why Should Machines Learn”, Herbert Simon (1983)
9. “Perceptron, Encyclopedia of Computer Science”, Laveen Kanal (2003)
11. “A Few Useful Things to Know About Machine Learning”, Pedro Domingos (2012)
14. “The Five Tribes of Machine-Learning: A Brief Overview”, Jens Pohl (2019)
18. "Learning representations by back-propagating errors", David E. Rumelhart et al. (1986)
21. "Deep learning", Yann LeCun et al. (2015)
22. "ImageNet Classification with Deep Convolutional Neural Networks", Alex Krizhevsky et al. (2012)

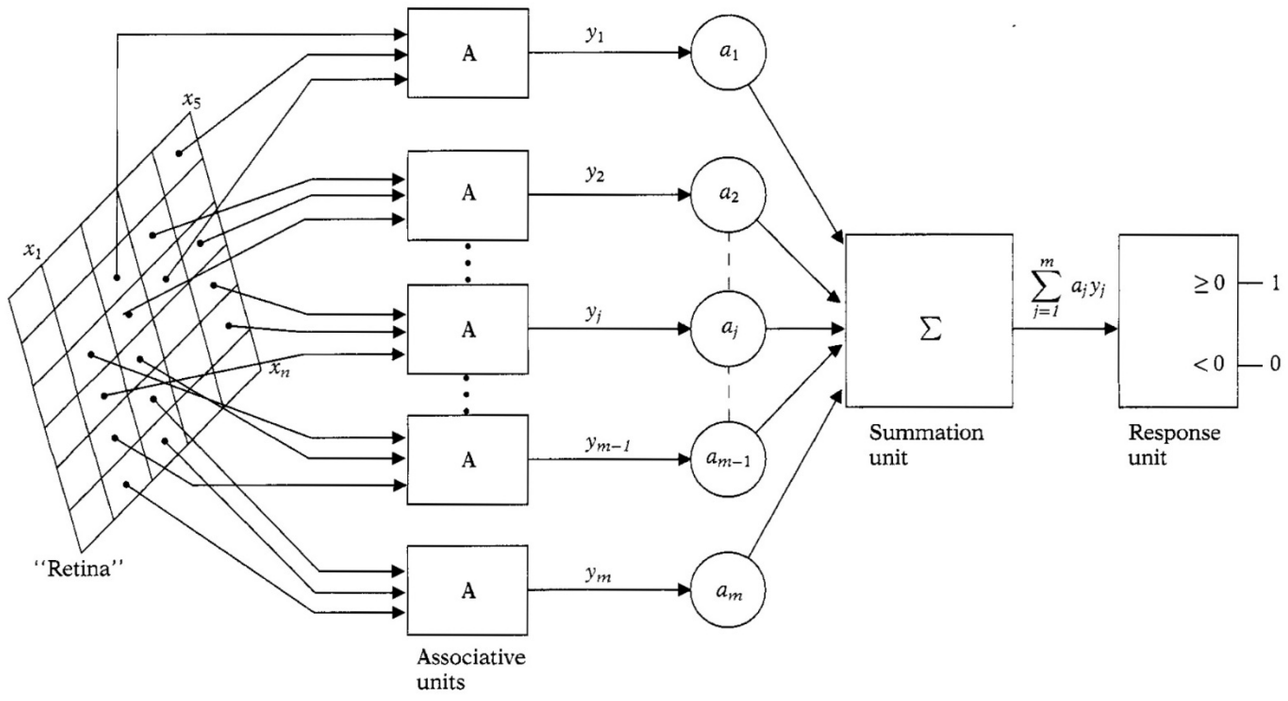
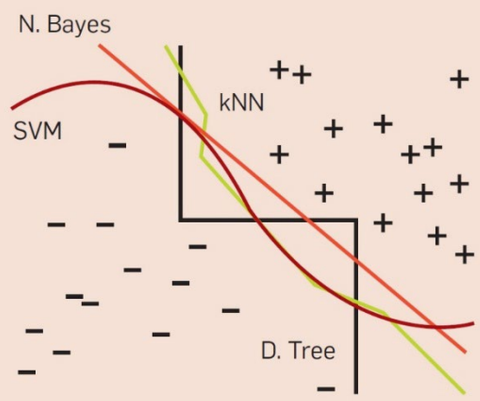


Figure 3. Very different frontiers can yield similar predictions. (+ and – are training examples of two classes.)



3. “Why Should Machines Learn”, Herbert Simon (1983)

Learning denotes changes in the system that are adaptive in the sense that they enable, the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time.

...tuning the evaluation function for positions on the basis of outcomes.

It's a lot more convenient to have the system index itself as it goes along, than to index it by hand.

Perception research... this line of research didn't get anywhere. ...strengthen our skepticism that the problems of AI are to be solved solely by building learning systems.

The ISAAC Program... uses its schemas and productions to produce an internal representation of the problem. ...While it understands problems, how about understanding physics! This would require the ability to use natural language information to construct new schemas and new productions.

...the only way to bring about continual modification and improvement of the program is by means of learning procedures that don't involve knowing the detail of the internal languages and programs.

9. “Perceptron, Encyclopedia of Computer Science”, Laveen Kanal (2003)

In 1957 the psychologist Frank Rosenblatt proposed “The Perceptron: a perceiving and recognizing automaton” as a class of artificial nerve nets, embodying aspects of the brain and receptors of biological systems.

A perceptron is a signal transmission network consisting of sensory units (S units), association units (A units), and output or response units (R units).

The design of a perceptron to distinguish between two given sets of patterns involves adjusting the weights ... and the threshold...

The output of the perceptron is monitored to determine whether a pattern is correctly classified. If not, the weights are adjusted according to... “error correction” procedure.

There is no layer of “hidden” elements – i.e. elements for which the adjustment is only indirectly related to the output error.

For series-coupled perceptrons with multiple R units, Rosenblatt proposed a “back-propagating error correction” procedure that used error from the R units to propagate correction back to the sensory end.

The demise of funding for perceptron-type networks should be blamed on the inadequate technology and training algorithms available for multilayer perceptrons and the premature, overblown results promised the funding agencies.

11. “A Few Useful Things to Know About Machine Learning”, Pedro Domingos (2012)

A *classifier* is a system that inputs (typically) a vector of discrete and/or continuous *feature values* and outputs a single discrete value, the *class*.

Learning = Representation + Evaluation + Optimization

The fundamental goal of machine learning is to generalize beyond … the training set.

The most common mistake … is to test on the training data and have the illusion of success. If the chosen classifier is then tested on new data, it is often no better than random guessing.

Contamination of your classifier by test data can occur in insidious ways, for example, if you use test data to tune parameters and do a lot of tuning.

Every learner must embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it.

...induction (what learners do) is a knowledge lever: it turns a small amount of input knowledge into a large amount of output knowledge.

…the similarity-based reasoning that machine learning algorithms depend on (explicitly or implicitly) breaks down in high dimensions.

As a rule of thumb, a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it.

14. “The Five Tribes of Machine-Learning: A Brief Overview”, Jens Pohl (2019)

Is there any way to learn something from the past that we can be confident will hold true for the future? This is a fundamental problem for machine-learning.

...the objective in machine-learning is to include just sufficient knowledge in our program so that the learner can continue to acquire knowledge ad infinitum by reading and synthesizing data.

It starts with the selection of the features of the data that we consider to be relevant.

...there are really only five principal approaches...: connectionist algorithms endeavor to simulate the way the human brain works; evolutionary algorithms draw on genetics and evolutionary biology; Bayesian algorithms rely on statistical probabilities; analogy algorithms rely on similarity judgments; and, symbolistic algorithms use inverse deduction.

An important part of the learning process of an infant is that the infant is actively engaged in the environment by being able to touch objects, play with them, experiment, use them as tools to pursue objectives, and so on.

...the infant’s learning capabilities develop as an integral part of the learning experience. This is a very different paradigm in comparison with a machine-learning algorithm, whose operational mechanism and capabilities are fixed from the start with the objective of using these capabilities to extract knowledge from data.

...a master algorithm that would bring machine learning to an Artificial General Intelligence (AGI) level will require the seamless integration of the distinctly different approaches pursued by the connectionists, evolutionists, Bayesians, analogists, and symbolists.

18. “Learning representations by back-propagating errors", David E. Rumelhart et al. (1986)

We describe a new learning procedure, back-propagation, for networks of neurone-like units.

The aim is to find a set of weights that ensure that for each input vector the output vector produced by the network is the same as (or sufficiently close to) the desired output vector.

The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector.

As a result of the weight adjustments, internal ‘hidden’ units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units.

The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure.

The learning procedure, in its current form, is not a plausible model of learning in brains.

However, applying the procedure to various tasks shows that interesting internal representations can be constructed by gradient descent in weight-space, and this suggests that it is worth looking for more biologically plausible ways of doing gradient descent in neural networks.

21. "Deep learning", Yann LeCun et al. (2015)

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.

These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.

Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer.

Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences.

We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data.

“Blind spots”

23."Datasheets for Datasets", Gebru et al. (2018)

24."There is a blind spot in AI research", Crawford & Calo (2016)

cago police use algorithmic systems to predict which people are most likely to be involved in a shooting, but they have proved largely ineffective.

23. "Datasheets for Datasets", Gebru et al. (2018)

Every machine learning model is trained and evaluated using data, quite often in the form of static datasets.

The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases.

Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice, hiring, critical infrastructure, and finance.

To address this gap, we propose datasheets for datasets. …we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on.

Datasheets for datasets have the potential to increase transparency and accountability within the ML community, mitigate unwanted societal biases in ML models, facilitate greater reproducibility of ML results, and help researchers and practitioners select more appropriate datasets for their chosen tasks.

24. "There is a blind spot in AI research", Crawford & Calo (2016)

Autonomous systems are already deployed in our most crucial social institutions, from hospitals to courtrooms. Yet there are no agreed methods to assess the sustained effects of such applications on human populations.

“People worry that computers will get too smart and take over the world, but the real problem is that they’re too stupid and they’ve already taken over the world.” (Pedro Domingos)

…the downsides of AI systems disproportionately affect groups that are already disadvantaged by factors such as race, gender and socio-economic background

So far, there have been three dominant modes of responding to concerns about the social and ethical impacts of AI systems: compliance, ‘values in design’ and thought experiments. All three are valuable. None is individually or collectively sufficient.

We believe that a fourth approach is needed. A practical and broadly applicable social-systems analysis thinks through all the possible effects of AI systems on all parties. It also engages with social impacts at every stage — conception, design, deployment and regulation.

Discussion

?