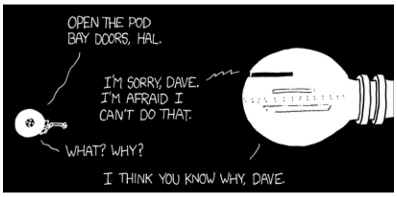
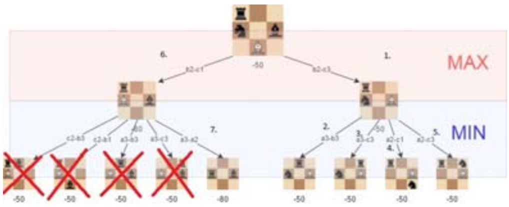

OPEN THE POD BAY DOORS, HAL.
I'M SORRY, DAVE. I'M AFRAID I CAN'T DO THAT.
WHAT? WHY?
I THINK YOU KNOW WHY, DAVE.

Explainable Artificial Intelligence

Johan Kwisthout, DCC / AI

Thank you for an enjoyable game!

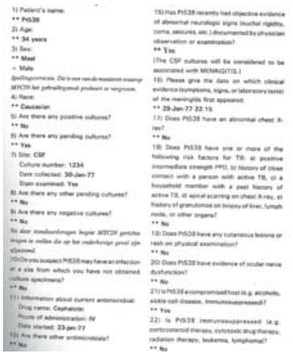
Minimax with alpha-beta pruning



I'm sorry, Frank, I think you missed it

- Frank missed, that the minimax algorithm returns -1, indicating a winning path for black
- Frank missed, that $\exists m_B \forall m_W \exists m_B \forall m_W$ ... black wins
- In the (science fiction) movie, HAL **explained** to Frank why he played a sure loss by **exemplifying** a 'natural' sequence of moves that made it **obvious** to **him** that Frank's position was hopeless
- (There's a whole lot more going on in this scene that is not relevant for this talk…)

Explanation in MYCIN expert system

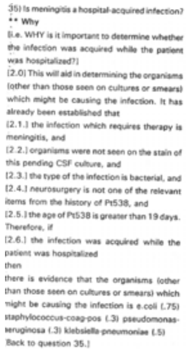


1) Patient's name:
** Pt538
2) Age:
** 34 years
3) Sex:
** Mael
= Male
Spellingcorrectie. Dit is een van de manieren waarop MYCIN het gebruiksgemak probeert te vergroten.
4) Race:
** Caucasian
5) Are there any positive cultures?
** No
6) Are there any pending cultures?
** Yes
7) Site: CSF
Culture number: 1234
Date collected: 30-Jan-77
Stain examined: Yes
8) Are there any other pending cultures?
** No
9) Are there any negative cultures?
** No
Na deze standaardvragen begint MYCIN gerichte vragen te stellen die op het onderliggende geval zijn afgestemd.
10) Do you suspect Pt538 may have an infection at a site from which you have not obtained culture specimens?
** No
11) Information about current antimicrobial:
Drug name: Cephalotin
Route of administration: IV
Date started: 23-Jan-77
12) Are there other antimicrobials?
** No

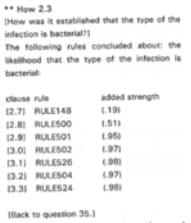
15) Has Pt538 recently had objective evidence of abnormal neurologic signs (nuchal rigidity, coma, seizures, etc.) documented by physician observation or examination?
** Yes
[The CSF cultures will be considered to be associated with MENINGITIS.]
16) Please give the date on which clinical evidence (symptoms, signs, or laboratory tests) of the meningitis first appeared.
** 29-Jan-77 22:15
17) Does Pt538 have an abnormal chest X-ray?
** No
18) Does Pt538 have one or more of the following risk factors for TB: a) positive intermediate strength PPD, b) history of close contact with a person with active TB, c) a household member with a past history of active TB, d) apical scarring on chest X-ray, e) history of granulomas on biopsy of liver, lymph node, or other organs?
** No
19) Does Pt538 have any cutaneous lesions or rash on physical examination?
** No
20) Does Pt538 have evidence of ocular nerve dysfunction?
** No
21) Is Pt538 a compromised host (e.g. alcoholic, sickle-cell-disease, immunosuppressed)?
** Yes
22) Is Pt538 immunosuppressed (e.g. corticosteroid therapy, cytotoxic drug therapy, radiation therapy, leukemia, lymphoma)?
** No

- Diagnosis of infections and suggested antibiotics
- Developed in 1972 (!)
- Set of 600 rules, reasoning with uncertainty
- Sometimes outperforming human diagnoses
- Can explain and answer 'why' and 'how' questions

Explanation in MYCIN expert system



35) Is meningitis a hospital-acquired infection?
** Why
[i.e. WHY is it important to determine whether the infection was acquired while the patient was hospitalized?]
[2.0] This will aid in determining the organisms (other than those seen on cultures or smears) which might be causing the infection. It has already been established that
[2.1.] the infection which requires therapy is meningitis, and
[2.2.] organisms were not seen on the stain of this pending CSF culture, and
[2.3.] the type of the infection is bacterial, and
[2.4.] neurosurgery is not one of the relevant items from the history of Pt538, and
[2.5.] the age of Pt538 is greater than 19 days.
Therefore, if
[2.6.] the infection was acquired while the patient was hospitalized
then
there is evidence that the organisms (other than those seen on cultures or smears) which might be causing the infection is e.coli (.75) staphylococcus-coag-pos (.3) pseudomonas-aeruginosa (.3) klebsiella-pneumoniae (.5)
[Back to question 35.]



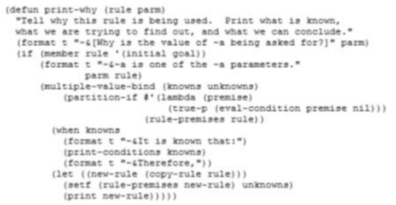
** How 2.3
[How was it established that the type of the infection is bacterial?]
The following rules concluded about: the likelihood that the type of the infection is bacterial:

clause	rule	added strength
[2.7]	RULE148	(.19)
[2.8]	RULE500	(.51)
[2.9]	RULE501	(.95)
[3.0]	RULE502	(.97)
[3.1]	RULE526	(.98)
[3.2]	RULE504	(.97)
[3.3]	RULE524	(.98)

[Back to question 35.]

Explanation in MYCIN expert system

- Maybe not so sophisticated as it looks…



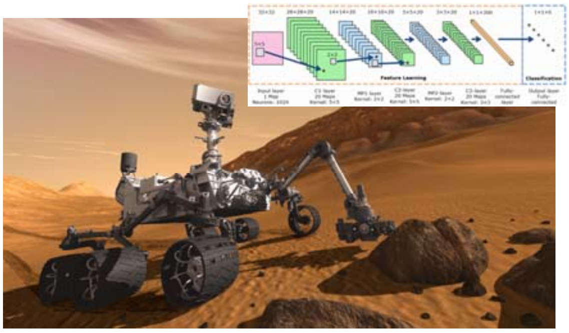
```
(defun print-why (rule parm)
  "Tell why this rule is being used.  Print what is known,
  what we are trying to find out, and what we can conclude."
  (format t "~&[Why is the value of ~a being asked for?]" parm)
  (if (member rule '(initial goal))
      (format t "~&~a is one of the ~a parameters."
              parm rule)
      (multiple-value-bind (knowns unknowns)
          (partition-if #'(lambda (premise)
                            (true-p (eval-condition premise nil)))
                        (rule-premises rule))
        (when knowns
          (format t "~&It is known that:")
          (print-conditions knowns)
          (format t "~&Therefore,"))
        (let ((new-rule (copy-rule rule)))
          (setf (rule-premises new-rule) unknowns)
          (print new-rule)))))
```

https://norvig.com/paip/mycin.lisp

MYCIN revisited

- MYCIN was a **rule-based** expert system within a very **specific domain**, all rules **hand-coded** based on expert knowledge elicitation
- Explanation in this system useful, but mechanistic
- In modern AI, information is mostly ***machine learned*** by discovering statistical patterns in data
- How can we let such systems explain their decisions to us?

Curiosity Mars Rover

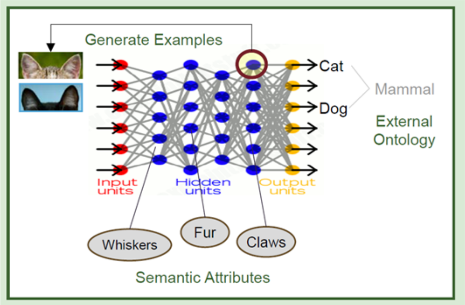


Explanations from Curiosity

- Under the hood, autonomous vehicles might base their decisions on (e.g.) deep neural networks
- Sub-symbolic AI: information decoded in weights between artificial neurons
- When asked 'why did you make that decision?' we don't want an answer like:

"because $f\left(\sum_{k=1}^{n} i_k \cdot W_k\right)$ > 0"!

Gunning (DARPA): XAI project



- Train the net to associate semantic attributes with hidden layer nodes and known ontologies

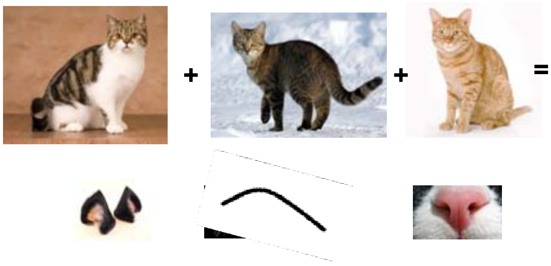
Generating examples

Learning to Generate Chairs with Convolutional Neural Networks

Alexey Dosovitskiy Jost Tobias Springenberg Thomas Brox
Department of Computer Science, University of Freiburg
{dosovits, springj, brox}@cs.uni-freiburg.de

Why is this a cat?

- Specific nodes in a deep layer of the neural network capture statistical abstractions – that may or may not correspond to salient features we recognize…



Pattern recognition ≠ explanation!

- Specific nodes in a deep layer of the neural network capture statistical abstractions – that may or may not correspond to salient features we recognize…
- Even granted that, and granted that we would provide the learning algorithm with a full ontology, just listing (selected) hidden nodes that are activated on an input is not so useful…
- **Inference to the best explanation**: what is the **cause** (or reason, principle, etc.) of the phenomena I can **observe**? Also called **abduction**

Deduction, induction, and abduction

- **Deduction**: from premises and a rule deduce a logically valid conclusion

  $(\forall x H(x) \rightarrow M(x), H(s) \vDash M(s))$

- **Induction**: from specific observations derive general principles

  (all biological life forms we know depend on liquid water to exist, so it is likely that there cannot be life without water)

- **Abduction**: from an observation derive a cause that best explains the observation

  (When I leave my house in the morning the grass is wet; given weather forecasts the best explanation is that it rained tonight)

Statistical association ≠ explanation!

- From data alone we cannot learn causal relations (“correlation is not causation”)
- Judea Pearl (2017): vital for AI to be able to ask questions “what if I do X” (intervention) and “what I had done Y” (counterfactuals)
- These questions are also important for Explainable AI!

Pearl at NIPS 2017

What is a good explanation?

- Explanation not just answers “**why this**”, but “why this, **rather than** that” (parsimoniously)
- Q “Why did Alice got tenure (while Bob didn’t)?”
- A1 “Alice had a good publication record”
  - But Bob had a good publication record as well! That doesn’t explain why she got tenure!
- A2: “Alice had a good publication record and did quality teaching”
  - Bob was a poor teacher, so this explains why Alice got tenure and Bob was denied tenure!

What is a good explanation?

- Explanation must be based on *relevant* information
- Q “Why did Alice got tenure (while Bob didn’t)?”
- A3 “Alice had a good publication record and wore glasses”
  - But Bob had a good publication record as well! And wearing glasses ought to be irrelevant for getting tenure! That doesn’t explain why she got tenure!
- But how do we decide that wearing glasses is not relevant *even if this might be statistically significant?*

Abduction ingredients

- Abduction is **making sense** of your observations in order to act accordingly and **motivate** your actions
- **Generate** candidate hypotheses that might explain the phenomena that have been observed
- Decide what is **relevant** in its sensory input and what is not
- Decide when to gather **new evidence** (e.g., re-orient your sensors, do additional tests) to reduce uncertainty
- From a set of candidate hypotheses, **select the best** one
- In a given context, **determine** what constitutes 'best'
- Try to infer **causal relationships** and test hypotheses by **interventions**, i.e., acting in the world
- Generate and reason through '**what if**' scenarios

Important

- Some or all of these 'ingredients' might be 'implemented' in modern sub-symbolic models (e.g. convolutional deep neural networks)
- Yet, for Explainable AI the challenges are the same as for (symbolic) GOFAI and philosophy of mind (e.g., symbol grounding, frame problem)
- It is one thing for AlphaGO to beat the world champion; explaining the rationale behind its moves in a way we can understand is a different ball game

Sherlock Holmes without Watson is 'useless'

Explainable AI as research method

- "What I cannot create, I do not understand"
- Try to implement a theory in order to identify its ambiguities, test consistency and completeness, and identify gaps
- Learn about human cognition by trying to implement your favorite theory of human sense-making in AI
- See also: Otworowska et al. (BNAIC 2015). "The Robo-havioral Methodology"

Conclusion and summary

- Explainable AI becomes more and more important particularly when the AI becomes a 'black box' as in deep neural networks and machine learning
- The challenges in creating Explainable AI are similar to the challenges in understanding human sense-making (e.g. frame problem)
- Apart from a research goal on its own, Explainable AI can be a research method to put theories of human sense-making to the test