Artificial & Computational Intelligence

AIML CLZG557

M7: Ethics in AI

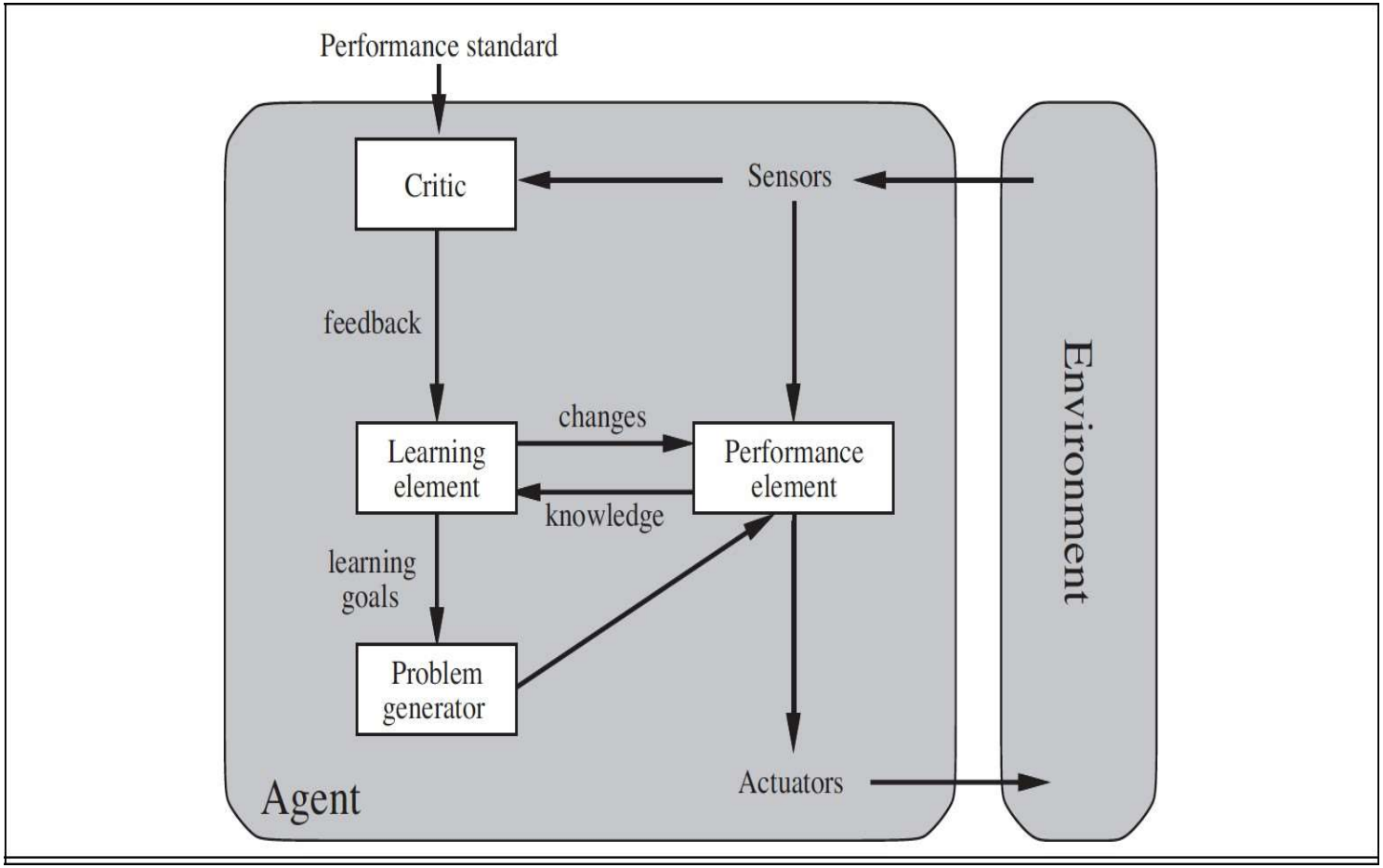
Dr. Sudheer Reddy

BITS Pilani
Pilani Campus

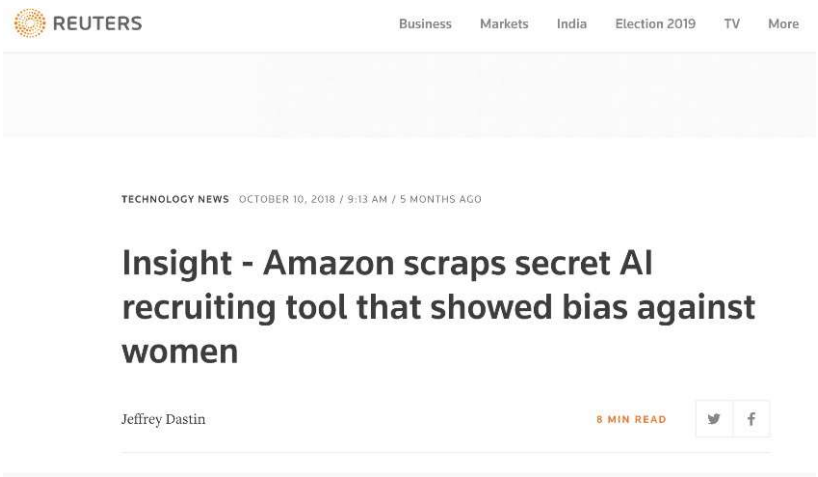
Course Plan

M1	Introduction to AI
M2	Problem Solving Agent using Search
M3	Game Playing
M4	Knowledge Representation using Logics
M5	Probabilistic Representation and Reasoning
M6	Reasoning over time, Reinforcement Learning
M7	Ethics in AI

Ethics in Artificial Intelligence


Performance standard
Critic
Sensors
feedback
changes
Learning element
Performance element
knowledge
learning goals
Problem generator
Actuators
Agent
Environment

Recommendation System



Amazon's Edinburgh engineering hub's goal was to develop AI that could rapidly crawl the web and spot candidates worth recruiting

Fairness : The absence of bias towards an individual or a group

Are the predictions _____?

- Fair
- Unbiased

Object Recognition System

-

Are the Inferences_____?

- Correct
- Unbiased

Are the Predictions _____?

- Fair
- Universally Applicable

Building a Fair Model

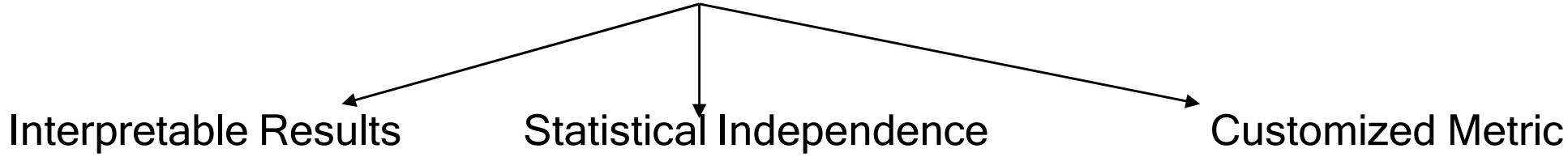
No artificial model is a perfect one. But every model significantly influence the social , economic, cultural ethics impacting humanity.

Justify the design modelled & metric used to validate the model, is in fact the right choices fit in the context.

1. Is it fair to make an AI-ML system?
2. Is there a better technical approach to convert an existing AI system fair?
3. Are the results obtained by the AI system fair?

Building a Fair Model

1. Is it fair to make an AI-ML system?
2. Is there a better technical approach to convert an existing AI system fair?



Interpretable Models

Are the results obtained by the AI system fair?

Interpretable models helps to trust the AI system by answering transparently to the specific questions like “Why the system is behaving under certain scenarios?”

- If a loan gets rejected, do we know the reasons?
- If a job application is accepted, is it biased towards a gender?
- If a bail is granted to an accused, is it based on their race?
- If a patient is diagnosed with a disease, what factors made the algorithm to classify it?

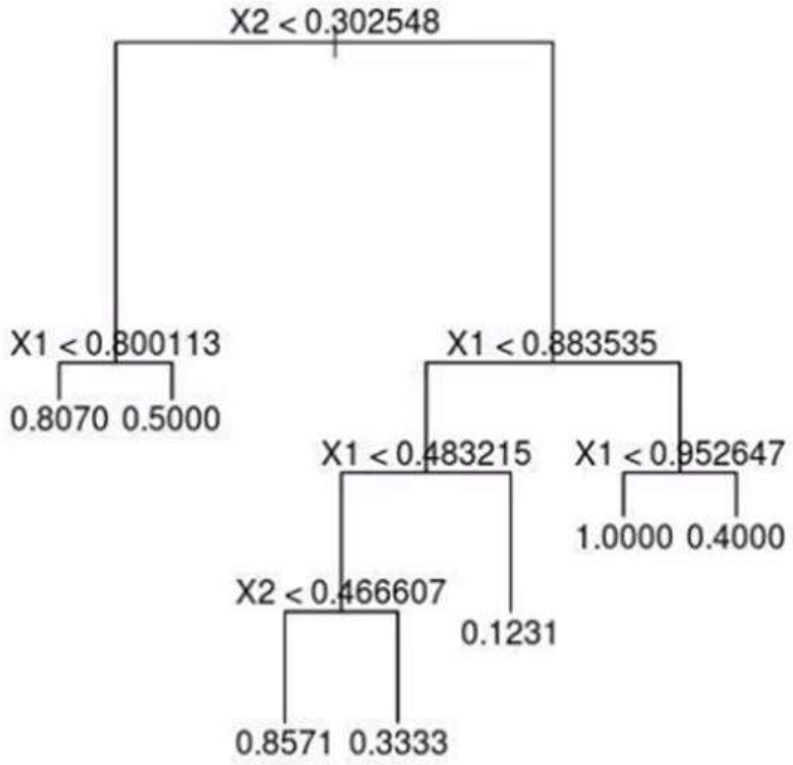
Interpretable Models

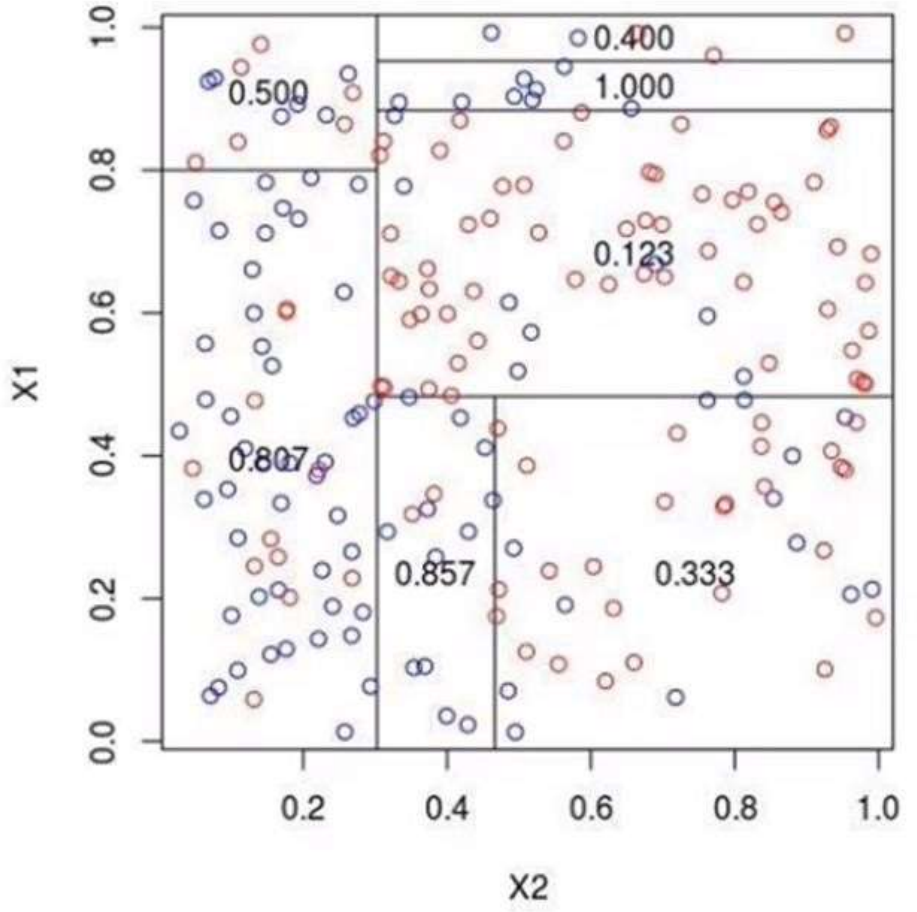
Example Based Explanations:
If SymptomInX ≡ SymptomInY
if DiseaseA infected X
then probably DiseaseA might have infected Y

If CustomerX ≡ CustomerY
if CustomerX purchased P1
then probably CustomerY will purchase P1

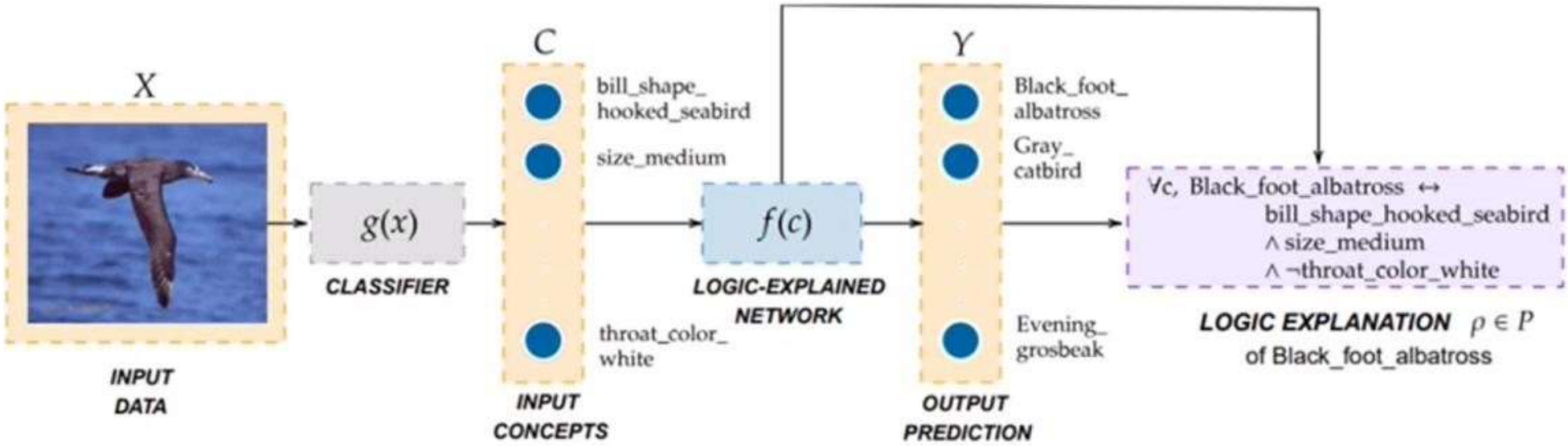
Counterfactual Explanations:
If customerX's income level had not been less than L3
then the customer's Loan might not have been rejected

Interpretable Models





In Deep Learning



Source Credit: 2021: **Logic Explained Networks**

Required Reading: Web Resources

Thank You for all your Attention

Note : Some of the slides are adopted from AIMA TB materials

Sample Problems

Logics

"In the marketing industry, all advertised products gain popularity. Not all profitable products have been popular, but all the popular products have been always profitable. Profitable products attract investments from corporates. "

1. Represent the knowledge base using propositional logic (without quantifiers)
2. convert KB into CNF and find any three sample complete BSAT (Binary Satisfiability) solutions to the variables using DPLL algorithm.
3. Using the result of part a), prove the below using equivalence laws and resolution.

"All the advertised products are invested by the corporates"

Logics

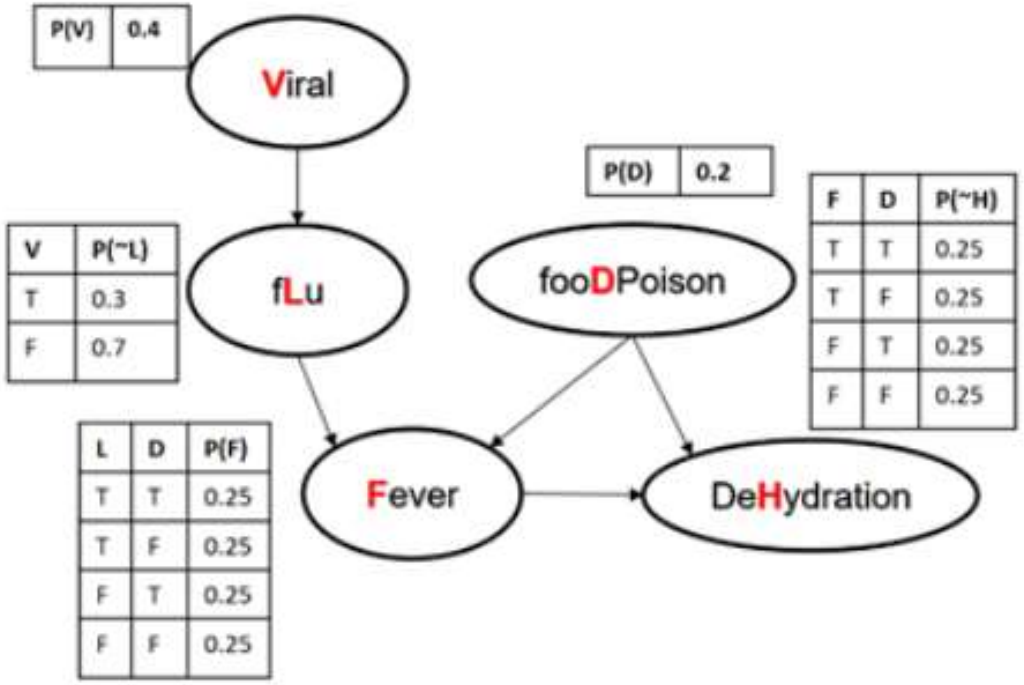
"A software intern is always trained in a project as a project member. Every Project member is involved in development team and support team. Every development member is involved in unit testing. Every support member is involved in user acceptance testing. A project member trained in both unit testing and acceptance testing is certified as skilled in testing."

Convert the above into predicate logic.

Prove by backward chaining that "*Prove that all the software interns will be certified as skilled in testing.*" using the results of part a. Show the steps by step inferences using neat diagram with direction.

Bayesian Network

Consider the below Bayesian Network and answer the following questions:

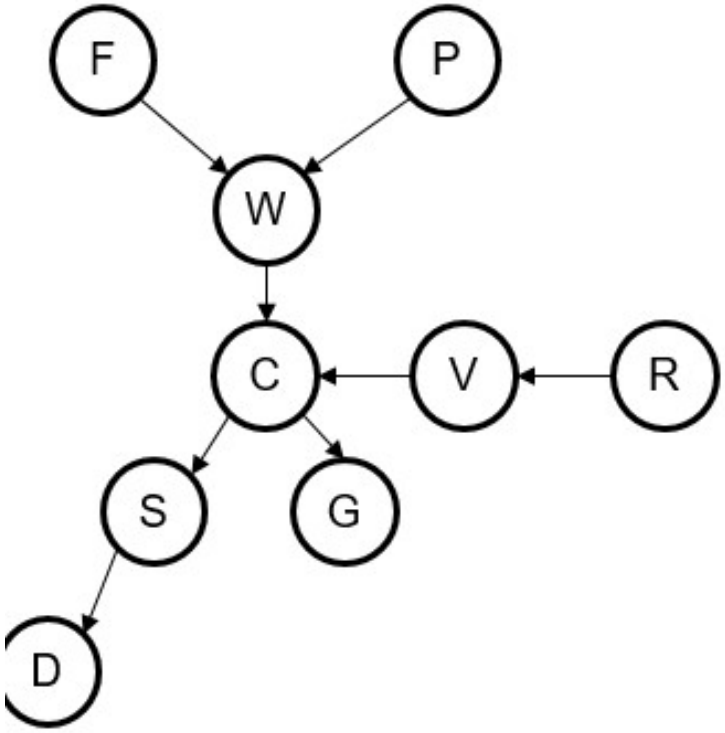


1. **Exact Inference : What is the chance that a person doesn't get fever given the evidence that his/her blood test results show viral infection and severe dehydration?**

2. Approximate inference – Prior Sampling, Rejection sampling, likelihood weighing, "0.3, 0.6, 0.2, 0.1, 0.7, 0.5, 0.5, 0.25, 0.45, 0.85, 0.35, 0.9, 0.15, 0.65, 0.51, 0.2, 0.7, 0.10, 0.6, 0.8"

Bayesian Network

Most of the WILP students are fans (F) of cricket irrespective of their gender. With the new season of IPL (Indian Premier League) having started on the exam month almost every cricket fans spend time to watch(W) the live play. Sometimes being a parent (P) reduces the probability of watching the IPL live season. A likely consequence of watching matches is reduced concentration(C) on the following day/s. A consequence of the reduced concentration is increased stress(S) with work environment leading to reduced productivity (D) in project. Lack of concentration might also be caused by viral (V) infection, which is common in this rainy season(R). WILP students have the comprehensive exams and reduced concentration would reduce the probability of good grades (G) in the exam which reflects the performance of students in examination. Assume an AI agent is fed this information and it answers to certain queries that can be inferred. Assume all the events(conditional or unconditional) are equally likely to occur:



Example Joint Prob.Distribution Query :

What is the chance that “an ardent fan of cricket who is a parent of two kids, never misses an IPL match, doesn’t get stressed in work environment, is affected by viral infection and performs well in the comprehensive examination”?

D-Seperation: Performance of in the examination is independent of stress in work environment given its known that the student is affected by viral infection