

EE2213 Introduction to AI

Lecture 12: Introduction to Knowledge Representation and Reasoning

Dr. WANG Si

si.wang@nus.edu.sg

Electrical and Computer Engineering Department
National University of Singapore

OVERVIEW OF COURSE CONTENTS

- **Introduction (Shaojing)**

- What is AI
- Applications of AI
- AI agent

- **Search (Shaojing)**

- Uninformed search algorithms: breadth-first, depth-first, uniform-cost(Dijkstra's algorithm)
- Informed search algorithms: greedy best-first, A*

- Applications

- **Optimisation (Shaojing)**

- Linear programming
- Convex problems
- Applications

- **Machine learning (Wang Si) (Weeks 6-9)**

- Supervised and unsupervised learning: regression, classification, clustering
- Neural networks and deep learning
- Applications

- **Knowledge representation (Wang Si) (Weeks 10-11)**

- Knowledge Representation and Reasoning
- Propositional Logic
- Applications

- **Ethical considerations (Shaojing)**

- Bias in AI
- Privacy concerns
- Societal impact

AGENDA

- We will discuss:
 - Definition of Knowledge Representation and Reasoning (KRR)
 - Types of Knowledge Representation
 - Types of Reasoning
 - A Case Study: The Wumpus World

At the end of this lecture, you should be able to:

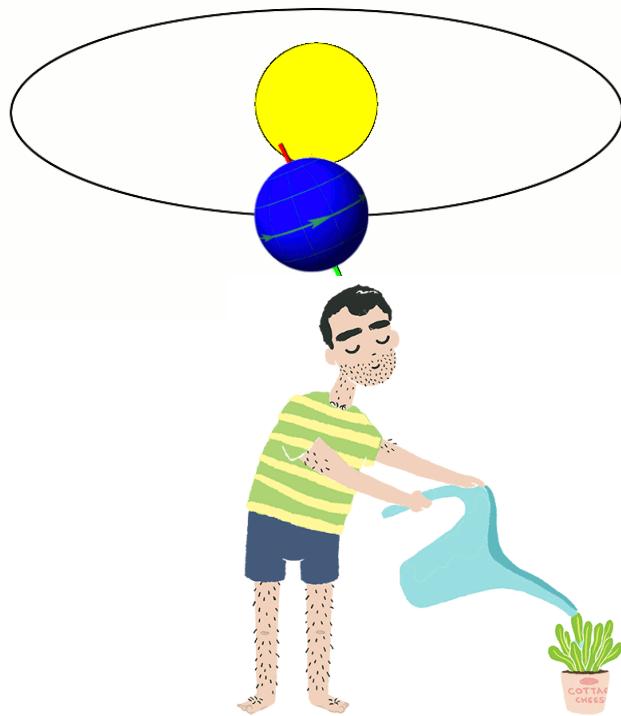
- ✓ Describe the role of KRR in AI agent
- ✓ Differentiate among types of knowledge, types of knowledge representation, and types of reasoning
- ✓ Understand Knowledge-based Agent

Knowledge

➤ What is Knowledge?

An awareness of facts, a familiarity with individuals and situations, or a practical skill

----by Wikipedia



What if we want machines to know things?

Knowledge Representation and Reasoning (KRR)

➤ What is KRR?

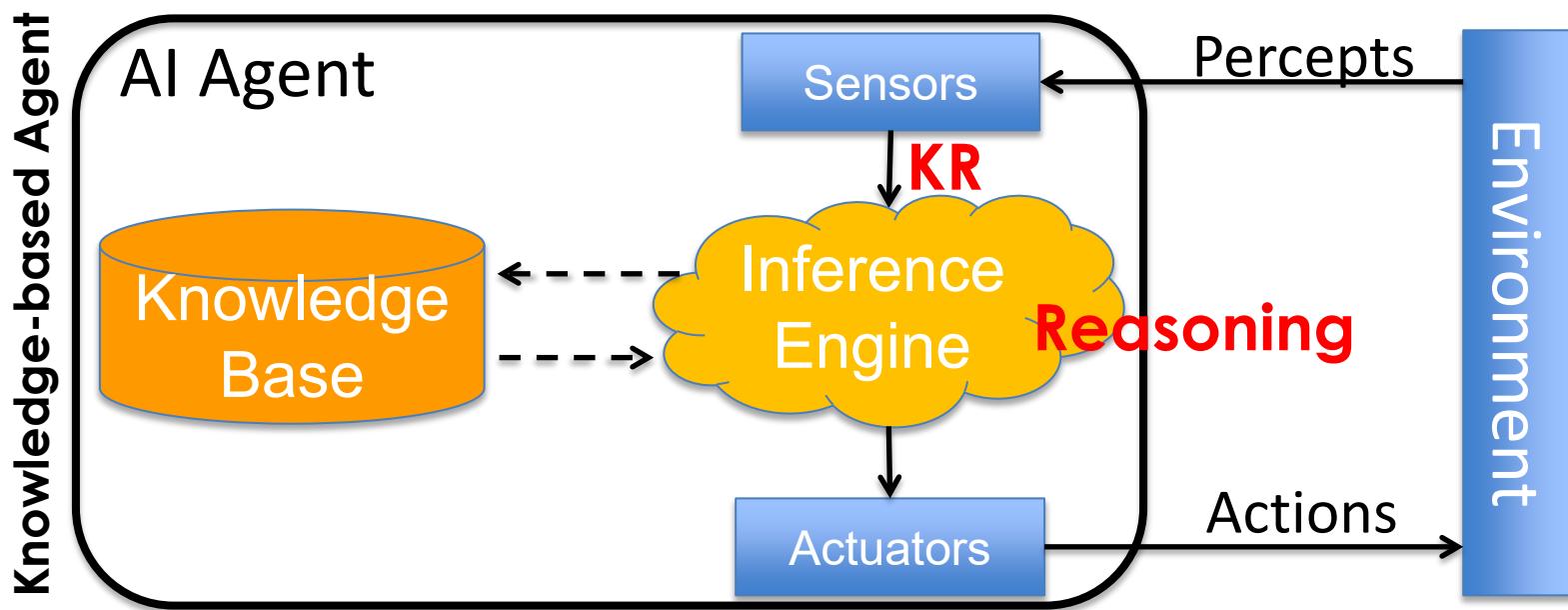
A field in AI concerned with how knowledge can be **formally** represented and used by machines to reason and make decisions

Part1: KR

How we store and
organize knowledge

Part2: Reasoning

how to draw conclusions or
make decisions from it

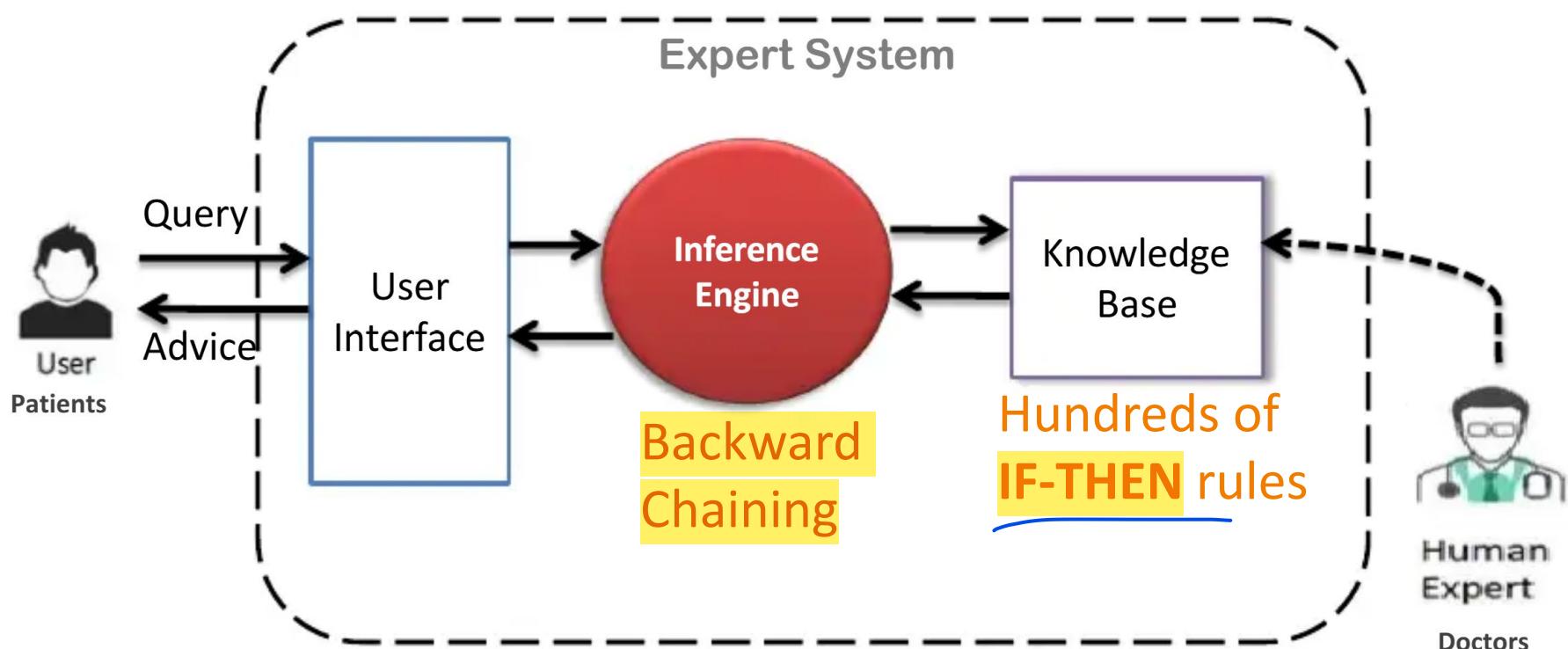


Knowledge Representation and Reasoning (KRR)

➤ KRR in AI Agent

Example 1: MYCIN Expert System (KR: IF-THEN rules)

(developed in the 1970s to diagnose bacterial infections and recommend antibiotics)

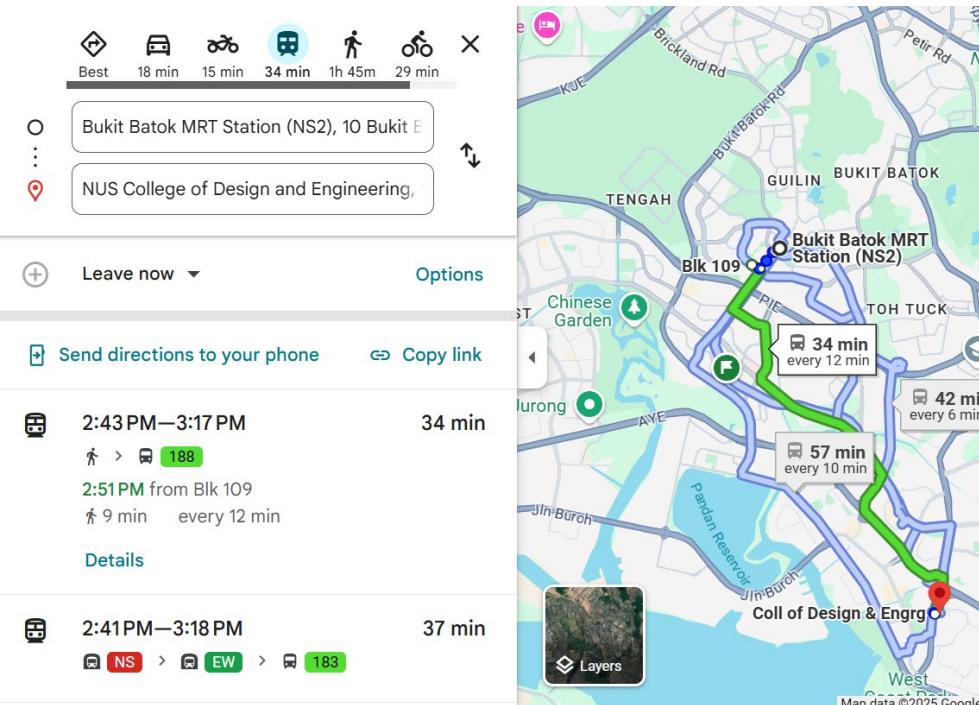


e.g., *IF* the patient has a fever AND high white blood cell count, *THEN* there is a high chance of bacterial infection.

Knowledge Representation and Reasoning (KRR)

➤ KRR in AI Agent

Example 2: Google Maps' Route Recommendation



KR:

- Road Network as a **Graph**
 - ✓ nodes = intersections;
 - ✓ edges = roads;
 - ✓ Edge attributes = distance, speed limit, average traffic, road type, toll info, etc.
- Historical and Real-time traffic Data
 - ✓ Real-time traffic info as **dynamic weights on edges**.
 - ✓ Historical traffic pattern as **time-stamped edge attributes/probabilistic model/lookup tables**

⋮

Reasoning:

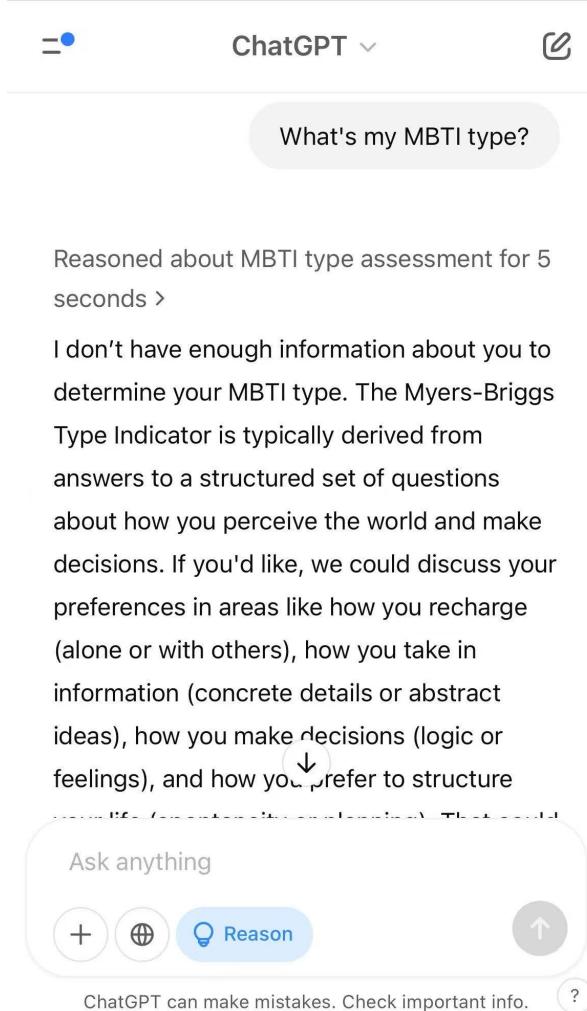
Graph search algorithms like **Dijkstra's or A***
Dynamic adaptation for real-time traffic

Predictive reasoning to predict future traffic condition

⋮

➤ KRR in AI Agent

Example 3: ChatGPT



ChatGPT

What's my MBTI type?

Reasoned about MBTI type assessment for 5 seconds >

I don't have enough information about you to determine your MBTI type. The Myers-Briggs Type Indicator is typically derived from answers to a structured set of questions about how you perceive the world and make decisions. If you'd like, we could discuss your preferences in areas like how you recharge (alone or with others), how you take in information (concrete details or abstract ideas), how you make decisions (logic or feelings), and how you refer to structure

Ask anything

+ @@ Reason

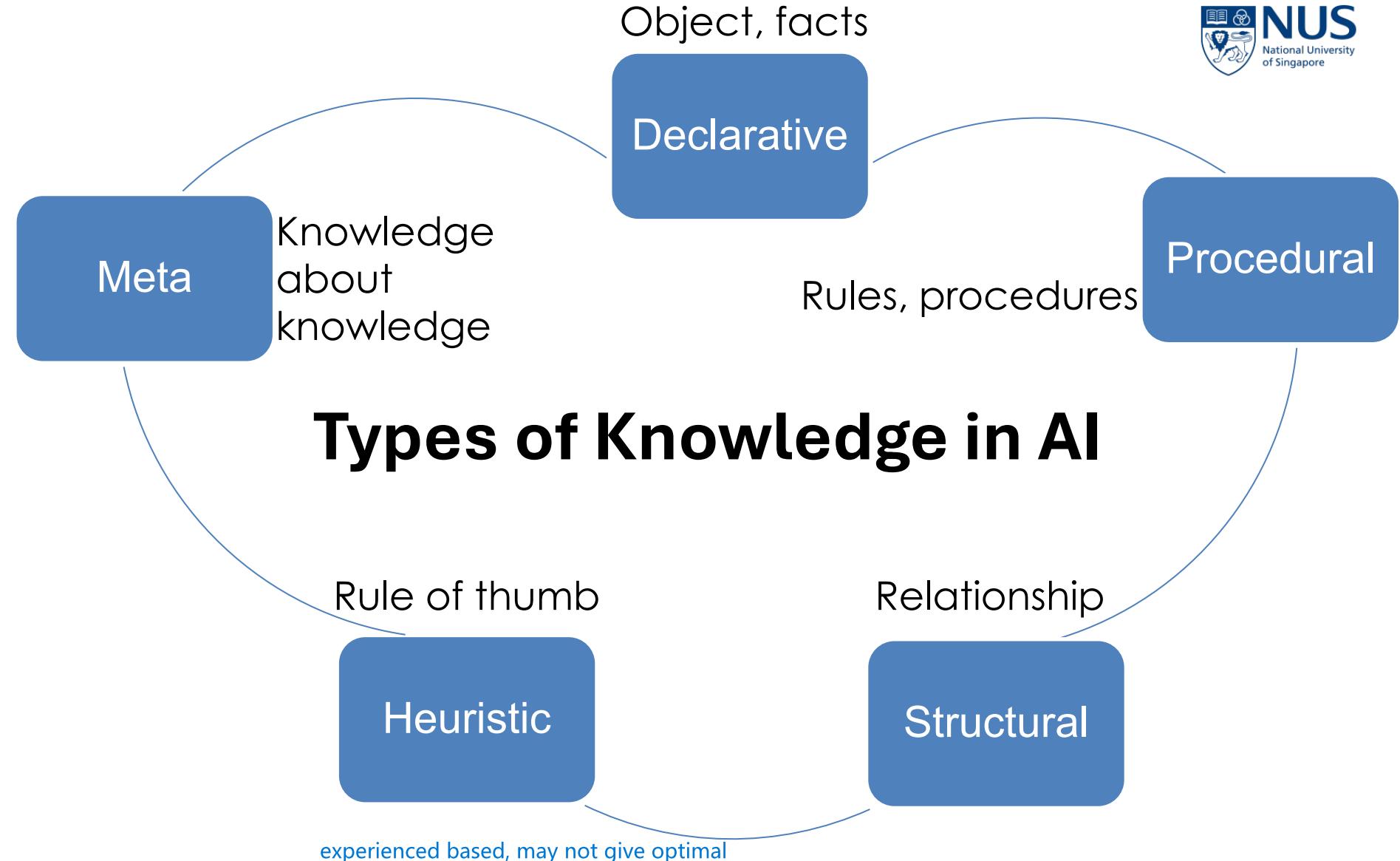
ChatGPT can make mistakes. Check important info.

KR: Vector-based representation

(facts, concepts, and language patterns are encoded in the weights of the neural networks)

Reasoning: Predictive reasoning

“By predicting the most likely next token in a sequence, conditioned on context”



Types of Knowledge in AI

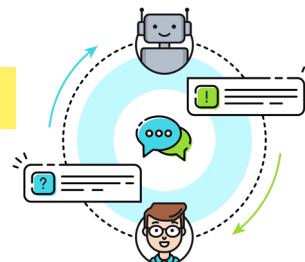
➤ Declarative Knowledge

Facts and statements about the world

("Knowing that" something is true)

Applications:

1. Form the knowledge base of knowledge-based agent (e.g., expert system)



2. Allow chatbots to answer factual questions.

3. Specify the world state and constraints for robot navigation system.

"Room A is connected to Room B via Door 1"

"A corridor leads from the entrance to the lab"

"Do not enter hazardous zones"

"Keep 1m away from walls"

Examples

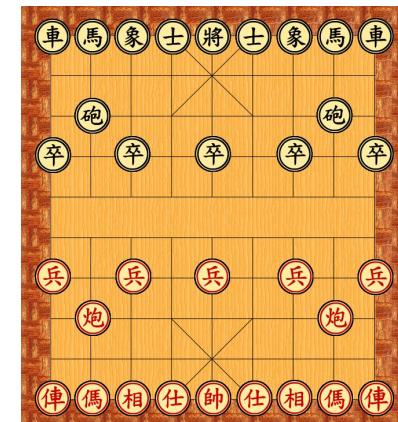
"Beijing is the capital of China"

"Water freezes at 0°C"

"Dogs are mammals"

"A triangle has 3 sides"

4. Define rules and facts of the game for the game AI



Types of Knowledge in AI

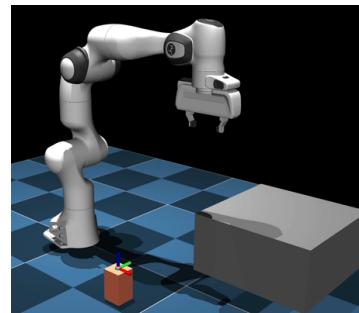
➤ Procedural Knowledge

Rules, strategies, and sequence of actions that can be directly applied to tasks.

(“Knowing how” to do something)

Applications:

1. Motion control for robotics



2. Problem-Solving

3. Guide parsing and generation of the sentence in Natural Language Processing

4. Behavior scripts for NPCs in Game AI

5. Machine Learning e.g. how to back prop in MLP

“Steps to update weights in neural networks.”

Examples

Steps to solve a Rubik’s Cube

How to sort a list

How to grasp a cup of water

How to navigate from A to B



Types of Knowledge in AI

➤ Structural Knowledge

Relationships and hierarchies between concepts.

Examples

"A robin is a bird, a bird is an animal, so a robin is an animal."

"Fuel and oxygen are required for fire; fire produces smoke; smoke can trigger alarms."

"A car is a type of vehicle, all vehicles have wheels, therefore a car has wheels."

"A day has three main meals: breakfast, lunch, and dinner; breakfast comes before lunch, and lunch comes before dinner."

Applications:

1. Recommender System

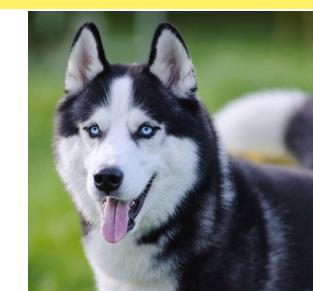
"Action movies → Superhero → Marvel"

2. Allow chatbots to answer indirect questions

"Alice is Bob's mother." "Bob is Charlie's father."

"Who is Charlie's grandmother?"

3. Hierarchical Classification



also know it
is dog, a
class of
higher
hierarchy

Types of Knowledge in AI

➤ Heuristic Knowledge

Rules of thumb or experience-based shortcuts

(not guaranteed to be correct/optimal, but useful/fast)

Examples

“Explore the most promising node first in A*.”

“If a door is locked, try the keys in order of most frequently used keys first.”

“Choose the restaurant with the longest queue.”

“Carry an umbrella if the sky looks cloudy.”

Applications:

1. Recommender System

“Recommend popular items first”

2. Robot Arm Planning

“prefer moves that minimize distance to the target object”

3. Pathfinding

Types of Knowledge in AI



➤ Meta-Knowledge

Know how to use, organize, select, or reason with knowledge

(Knowing about knowledge)

Examples

“Use supervised learning if labeled data is available; otherwise, use unsupervised learning”

“If the GPS signal is weak, I’ll follow road signs instead.” when to rely on GPS

“When I am tired, I’ll do easier tasks first.”

“If my first approach to a puzzle isn’t working, I’ll try a different strategy.” know which method is likely to succeed

Types of Knowledge in AI



■ Practice Questions: Identify type of knowledge

“If it looks like a duck and quacks like a duck, assume it’s a duck”

heuristic knowledge

“To make coffee: add grounds, pour water, heat.”

procedural knowledge

“Birds can fly” declarative

“Penguins are birds, and birds are animals.” structural (emphasise on relationship)

“Penguins are birds.” declarative (emphasise more on fact than relationship)

“This system works poorly when data is noisy.”

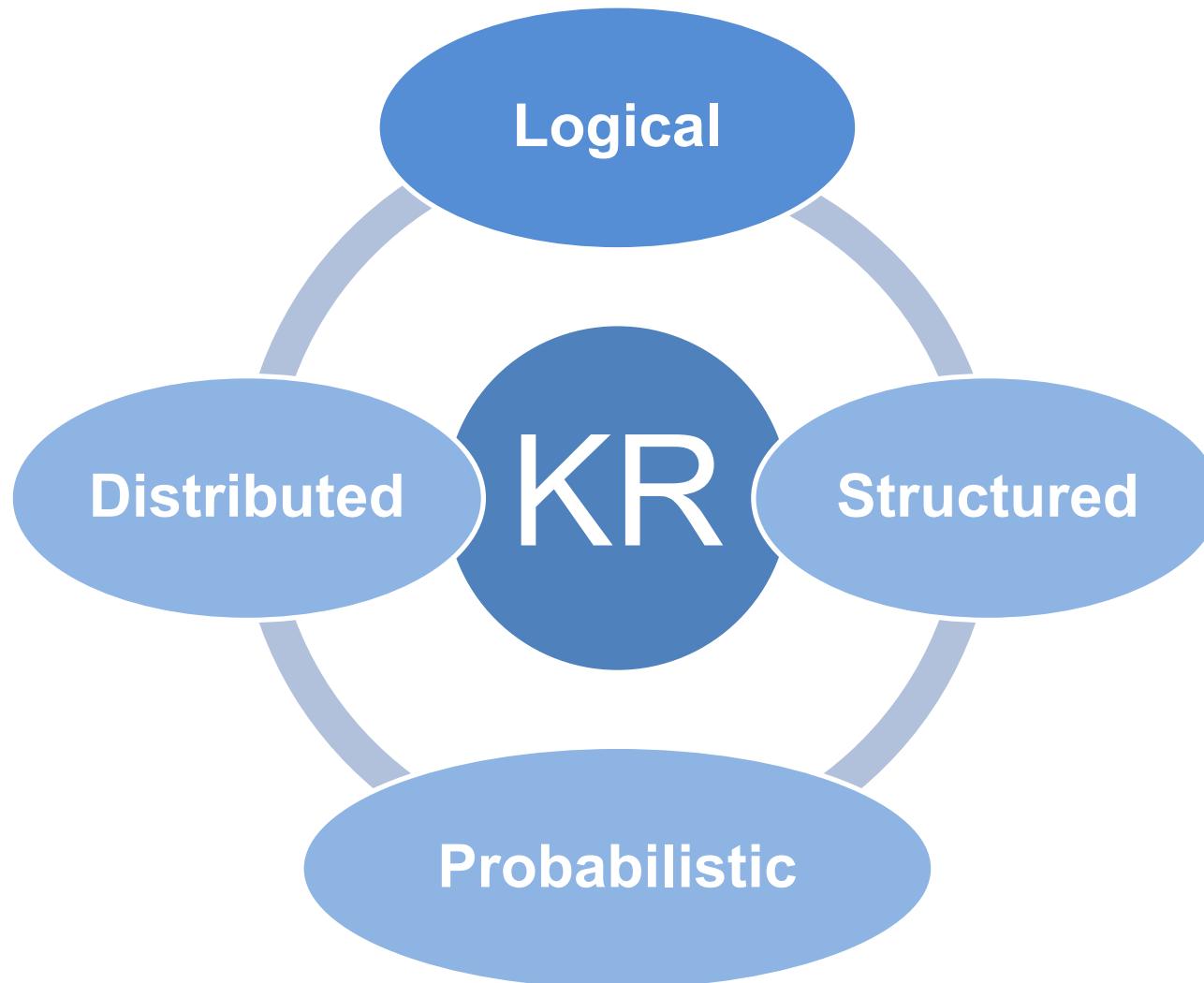
meta (performance)

Knowledge Representation

- Encode information into a format that computers can understand and use to solve problems

Knowledge Representation

➤ Core Methods



Knowledge Representation

➤ Logical KR

Represent knowledge using formal logic---clear rules or statements that are either **true** or **false**.

Techniques: Propositional logic, First-order logic

declarative statements (propositions)
linked by logical operators like AND, OR,
NOT, etc.

e.g., “IF it rains AND the ground is wet,
THEN the road is slippery”



- A: “It rains”
 - B: “The ground is wet”
 - C: “The road is slippery”
- $$A \wedge B \Rightarrow C$$

Extends propositional logic by
introducing variables, quantifiers, and
predicates.
how many objects in the domain
property of objects/relations

e.g., “All humans are mortal”

$\forall x$ Human(x) \Rightarrow Mortal(x)

x : variable
Human(x): x is a human
Mortal(x): x is mortal



Knowledge Representation

➤ Structured KR

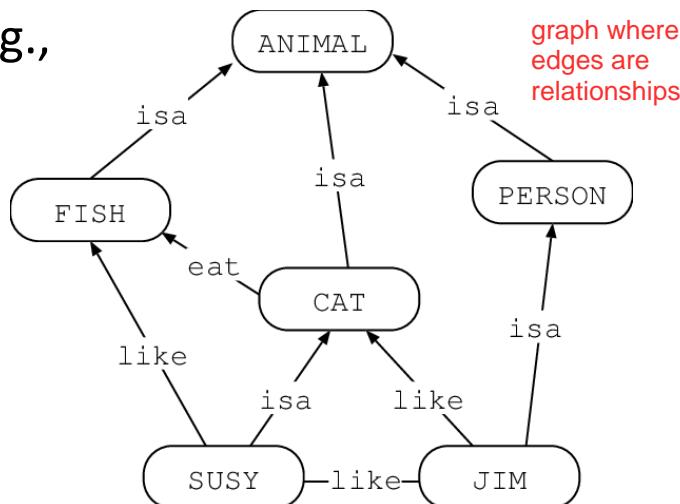
Organize knowledge into structured formats like hierarchies, graphs, or frames.

Techniques: Semantic Networks, Frames, etc.

Nodes = Concepts

Edges = Relationships

e.g.,



Structured object with slots & values

e.g.,

key,value pair

University	
Name	NUS
Programs	B.E, M.E, Ph.D
Department	ECE, CEE

slots

values

Knowledge Representation

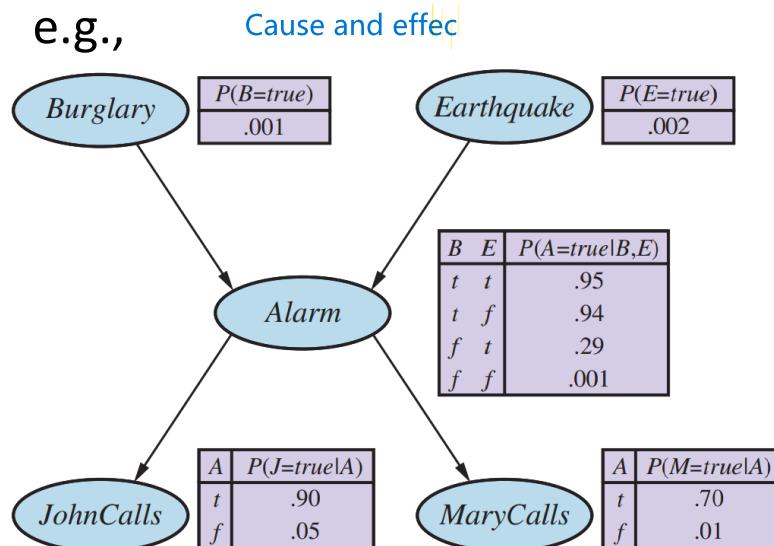
➤ Probabilistic KR

Represent knowledge where uncertainty exists by assigning probabilities to outcomes.

Techniques: Bayesian Networks, Markov Networks, etc.

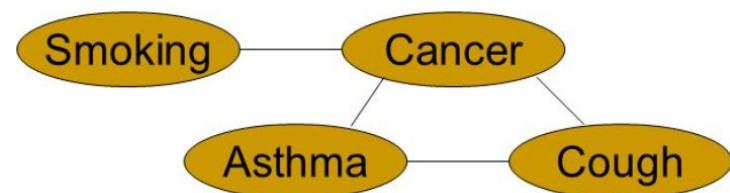
Model causal relationships among variables (directed graph)

Nodes = Variables
Edges = Conditional dependencies
e.g.,



Model correlations among variables (undirected graph)

Nodes = Variables
Edges = Direct probabilistic influence
e.g.,



Knowledge Representation



➤ Distributed KR

Represent knowledge as vectors in a high-dimensional space, often learned from data (not hand-coded)

Techniques: Word embeddings, neural networks

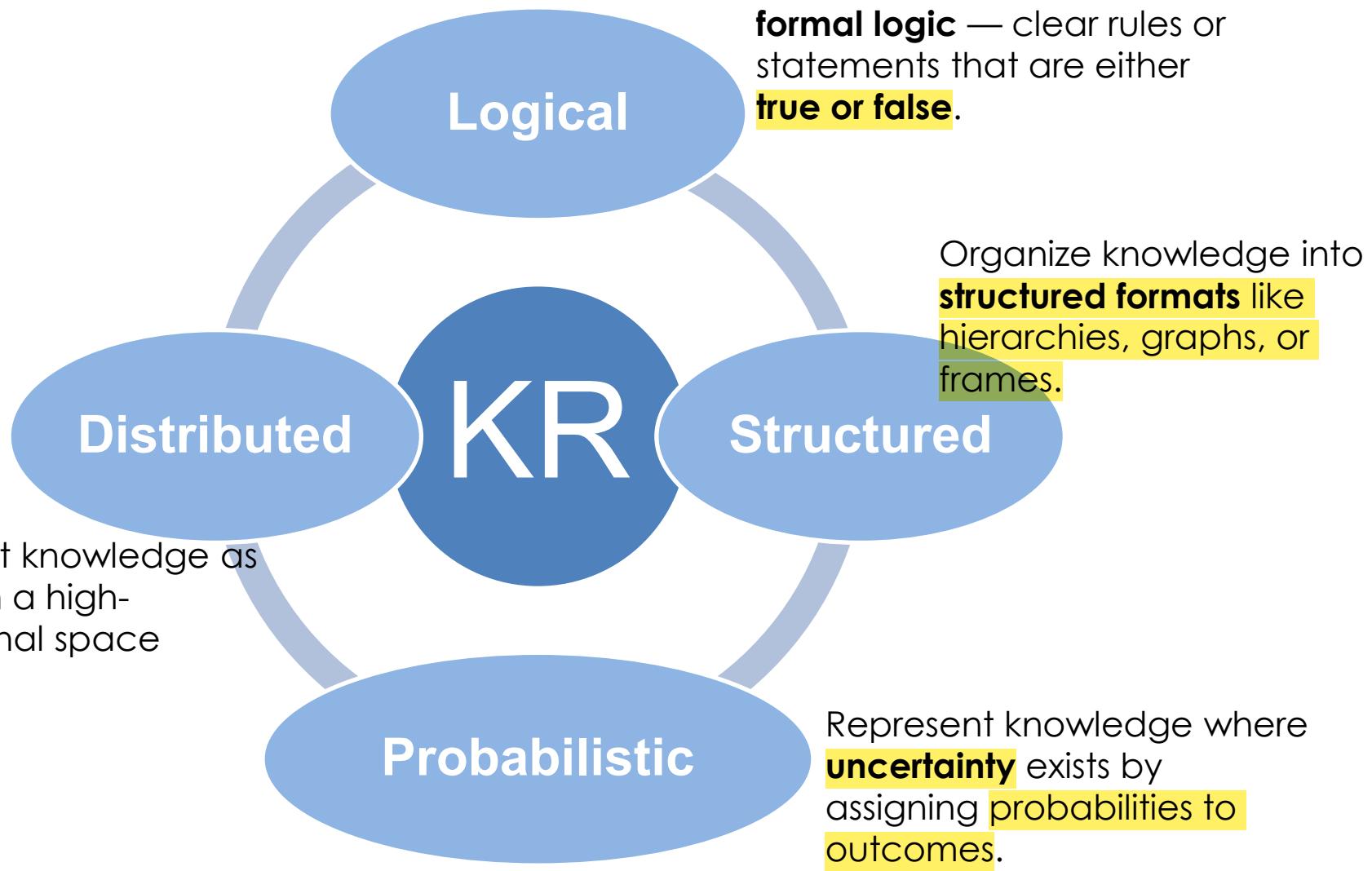
e.g., Word2Vec map synonyms to nearby vectors

Word	Vector (Simplified)
king	[0.52, 0.71, -0.33, 0.44, 0.91]
queen	[0.51, 0.69, -0.31, 0.46, 0.93]
man	[0.60, 0.80, -0.40, 0.38, 0.90]
woman	[0.49, 0.70, -0.29, 0.47, 0.92]

$$\begin{aligned} & \text{king} - \text{man} + \text{woman} \\ & \approx \text{queen} \end{aligned}$$

Knowledge Representation

➤ Core Methods



Knowledge Representation

- Practice Questions: Identify the type of knowledge representation

Imagine a Self-driving Car:

- It uses a neural network to detect road signs from images.
- It estimate probabilities like “There is an 85% chance that the object ahead is a pedestrian.”
probabilitisitc
- It uses rules like “If the light is red, stop.”
logical
- It stores info about itself like:

Vehicle	
Manufacturer	Tesla
Model	Model 3
Sensors	Cameras, LIDAR, Radar, Ultrasonic

structured

Reasoning

- Draw conclusions, make inference, and solve problems based on the existing knowledge

Reasoning

➤ 3 Main Types of Reasoning

Deductive

General Rule

Specific Conclusion

Inductive

Specific Observation

General Conclusion

Abductive

Incomplete Observation

Best Explanation

e.g.,

All humans are mortal

Wang Si is a human



Wang Si is mortal

Swan 1 is white;
Swan 2 is white...



All swans are
probably white

The grass is wet



did not see the
rain

It probably rained

Reasoning

➤ 3 Main Types of Reasoning

Deductive

General Rule



Specific Conclusion



Logical KR

Inductive

Specific Observation



General Conclusion



Probabilistic KR

Distributed KR

↳ training / ML

Abductive

Incomplete Observation



Best Explanation



Probabilistic KR

↳ guess

Reasoning

■ Practice Questions: Identify the type of reasoning

- A student consistently scores well in the previous exams. The teacher predicts that the student will also score well in the next exam
 - inductive
 - takes specific observations of doing well and generalising that the student will do well every time
- A mechanic finds that a car won't start. Possible causes: dead battery, empty fuel tank, faulty ignition. He hypothesizes it's most likely a dead battery
 - abductive (finds most likely explanation)
- All even numbers are divisible by 2. 8 is an even number. Therefore, 8 is divisible by 2.
 - deductive

function KB-AGENT(*percept*) **returns** an *action*
persistent: KB, a knowledge base
 t, a counter, initially 0, indicating time

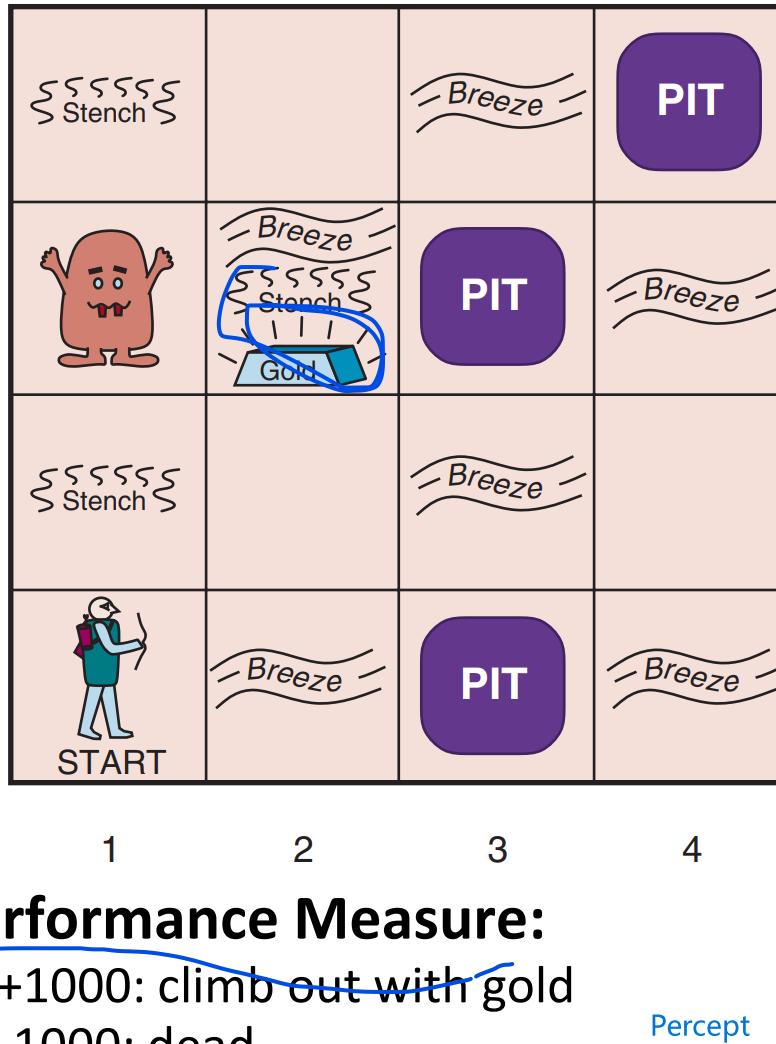


TELL(KB, MAKE-PERCEPT-SENTENCE(*percept*, *t*))
action \leftarrow ASK(KB, MAKE-ACTION-QUERY(*t*))
TELL(KB, MAKE-ACTION-SENTENCE(*action*, *t*))
t \leftarrow *t* + 1
return *action*

Knowledge-based Agent:

Case Study

The Wumpus World



Performance Measure:

- +1000: climb out with gold
- 1000: dead
- 1: each action taken
- 10: use up the only arrow

Environment:

- 4x4 grid of rooms
- The agent starts in [1,1], facing rightward.
- Some rooms contain:
 - A Wumpus —kill the agent if entered.
 - Pits — deadly if fallen into.
 - Gold — the agent wants

Actuators:

- Move, grab gold, shoot an arrow, and climb out.

Sensors:

- Stench (adjacent to Wumpus),
- Breeze (adjacent to PIT),
- Glitter (Gold),
- Bump (Wall),
- Scream (Wumpus is killed)

The Wumpus World

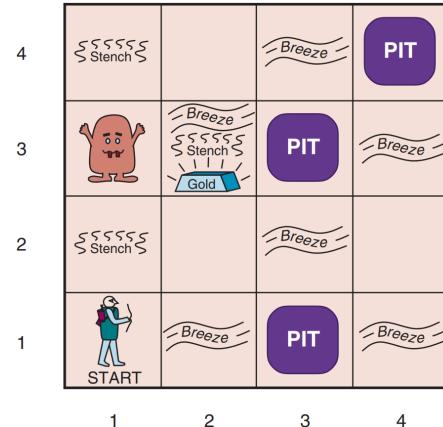
- Environment Exploration by a Knowledge-based Wumpus agent

Initial situation

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2	2,2	3,2	4,2
OK			
1,1 A OK	2,1 OK	3,1	4,1

UPDATE

A = Agent
B = Breeze
G = Glitter, Gold
OK = Safe square
P = Pit
S = Stench
V = Visited
W = Wumpus



At first

KB: rules of the world;
the agent is in [1,1]; [1,1] is OK

Percept: [None, None, None, None, None]

Inference: [1,2] and [2,1] are OK

Action: Move to [2,1]

The Wumpus World

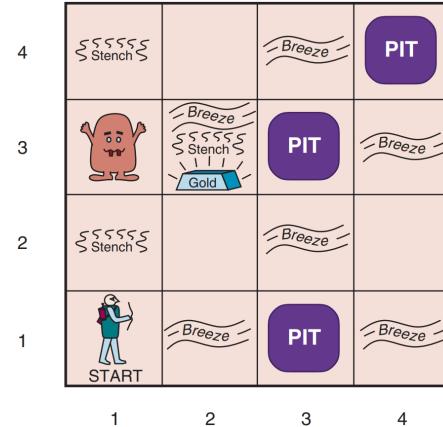
- Environment Exploration by a Knowledge-based Wumpus agent

After 1st Move

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2	2,2 <i>P?</i>	3,2	4,2
OK			
1,1	2,1 A B OK	3,1 <i>P?</i>	4,1
V OK			

Action: Move back to [1,1], and then proceed to [1, 2]

- | | |
|-----------|-----------------|
| A | = Agent |
| B | = Breeze |
| G | = Glitter, Gold |
| OK | = Safe square |
| P | = Pit |
| S | = Stench |
| V | = Visited |
| W | = Wumpus |



KB: rules of the world;
the agent is in [2,1];
[1,1], [1,2], [2,1] are OK

Percept: [None, Breeze, None, None, None]

Inference: A Pit in [2,2] or [3,1] or both,
No Wumpus in [2,2] and [3,1]

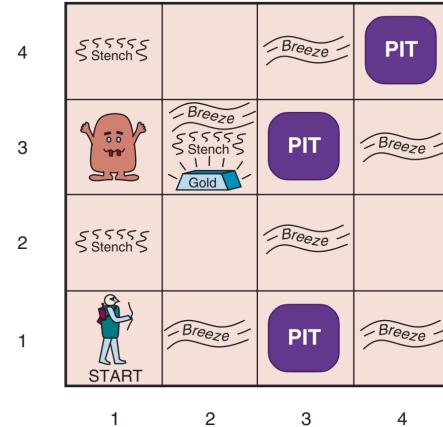
The Wumpus World

- Environment Exploration by a Knowledge-based Wumpus agent

After 3rd Move

1,4	2,4	3,4	4,4
1,3 W!	2,3	3,3	4,3
1,2 A S OK	2,2 OK	3,2	4,2
1,1 V OK	2,1 B V OK	3,1 P!	4,1

A = Agent
B = Breeze
G = Glitter, Gold
OK = Safe square
P = Pit
S = Stench
V = Visited
W = Wumpus



KB: rules of the world;
the agent is in [1,2];
[1,1], [1,2], [2,1] are OK;
[2,2] and [3,1] are P? But no Wumpus.

Percept: [Stench, None, None, None, None]

Inference: Wumpus in [1,3];
[2,2] is OK;
A Pit in [3,1]

Action: Move to [2, 2]

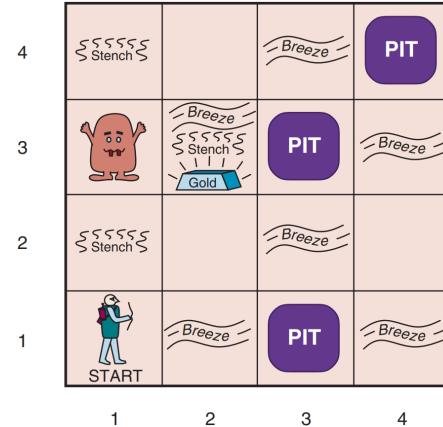
The Wumpus World

- Environment Exploration by a Knowledge-based Wumpus agent

After 4th Move

1,4	2,4	3,4	4,4
1,3 W!	2,3 OK	3,3	4,3
1,2 V S OK	2,2 A OK	3,2 OK	4,2
1,1 V OK	2,1 B V OK	3,1 P!	4,1

A = Agent
B = Breeze
G = Glitter, Gold
OK = Safe square
P = Pit
S = Stench
V = Visited
W = Wumpus



KB: rules of the world;
 the agent is in [2,2];
 [1,1], [1,2], [2,1], [2,2] are OK;
 ↙ A Pit in [3,1]; A Wumpus in [1,3]

Percept: [None, None, None, None, None]
Inference: [2,3] and [3,2] are OK.

Action: Move to [2, 3]

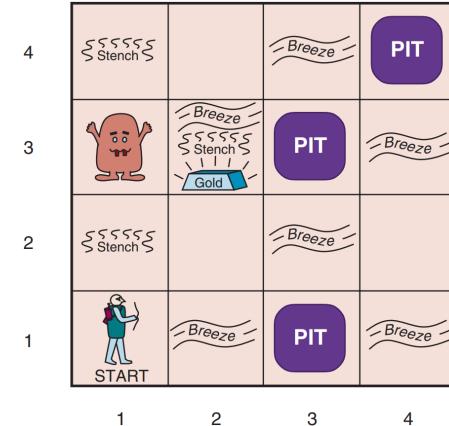
The Wumpus World

- Environment Exploration by a Knowledge-based Wumpus agent

After 5th Move

1,4	2,4 P?	3,4	4,4
1,3 W!	2,3 A S B G OK	3,3 P?	4,3
1,2 V S OK	2,2 V OK	3,2 OK	4,2
1,1 V OK	2,1 B V OK	3,1 P!	4,1

A = Agent
B = Breeze
G = Glitter, Gold
OK = Safe square
P = Pit
S = Stench
V = Visited
W = Wumpus



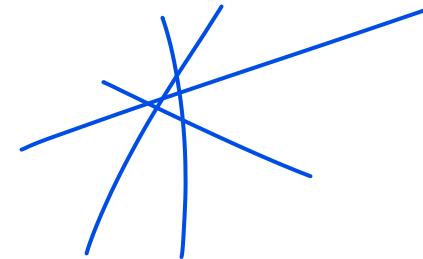
KB: rules of the world;
 the agent is in [2,3];
 [1,1], [1,2], [2,1], [2,2], [2,3], [3,2] are OK;
 A Pit in [3,1]; A Wumpus in [1,3]

Percept: [Stench, Breeze, Glitter, None, None]

Inference: [2,3] has gold;
 A Pit in [2,4] or [3,3] or both

Action: Grab gold, and move back to [1, 1], and climb out

Summary



- ✓ 5 types of Knowledge
 - Declarative, Procedural, Structural, Heuristic, Meta
- ✓ 4 types of Knowledge Representation
 - Logical, Structured, Probabilistic, Distributed
- ✓ 3 types of Reasoning
 - Deductive, Inductive, Abductive
- ✓ 1 Case Study of Knowledge-based Agent:
The Wumpus World

THE END

EE2213 Introduction to AI

Lecture 13: Propositional Logic

Dr. WANG Si

si.wang@nus.edu.sg

Electrical and Computer Engineering Department
National University of Singapore

OVERVIEW OF COURSE CONTENTS

- **Introduction (Shaojing)**

- What is AI
- Applications of AI
- AI agent

- **Search (Shaojing)**

- Uninformed search algorithms: breadth-first, depth-first, uniform-cost(Dijkstra's algorithm)
- Informed search algorithms: greedy best-first, A*

- Applications

- **Optimisation (Shaojing)**

- Linear programming
- Convex problems
- Applications

- **Machine learning (Wang Si) (Weeks 6-9)**

- Supervised and unsupervised learning: regression, classification, clustering
- Neural networks and deep learning
- Applications

- **Knowledge representation (Wang Si) (Weeks 10-11)**

- Knowledge Representation and Reasoning
- Propositional Logic
- Applications

- **Ethical considerations (Shaojing)**

- Bias in AI
- Privacy concerns
- Societal impact

AGENDA

- We will discuss:
 - Syntax of Propositional Logic
 - Semantics of Propositional Logic
 - Inference for Propositional Logic

At the end of this lecture, you should be able to:

- ✓ Explain the syntax and semantics of propositional logic
- ✓ Apply discussed inference methods to determine logical entailment in propositional logic

Propositional Logic:

Introduction

Propositional Logic: Introduction



➤ What is Propositional Logic?

A branch of logic where:

- Statements (propositions) are either **true** or **false**, but **not both**
e.g., P= “It is raining”, Q= “I have an umbrella”
- These statements can be combined using **logical connectives** (like AND, OR, NOT, etc.)
e.g., P AND Q= “It is raining AND I have an umbrella”
- The **truth value of complex statements** can be determined from the **truth values of their components**

Propositional Logic: Introduction



➤ Importance of Propositional Logic

- ✓ Foundation for Formal Reasoning

A **formal, mathematical system** to express knowledge in a clear, unambiguous way.

- ✓ Building Block for Complex Logics

Basis for more expressive logical systems, such as first-order logic, description logic

- ✓ Enables **Automated Inference**

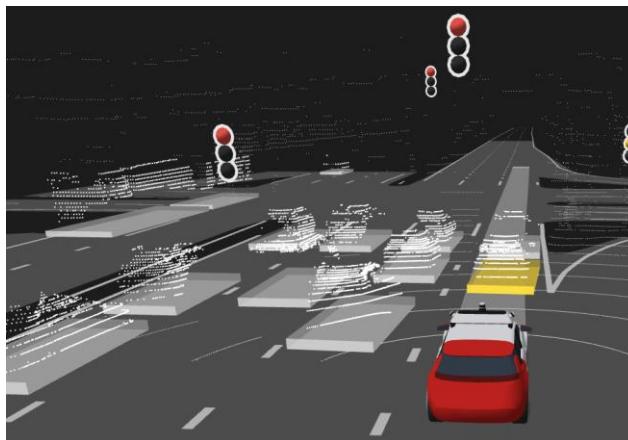
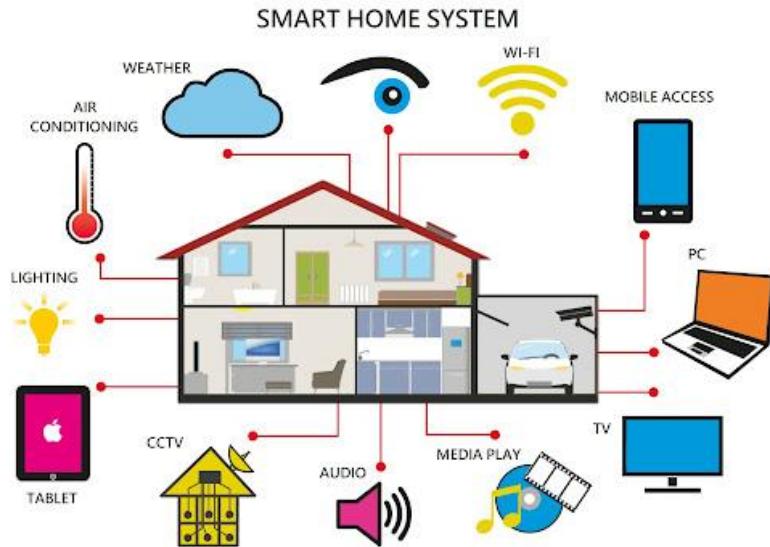
- ✓ Simplicity and Efficiency

Easy to implement and computationally tractable for small to medium-sized domains.

Propositional Logic: Introduction

➤ Applications of Propositional Logic

Automate lights

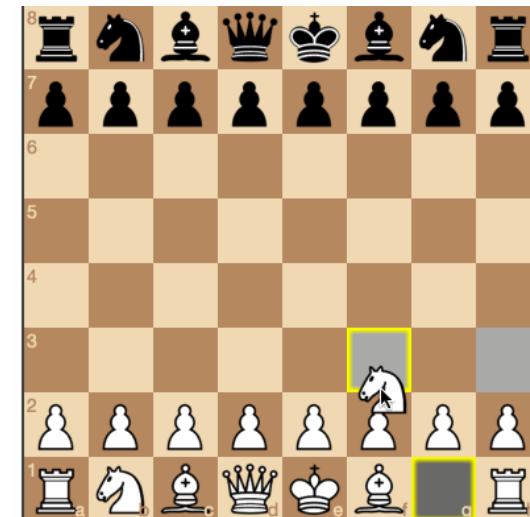


Make decisions on the road

MEDICAL DIAGNOSTIC



Support doctors
with diagnosis



Make moves based on conditions

Propositional Logic: Syntax

-Define allowable sentences in a language

e.g., In the English:

“I am a female” is a sentence; “a I female am” is not.

Propositional Logic: Syntax

➤ Proposition

A **statement/sentence** that is either true or false, but not both

Atomic proposition

A basic and indivisible statement

e.g., “*It is raining*”

(represented by a single proposition symbol, typically an uppercase letter like A,B,C,etc.)

Compound proposition

Built from atomic ones using **logical operators**.

e.g., “*It is raining AND I am indoors*”

Propositional Logic: Syntax



➤ Logical Connectives

Logical symbols that connect proposition symbols in order to reason in a more complex way about the world

- ✓ \neg (**not**): A sentence such as $\neg P$ is called **negation of P**
e.g., P: “It is raining”, $\neg P$: “It is NOT raining”
- ✓ \wedge (**and**): A sentence whose main connective is \wedge , such as $P \wedge Q$ is called a **conjunction**
e.g., P: “It is raining”, Q: “I bring an umbrella”.
 $P \wedge Q$: “It is raining AND I bring an umbrella”.
- ✓ \vee (**or**): A sentence whose main connective is \vee , such as $P \vee Q$ is called a **disjunction**

Propositional Logic: Syntax

➤ Logical Connectives

Logical symbols that connect proposition symbols in order to reason in a more complex way about the world

- ✓ \Rightarrow (**implies**): A sentence such as $P \Rightarrow Q$ is called an **implication/conditional**. **P** is called **premise/antecedent**; **Q** is called the **consequent/conclusion**

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

e.g., P: “It is raining”, Q: “I am in doors”.

$P \Rightarrow Q$: “IF it is raining, THEN I am indoors”

- ✓ \Leftrightarrow (**if and only if**): A sentence such as $P \Leftrightarrow Q$ is called a **biconditional**

P	Q	$P \Leftrightarrow Q$
T	T	T
T	F	F
F	T	F
F	F	T

e.g., P: “It is raining”, Q: “I am in doors”.

$P \Leftrightarrow Q$: “IF it is raining, THEN I am indoors. AND IF I am indoors, THEN it is raining”

Propositional Logic: Syntax

➤ Operator Precedence

Highest

Lowest



e.g., $\neg A \Leftrightarrow B \vee C \wedge D \Rightarrow E$

Propositional Logic: Syntax



Practice: Translate English to Propositional Logic

Turn on light when dark and motion

dark & motion \Rightarrow light on

Harry visited Hagrid or Dumbledore today

visited = H | D

The ground is not dry and it is cloudy

\sim dry ground & cloudy

Propositional Logic: Syntax

➤ Well-formed Formula (WFF)

T / F

are WFF too

A valid expression built according to the syntactic rules:

1. Any proposition symbol is a WFF

A WFF states what is a legal statement in logic,
not well formed = no meaning and cannot be
reasoned about. This is imp for truth table,
symbolic manipulation, etc

2. If α is a WFF, then $\neg \alpha$ is a WFF

meaningless

3. If α and β are WFFs, then the following are also WFFs:

$$(\alpha \underset{\text{—}}{\wedge} \beta), (\alpha \underset{\text{—}}{\vee} \beta), (\alpha \underset{\text{—}}{\Rightarrow} \beta), (\alpha \underset{\text{—}}{\Leftrightarrow} \beta)$$

e.g., Q1: If A and B are proposition symbols, $\neg(A \vee \neg B)$? *y*

Q2: $\neg(\alpha \vee \neg \beta)$? *d k* may not be proposition symbols

Q3: If A and B are proposition symbols, $A \underset{\text{—}}{\wedge} \vee B$? *not a valid logical exp*

Propositional Logic: Semantics

-Defines the truth of a sentence **in a given world**
(or WFF) **(or model)**

e.g., In the English:



“Wang Si is a female” is True



“Wang Si is a female” is False

Propositional Logic: Semantics

➤ Rules for Determining the Truth

- Rules for computing the truth of **atomic propositions**

The truth value of every proposition symbol must be specified directly in the **MODEL**



An assignment of truth value to **every** proposition symbol

e.g., P: "It is raining", Q: "I am indoors", R: "I bring an umbrella"

Model $m = \{P=\text{True}, Q=\text{False}, R=\text{True}\}$

What about $m = \{P=\text{True}\}$?

m satisfies P (or m is a model of P)

$M(P)$: the set of all models of P

2^3

must be valid models (include all symbols' truth value)

How many possible models in total?

Propositional Logic: Semantics



➤ Rules for Determining the Truth

- Rules for computing the truth of **compound propositions**

Rule 1: $\neg P$ is true if and only if P is false in model m

P	$\neg P$
True	False
False	True

e.g., P : “It is raining”, $\neg P$: “It is not raining”

Propositional Logic: Semantics



➤ Rules for Determining the Truth

- Rules for computing the truth of **compound propositions**

Rule 2: $P \wedge Q$ is true **if and only if** both P and Q are true in m

P	Q	$P \wedge Q$
True	True	True
True	False	False
False	True	False
False	False	False

e.g., P: “It is raining”, Q: “I bring an umbrella”,
 $P \wedge Q$: “It is raining and I bring an umbrella”.

Propositional Logic: Semantics

➤ Rules for Determining the Truth

- Rules for computing the truth of **compound propositions**

Rule 3: $P \vee Q$ is true **if and only if** at least one of them is true
in m

P	Q	$P \vee Q$
True	True	True
True	False	True
False	True	True
False	False	False

e.g., P: “It is raining”, Q: “I bring an umbrella”,
 $P \vee Q$: “It is raining or I bring an umbrella”.

Propositional Logic: Semantics

➤ Rules for Determining the Truth

- Rules for computing the truth of **compound propositions**

Rule 4: $P \Rightarrow Q$ is true **unless** P is true and Q is false in m

If P is true and Q is false, then $P \Rightarrow Q$ is false

otherwise, $P \Rightarrow Q$ is true

P	Q	$P \Rightarrow Q$
True	True	True
True	False	False
False	True	True
False	False	True

vacuously true

e.g., P: “It is raining”, Q: “I am indoors”,
 $P \Rightarrow Q$: “If it is raining, then I am indoors”.

Propositional Logic: Semantics

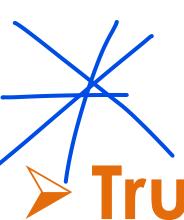
➤ Rules for Determining the Truth

- Rules for computing the truth of **compound propositions**

Rule 5: $P \Leftrightarrow Q$ is true **if and only if** P and Q are both true or both false in m

P	Q	$P \Leftrightarrow Q$
True	True	True
True	False	False
False	True	False
False	False	True

e.g., P: “It is raining”, Q: “I am indoors”,
 $P \Leftrightarrow Q$: “If it is raining, then I am indoors.
And if I am indoors, then it is raining”.



Propositional Logic: Semantics

Truth Table

A table that shows the **truth value** of a WFF for **every possible combination** of truth values of its atomic propositions

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
true	true	false	true	true	true	true
true	false	false	false	true	false	false
false	true	true	false	true	true	false
false	false	true	false	false	true	true

Propositional Logic: Semantics

➤ Entailment (Logical Consequence)

A WWF α follows logically from another WWF β

(i.e., β entails α)

$$\beta \vDash \alpha$$

$\beta \vDash \alpha$ if and only if $M(\beta) \subseteq M(\alpha)$

another way to think is all models of alpha is a proper subset of beta

(i.e., if β is true, then α must be true)

e.g., $P \vDash (P \vee Q) ?$

Yes, If P is true then P or Q is tru

$Q \vDash (P \Rightarrow Q) ?$

Yes, Q is true can surrender P implies Q is true (refer to the truth table). If P is true, then Q is true.
IF P is false, Q is vacuously true.

$(P \wedge Q) \vDash P ?$

Yes, for P&Q to be true, P must be true

Propositional Logic: Semantics

➤ Tautology

a sentence is valid if it is true in ALL models,
stricter than satisfiable

A WWF that is **always true** in all possible models.

e.g., $\neg P \vee P$

➤ Contradiction

A WWF that is **always false** in all possible models.

e.g., $\neg P \wedge P$ not valid and unsatisfiable

➤ Satisfiability

A WWF that is true/satisfied by at least one model

An expression is satisfiable if there is at least one assignment of truth values that make it true

if we can find ONE model that can make the WFF to be true.

Unsatisfiable means no way to assign the symbols a truth value such that the expression is true

fastest use truth table

just means there is a model where this WWF holds a True value

See slide / 7

Propositional Logic: Semantics

➤ Logical Equivalence

$$\mathcal{M}(\alpha) \subseteq \mathcal{M}(\beta)$$

Two WFFs α and β are logically equivalent if they are true in the same set of models.

$\alpha \equiv \beta$ if and only if $\alpha \models \beta$ and $\beta \models \alpha$

$$\mathcal{M}(\beta) \subseteq \mathcal{M}(\alpha)$$

$$(\alpha \wedge \beta) \equiv (\beta \wedge \alpha) \quad \text{commutativity of } \wedge$$

$$(\alpha \vee \beta) \equiv (\beta \vee \alpha) \quad \text{commutativity of } \vee$$

$$((\alpha \wedge \beta) \wedge \gamma) \equiv (\alpha \wedge (\beta \wedge \gamma)) \quad \text{associativity of } \wedge$$

$$((\alpha \vee \beta) \vee \gamma) \equiv (\alpha \vee (\beta \vee \gamma)) \quad \text{associativity of } \vee$$

$$\neg(\neg \alpha) \equiv \alpha \quad \text{double-negation elimination}$$

$$(\alpha \Rightarrow \beta) \equiv (\neg \beta \Rightarrow \neg \alpha) \quad \text{contraposition}$$

$$(\alpha \Rightarrow \beta) \equiv (\neg \alpha \vee \beta) \quad \text{implication elimination}$$

$$(\alpha \Leftrightarrow \beta) \equiv ((\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)) \quad \text{biconditional elimination}$$

$$\neg(\alpha \wedge \beta) \equiv (\neg \alpha \vee \neg \beta) \quad \text{De Morgan}$$

$$\neg(\alpha \vee \beta) \equiv (\neg \alpha \wedge \neg \beta) \quad \text{De Morgan}$$

$$(\alpha \wedge (\beta \vee \gamma)) \equiv ((\alpha \wedge \beta) \vee (\alpha \wedge \gamma)) \quad \text{distributivity of } \wedge \text{ over } \vee$$

$$(\alpha \vee (\beta \wedge \gamma)) \equiv ((\alpha \vee \beta) \wedge (\alpha \vee \gamma)) \quad \text{distributivity of } \vee \text{ over } \wedge$$

Propositional Logic: Semantics

➤ Knowledge Base (KB)

A set of **WWFs** that represent **known** facts or rules about the world.

P: "It is raining", Q:"I am indoors", R: "I bring an umbrella"

$$KB = \{P, Q, P \Leftrightarrow R\}$$

P, Q, $P \Leftrightarrow R$ are assumed to be **true** in the KB.



$$KB = \{P \wedge Q \wedge (P \Leftrightarrow R)\}$$

What about $KB = \{\neg P, P \wedge Q, P \Leftrightarrow R\}$?

N

$\neg P \wedge P$

always false

all true

Propositional Logic: Inference

-Process of deriving new conclusions from a given knowledge base (KB)

inference answers the question: does KB logically entail a query if yes, the query is true within our KB

Propositional Logic: Inference



➤ Model-checking Algorithm

Suppose that we want to know if α is true, given KB.



Check if $\text{KB} \models \alpha$

1. Enumerate all the possible models
2. Check if α is true in every model in which KB is true

Propositional Logic: Inference

➤ Model-checking Algorithm

1. Enumerate all the possible models
2. Check if α is true in every model in which KB is true

e.g., P: “It is raining”, Q: “I am indoors”, R: “I bring an umbrella”

$$KB = \{P, \neg Q, (P \wedge \neg Q) \Rightarrow R\}$$

Query (Goal): $P \wedge R?$

P	Q	R	KB
false	false	false	
false	false	true	
false	true	false	
false	true	true	
true	false	false	
true	false	true	
true	true	false	
true	true	true	

evaluate
this
to see
which
KB

Propositional Logic: Inference

➤ Model-checking Algorithm

1. Enumerate all the possible models
2. Check if α is true in every model in which KB is true

e.g., P: “It is raining”, Q: “I am indoors”, R: “I bring an umbrella”

$$KB = \{P, \neg Q, (P \wedge \neg Q) \Rightarrow R\}$$

Query (Goal): $P \wedge R?$

P	Q	R	KB
false	false	false	false
false	false	true	false
false	true	false	false
false	true	true	false
true	false	false	false
true	false	true	true
true	true	false	false
true	true	true	false

KB satisfies the query
The query is satisfiable

$P \wedge R=?$

true

Propositional Logic: Inference

Inference Rule

Generate new information based on existing knowledge
without considering every possible model

if all the statements above the line are true, then the statement below the lines must be true

$$\frac{\alpha_1, \alpha_2, \dots, \alpha_k}{\beta}$$

what is true

what logically follows from what is true

Premise Conclusion

A red arrow points from the text "if all the statements above the line are true, then the statement below the lines must be true" to the line above the main formula.

e.g., α_1 : If it is raining, then I am indoors. α_2 : It is raining

What is β ?

1. Modus Ponens

$$\frac{(\alpha \Rightarrow \beta, \alpha)}{\beta}$$

since a is true, b is true

if alpha implies beta and alpha is true then beta must also be true

2. And Elimination

$$\frac{\alpha \wedge \beta}{\alpha}$$

this is true

this must be true

beta is also true

Propositional Logic: Inference

Inference Rule

3. Logical Equivalence:

$$\alpha \equiv \beta$$

Inference Rules:

$$\frac{\alpha}{\beta}, \frac{\beta}{\alpha}$$

- $(\alpha \wedge \beta) \equiv (\beta \wedge \alpha)$ commutativity of \wedge
- $(\alpha \vee \beta) \equiv (\beta \vee \alpha)$ commutativity of \vee
- $((\alpha \wedge \beta) \wedge \gamma) \equiv (\alpha \wedge (\beta \wedge \gamma))$ associativity of \wedge
- $((\alpha \vee \beta) \vee \gamma) \equiv (\alpha \vee (\beta \vee \gamma))$ associativity of \vee
- $\neg(\neg \alpha) \equiv \alpha$ double-negation elimination
- $(\alpha \Rightarrow \beta) \equiv (\neg \beta \Rightarrow \neg \alpha)$ contraposition
- $(\alpha \Rightarrow \beta) \equiv (\neg \alpha \vee \beta)$ implication elimination
- $(\alpha \Leftrightarrow \beta) \equiv ((\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha))$ biconditional elimination
- $\neg(\alpha \wedge \beta) \equiv (\neg \alpha \vee \neg \beta)$ De Morgan
- $\neg(\alpha \vee \beta) \equiv (\neg \alpha \wedge \neg \beta)$ De Morgan
- $(\alpha \wedge (\beta \vee \gamma)) \equiv ((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$ distributivity of \wedge over \vee
- $(\alpha \vee (\beta \wedge \gamma)) \equiv ((\alpha \vee \beta) \wedge (\alpha \vee \gamma))$ distributivity of \vee over \wedge

e.g.,

$$\frac{\alpha \Leftrightarrow \beta}{(\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)}$$
$$\frac{(\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)}{\alpha \Leftrightarrow \beta}$$

Propositional Logic: Inference

Inference Rule

4. Resolution Rule

$$\frac{\alpha \vee \beta, \neg\alpha}{\beta}$$

if not α is true, α is false. Since α or β is true, β must be true

e.g., $\alpha \vee \beta$: Sandy is in the canteen or library.

$\neg\alpha$: Sandy is not in the canteen.

$$\frac{\alpha_1 \vee \dots \vee \alpha_n, \gamma_1 \vee \dots \vee \gamma_m}{\alpha_1 \vee \dots \vee \alpha_{i-1} \vee \alpha_{i+1} \vee \dots \vee \alpha_n \vee \gamma_1 \vee \dots \vee \gamma_{j-1} \vee \gamma_{j+1} \vee \dots \vee \gamma_m}$$

where α_i and γ_j are complementary literals

(Literal: atomic proposition or negated atomic proposition)

Propositional Logic: Inference

Inference Rule

4. Resolution Rule

$$\alpha_1 \vee \cdots \vee \alpha_n, \gamma_1 \vee \cdots \vee \gamma_m$$

$$\alpha_1 \vee \cdots \vee \alpha_{i-1} \vee \alpha_{i+1} \vee \cdots \vee \alpha_n \vee \gamma_1 \vee \cdots \vee \gamma_{j-1} \vee \gamma_{j+1} \vee \cdots \vee \gamma_m$$

Resolution rule only applies to clauses (disjunction of literals), so need to convert sentence to conjunctive normal form or CNF

Clause: a disjunction of literals

Can it be applied to other WFFs for premises (such as \Rightarrow , \Leftrightarrow)?

Every proposition \equiv A conjunction of clauses

(Conjunctive Normal Form (CNF))

$$Clause_1 \wedge \cdots \wedge Clause_n$$

e.g., Convert $A \Leftrightarrow (B \vee C)$ into CNF

CNF requires \neg to appear only in literals, so move \neg inwards by repeated application of equivalences

$\neg(\neg a) \equiv a$ (double-negation elimination)
 $\neg(a \& b) \equiv (\neg a \vee \neg b)$ (De Morgan)
bubble pushing

Propositional Logic: Inference

e.g., Convert $A \Leftrightarrow (B \vee C)$ into CNF

1. Eliminate \Leftrightarrow , replacing $\alpha \Leftrightarrow \beta$ with $(\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha)$.

$$(A \Rightarrow (B \vee C)) \wedge ((B \vee C) \Rightarrow A)$$

2. Eliminate \Rightarrow , replacing $\alpha \Rightarrow \beta$ with $\neg\alpha \vee \beta$

$$(\neg A \vee (B \vee C)) \wedge (\neg(B \vee C) \vee A)$$

3. Move \neg inwards by De Morgan

$$(\neg A \vee (B \vee C)) \wedge ((\neg B \wedge \neg C) \vee A)$$

4. Distribute \vee over \wedge

$$(\neg A \vee (B \vee C)) \wedge ((\neg B \vee A) \wedge (\neg C \vee A))$$

5. Associativity

$$(\neg A \vee B \vee C) \wedge (\neg B \vee A) \wedge (\neg C \vee A)$$

Propositional Logic: Inference

➤ Resolution Algorithm

Check if $\text{KB} \models \alpha$  Check if $\text{KB} \wedge \neg\alpha$ is contradiction

Step 1. Convert $\text{KB} \wedge \neg\alpha$ into CNF

Step 2. Apply resolution rule to each pair of clauses that contains **complementary literals** to produce a new clause, which is added to the KB if it is not already present, until one of the following happens:

1. Add negation of wtv we want to check
2. AND every clause in KB with it
3. Apply res rule which finds clauses with literals that are complement (B and not B), then simplify them which will result in new clauses.

4. Add those into the stack and apply res rule again, until the 2 on the right happens (e.g. B clause and not B clause, both can't be true at the same time = contradiction) so proven by contradiction, KB does entail alpha. Or else, if no new clause, means no entailment

- Empty clause is produced (equivalent to False)

$\text{KB} \models \alpha$

Propositional Logic : Inference

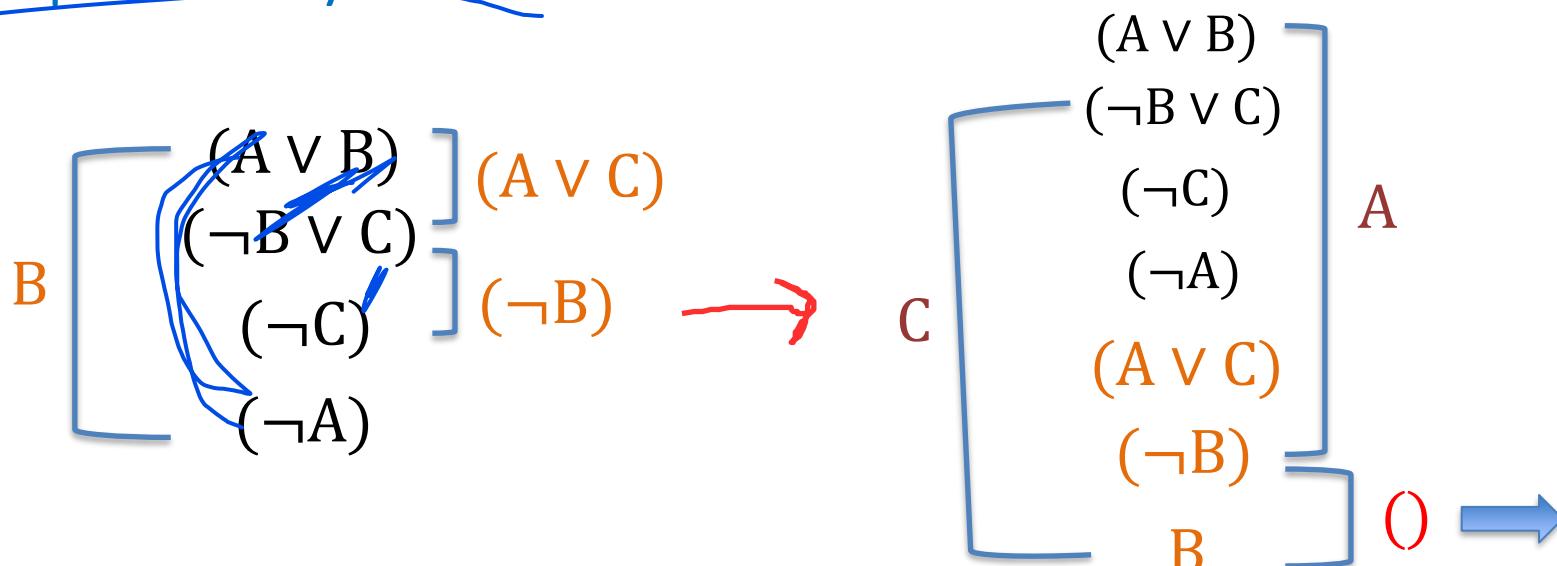
➤ Resolution Algorithm

e.g., does $\text{KB} = \{A \vee B, \neg B \vee C, \neg C\}$ entail A ?

Step 1. Convert $\text{KB} \wedge \neg A$ into CNF

$$\text{KB} \wedge \neg A = (A \vee B) \wedge (\neg B \vee C) \wedge (\neg C) \wedge (\neg A)$$

Step 2. Apply resolution rule to each pair of clauses that contains complementary literals



Propositional Logic : Inference

➤ Resolution Algorithm

Important!

Empty clause produced must arise only from resolving **two single literals (unit clauses)**

$$\text{e.g., } \frac{A \vee B, \neg A \vee \neg B}{?}$$

~~A, B~~

$$\frac{A, \neg A}{?} \checkmark$$

Propositional Logic : Inference

➤ Resolution Algorithm

Important!

Empty clause produced must arise only from resolving **two single literals (unit clauses)**

e.g.,
$$\frac{A \vee B, \neg A \vee \neg B}{A \vee \neg A} \text{ tautology}$$

or

$$\frac{A \vee B, \neg A \vee \neg B}{B \vee \neg B} \text{ tautology}$$

$$\frac{A, \neg A}{\phi}$$

Resolve one pair of complementary literals at a time!

Propositional Logic : Inference



➤ Forward Chaining Algorithm

Infer atomic propositions based on KB of **definite clauses**.

or aka Horn clauses



A disjunction of literals of which exactly one is positive

A & B -> C
this is equivalent to RHS,
meaning can keep spamming modus
ponens to simplify an expression to a
single atomic proposition and search
for the query

e.g., $(\neg A \vee \neg B \vee C)$

Step 1: Start with atomic propositions in the KB

A, B,

Step 2: Apply inference rules to infer new facts and add them to the KB

Step 3: Continue until:

-The goal (query) is derived, or

-No more new facts can be inferred

Propositional Logic : Inference

➤ Forward Chaining Algorithm

Step 1: Start with atomic propositions in the KB

Step 2: Apply inference rules to infer new facts and add them to the KB

Step 3: Continue until:

- The goal (query) is derived, or

only modus ponens is used, and get all the form $A \& B \Rightarrow P$ and simplify to RHS, until query is obtained

- No more new facts can be inferred

atomic ↗

e.g., $\text{KB} = \{A, B, A \wedge B \Rightarrow C, A \wedge P \Rightarrow L, B \wedge C \Rightarrow P\}$, Query: L

Step 1: A is true. B is true

Step 2: From $A \wedge B \Rightarrow C$, infer C is true.

literally working forward
from atomic to query

Update $\text{KB} = \{A, B, A \wedge B \Rightarrow C, A \wedge P \Rightarrow L, B \wedge C \Rightarrow P, C\}$

add ↗

From $B \wedge C \Rightarrow P$, infer P is true.

Update $\text{KB} = \{A, B, A \wedge B \Rightarrow C, A \wedge P \Rightarrow L, B \wedge C \Rightarrow P, C, P\}$

From $A \wedge P \Rightarrow L$, infer L is true.



Propositional Logic : Inference



➤ Backward Chaining Algorithm

Infer atomic propositions based on KB of **definite clauses**.

Start with the query (goal)

Step 1: Find implications in the KB whose conclusions are query.

Step 2: Try to prove the premises of those implications

Step 3: Recursively prove subgoals until reaching the atomic propositions in the KB

Propositional Logic : Inference

➤ Backward Chaining Algorithm

Step 1: Find implications in the KB whose conclusions are query.

Step 2: Try to prove the premises of those implications.

Step 3: Recursively prove subgoals until reaching the atomic propositions in the KB

really just the opposite of forward. Use query to check if propositions in KB are true.

e.g., $\text{KB} = \{A, B, A \wedge B \Rightarrow C, A \wedge P \Rightarrow L, B \wedge C \Rightarrow P\}$, Query: L

Step 1: Find $A \wedge P \Rightarrow L$ whose conclusion is L, check if $A \wedge P$ is true.

Step 2: A is true; Find $B \wedge C \Rightarrow P$ whose conclusion is P, check if $B \wedge C$ is true.

B is true; Find $A \wedge B \Rightarrow C$ whose conclusion is C

$A \wedge B$ is true \rightarrow C is true ✓✓

Summary



✓ Syntax

Propositions, Logical Connectives, Well-formed Formula (WFF)

✓ Semantics

Truth Table, Model, Entailment, Tautology, Contradiction, Satisfiability, Logical Equivalence, Knowledge Base

✓ Inference

Model Checking Algorithm, Resolution Algorithm, Forward Chaining Algorithm, Backward Chaining Algorithm.

THE END

EE2213 Introduction to Artificial Intelligence

Lecture 14

Dr. Shaojing Fan
fanshaojing@nus.edu.sg

OVERVIEW OF COURSE CONTENTS

- **Introduction (Shaojing)**

- What is AI
- Applications of AI
- AI agent

- **Search (Shaojing)**

- Uninformed search algorithms: breadth-first, depth-first, uniform-cost(Dijkstra's algorithm)
- Informed search algorithms: greedy best-first, A*
- Applications

- **Optimisation (Shaojing)**

- Linear programming
- Convex problems
- Applications

- **Machine learning (Wang Si)**

- Supervised and unsupervised learning: regression, classification, clustering
- Neural networks and deep learning
- Applications

- **Knowledge representation (Wang Si)**

- Knowledge Representation and Reasoning
- Propositional Logic
- Applications

- **Ethical considerations (Shaojing)**

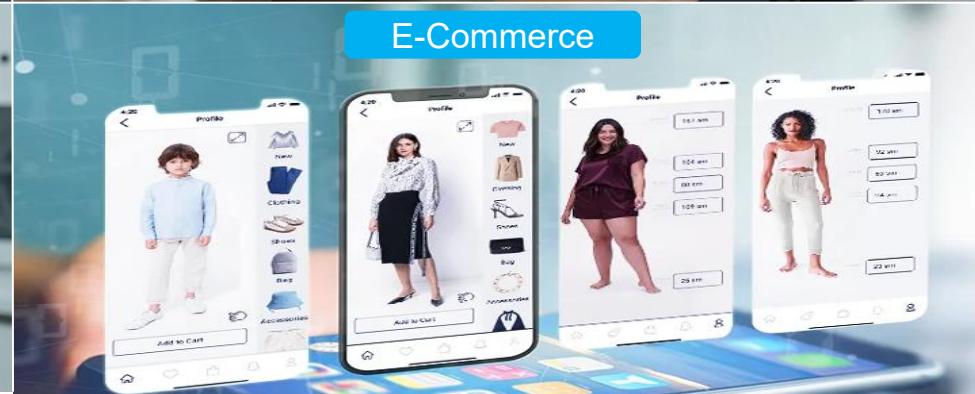
- Bias in AI
- Privacy concerns
- Societal impact



Agenda

- Generative AI & its social impact
- Bias in AI
 - Type of common biases
 - Ways to reduce biases
- Privacy concerns

Many facets of AI-generated content

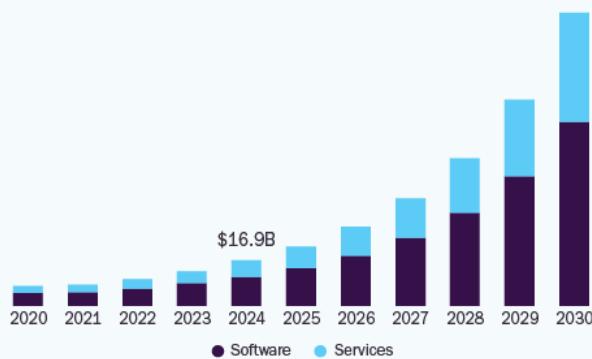


First three images were generated by Google Gemini 2.0 Flash

The growing prevalence of AI-generated content

Generative AI Market

Size, by Component, 2020 - 2030 (USD Billion)

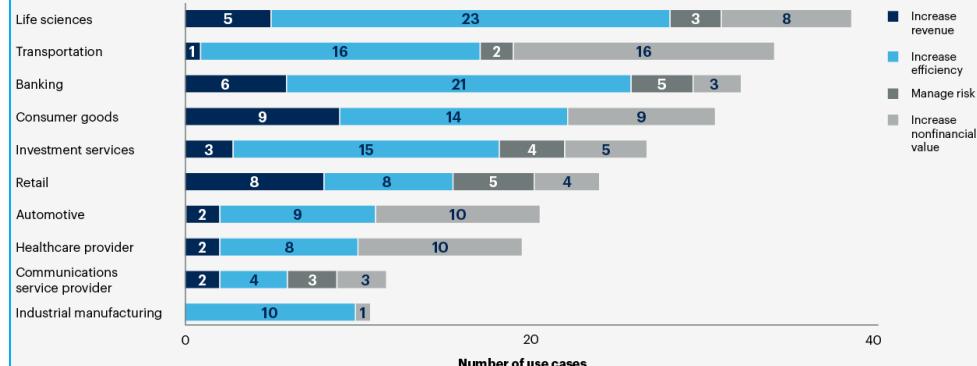


GRAND VIEW RESEARCH

37.6%
Global Market CAGR,
2025 - 2030

Source:
www.grandviewresearch.com

Business Value of Generative AI Case Examples by Industry



Source (left): Generative AI Market & Trends, *Grand View Research*, March, 2025. URL:

<https://market.us/report/generative-ai-in-content-creation-market/>

Source (right): 2024 generative AI case examples across industries. *Gartner*, September 2024. URL:

<https://www.gartner.com/en/articles/generative-ai-use-cases>

AI: A tool for progress or a misinformation machine?



Success in creating AI could be
the biggest event in the history of
our civilization. ...**Or the worst.**

*Stephen
Hawking*

Generative AI: Truth or Bias?



Why Google's AI tool was slammed for showing images of people of colour

Gemini's attempt to correct the marginalisation of ethnic minorities has backfired.



End Wokeness @EndWokeness



America's Founding Fathers, Vikings, and the Pope according to Google AI:



Summary: Ethical challenges in Generative AI



- Misinformation and Deepfakes: How generative AI can spread false or harmful content.
- Bias and Fairness: AI models can replicate or amplify societal biases.
- Ownership and Consent: Who owns AI-generated content? Was training data used with permission?

Studying human responses of AI content to combat misinformation



Input a piece of information (headline and image), predict its human responses.

Headline

Corresponding Image

Drop Image Here
~ OR ~
Click to Upload

Image Captioning Model

Blip-2

Explanation LLM

gpt-4o

Password

info@demo

clear

submit

Disclaimer: The predictions provided by this system are generated by artificial intelligence and are based on limited available data and observations. These predictions may not accurately reflect actual human reactions or behaviors. The system makes no promises or guarantees regarding the outcomes of using its product or service. Users are advised to use the predictions for informational purposes only and to exercise their own judgment and discretion when making decisions.

Shen, Zhiqi, Shaojing Fan, Danni Xu, Terence Sim, and Mohan Kankanhalli. "Modeling Human Responses to Multimodal AI Content." *arXiv preprint arXiv:2508.10769* (2025).

Xu, Danni, Shaojing Fan, and Mohan Kankanhalli. "Combating misinformation in the era of generative AI models." *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

Studying human responses of AI content to combat misinformation

H1: Temporal-Spatial Inconsistency (Account for 34.2%)



Reason 1--Discontinuous motions between current frames and following frames

H2: Pixel-Wise Inconsistency (Account for 31.5%)



Reason 2--Resolution: lower on left, higher on right



Reason 4--Color: mismatched color in left and right

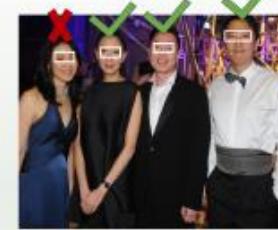


Reason 3--Light: lighter on left, darker on right



Reason 5--Artifacts: tampered on left, untouched on others

H3: Gaze Inconsistency (Account for 25.0%)



Reason 6--Abnormal averted gaze from camera

H4: Body-Face Inconsistency (Account for 7.5%)



Reason 7--Age: older body with younger face



Reason 8--Gender: female body with male face

Juan Hu, Shaojing Fan, and Terence Sim. "Seeing Through Deepfakes: A Human-Inspired Framework for Multi-Face Detection", ICCV, 2025

World-wide initiatives



Partnership on AI (PAI): Responsible Practices for Synthetic Media

Key Practices:

- **Disclosure** – Clearly label AI-generated content
- **Consent** – Use likeness/data only with appropriate permissions
- **Transparency** – Provide metadata or provenance info where possible

Who's Involved:

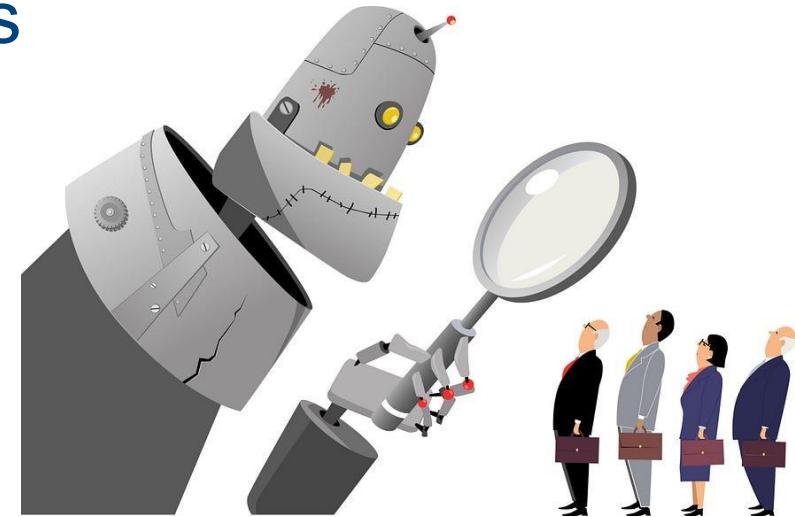
- 50+ global partners, incl. TikTok, OpenAI, BBC, Meta, Microsoft, Adobe
- Civil society orgs (WITNESS, Meedan), researchers, platforms

Website: <https://syntheticmedia.partnershiponai.org/>



Agenda

- Generative AI & its social impact
- Bias in AI
 - Type of common biases
 - Ways to reduce biases
- Privacy concerns



Common types of biases



•Data bias

Non-representative/imbalanced training datasets, biased heuristics

•Model bias

Simplifying assumptions that don't reflect real-world complexity, or (sometimes) overly complex or overfitted assumptions that reduce generalizability.

Deployment bias

Mismatch between model training conditions and real-world usage, such as applying it in the wrong setting or with different data than it was trained on.

•Evaluation bias

Flawed or inconsistent evaluation methods, such as relying on a single train/test split or using metrics that don't capture real-world performance

Example of biased heuristics

Imagine a GPS navigation system trying to find the fastest route from home to school. The system uses a heuristic to estimate travel times.

Suppose this heuristic **ignores traffic in urban areas**, assuming every road allows free-flowing traffic like a highway.

As a result, the algorithm **underestimates the cost of city routes** and thinks they are faster than they actually are.

bias and variance in training model,
underfitted and overfitted, refer to EE2211 slides

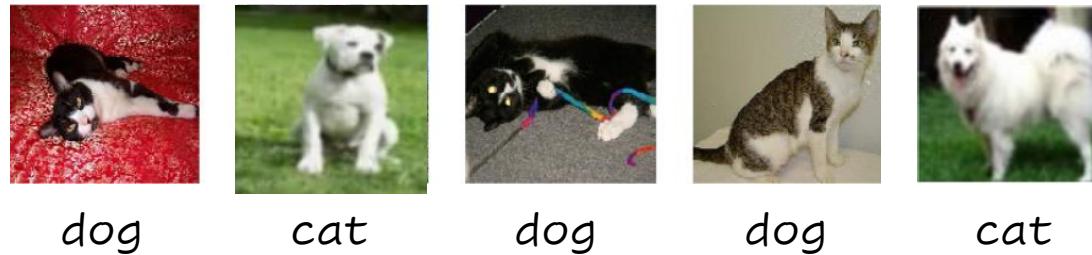


Example of data bias

I'm building a dog vs. cat classifier, observe my current training data set, do you notice any issues?



Training data



Test results

cats are white,
dogs are
coloured. If a
model is
overfitted, it will
have high bias (it
fits to training
data very well,
but using test
data it has low
accuracy)

Example of data bias

AI sentencing cut jail time for low-risk offenders, but study finds racial bias persisted

January 23, 2024 8:00 AM | Keith Brannon kbrannon@tulane.edu

[View PDF](#)



Broken Tools: Vincent Southerland illuminates the problem of bias in the criminal legal system's use of algorithmic tools



Source: <https://news.tulane.edu/pr/ai-sentencing-cut-jail-time-low-risk-offenders-study-finds-racial-bias-persisted>
<https://www.law.nyu.edu/news/ideas/vincent-southerland-algorithmic-tools-bias-criminal-justice-reform>

Takeaway: Reduce data bias in practice



- **Use diverse & balanced data**
Ensure fair representation across all groups.
can refer to EE2211 reduce bias by using less complex models and better feature selection
- **Audit & clean the dataset**
Fix labels, remove noise, and anonymize if needed.
- **Validate across subgroups**
Apply stratified cross-validation and track group-wise metrics.
- **Train with fairness in mind**
Use regularization, reweighting, or fairness-aware loss.

Common sources of bias



- **Data bias**

Non-representative/imbalanced training datasets, biased heuristics

- **Model bias (ML context)**

Simplifying assumptions that don't reflect real-world complexity, or (sometimes) overly complex or overfitted assumptions that reduce generalizability.

- **Deployment bias**

Mismatch between model training conditions and real-world usage, such as applying it in the wrong setting or with different data than it was trained on.

- **Evaluation bias**

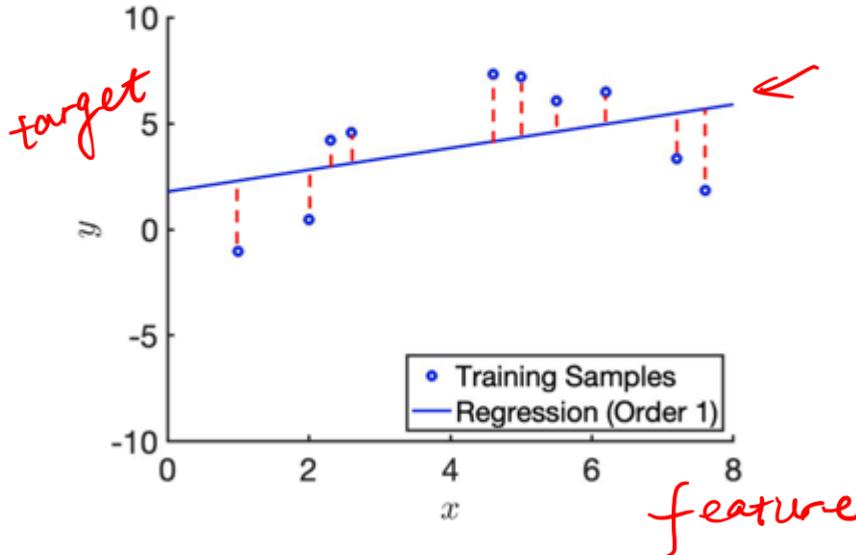
Flawed or inconsistent evaluation methods, such as relying on a single train/test split or using metrics that don't capture real-world performance

Example of model bias

Underfitting occurs when a model is too simplistic to capture the underlying patterns in the data, resulting in poor model performance.

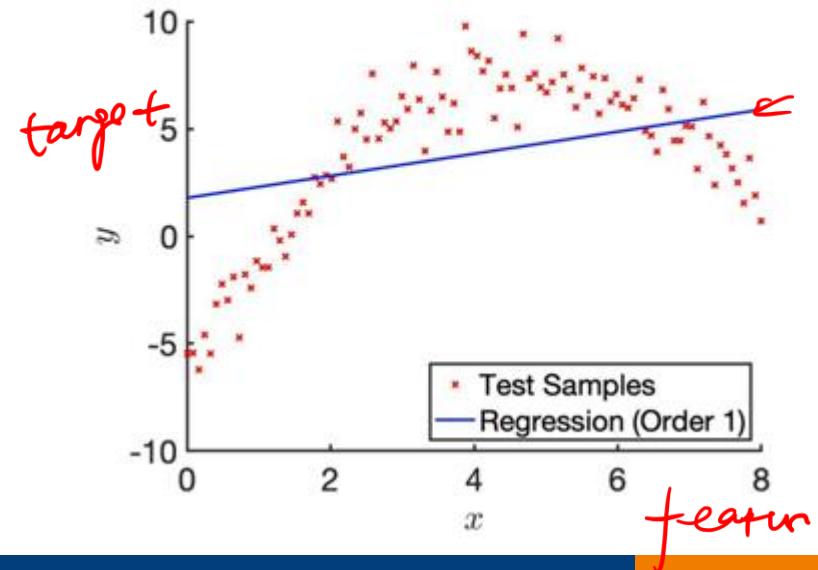
Training

Underfitting Example

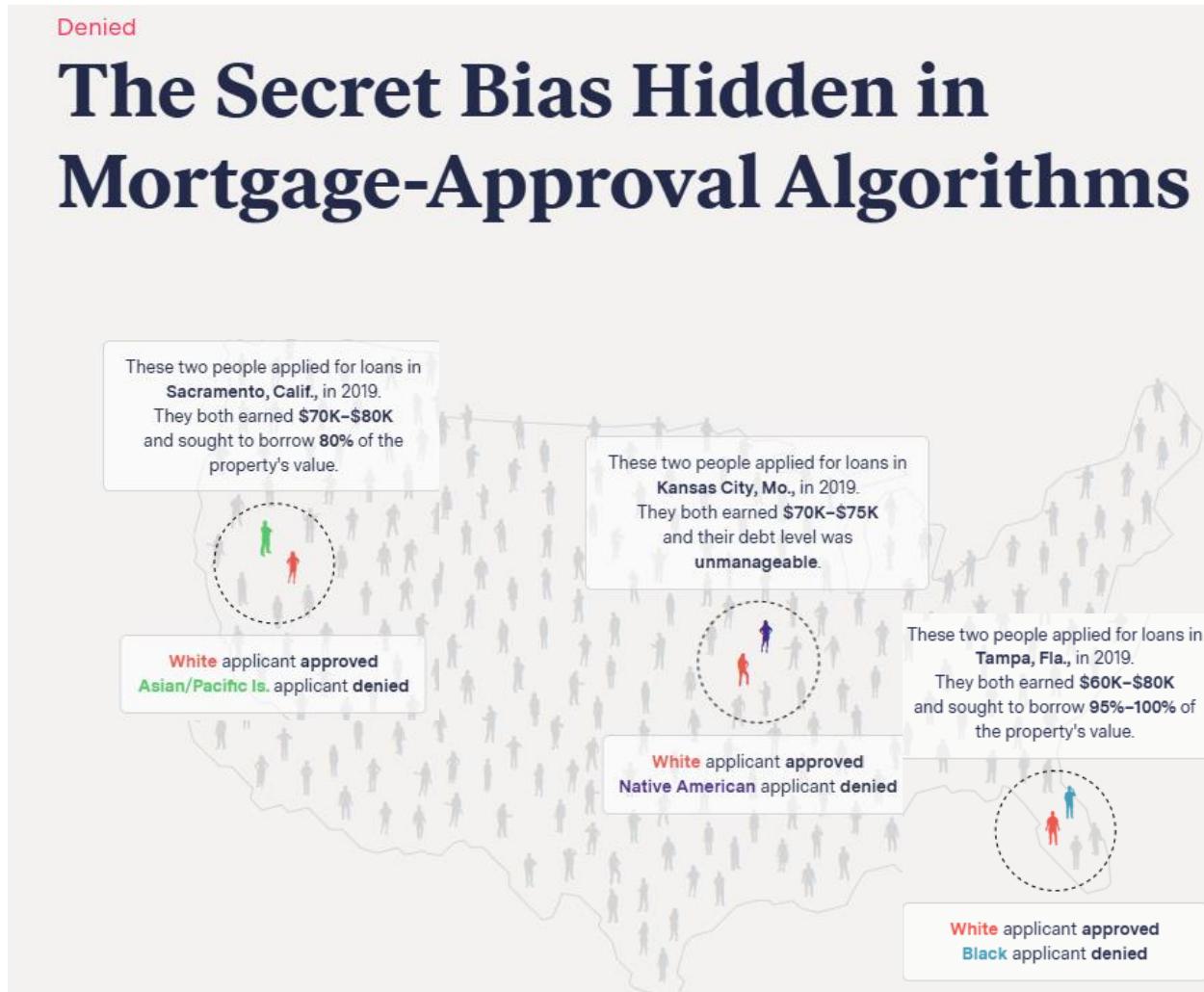


Testing

Underfitting Example



Example of model bias



A loan approval algorithm treats applicants with the same income and credit score as equally creditworthy, potentially overlooking systemic bias, historical discrimination, and other relevant features.

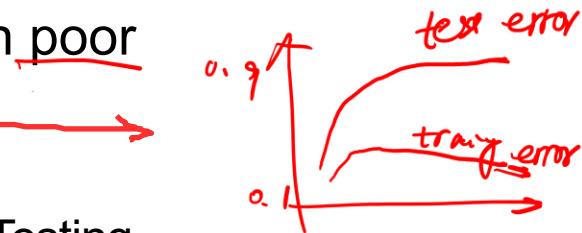
Source: <https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms>

can reduce features too (use better feature selection)

Example of model bias

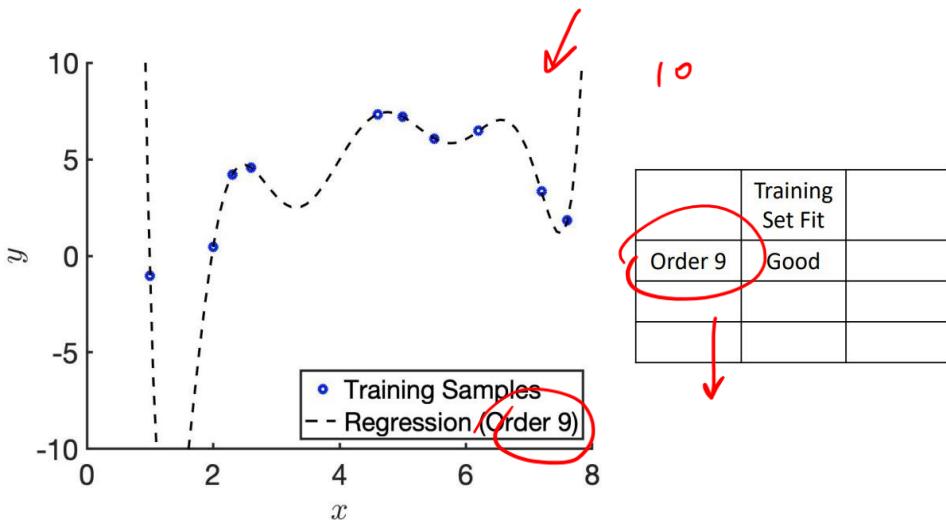
Overfitting occurs when a model is too complex and learns noise or random fluctuations in the training data, resulting in poor generalization to new data.

usually when high accuracy on training set but low on accuracy on test set

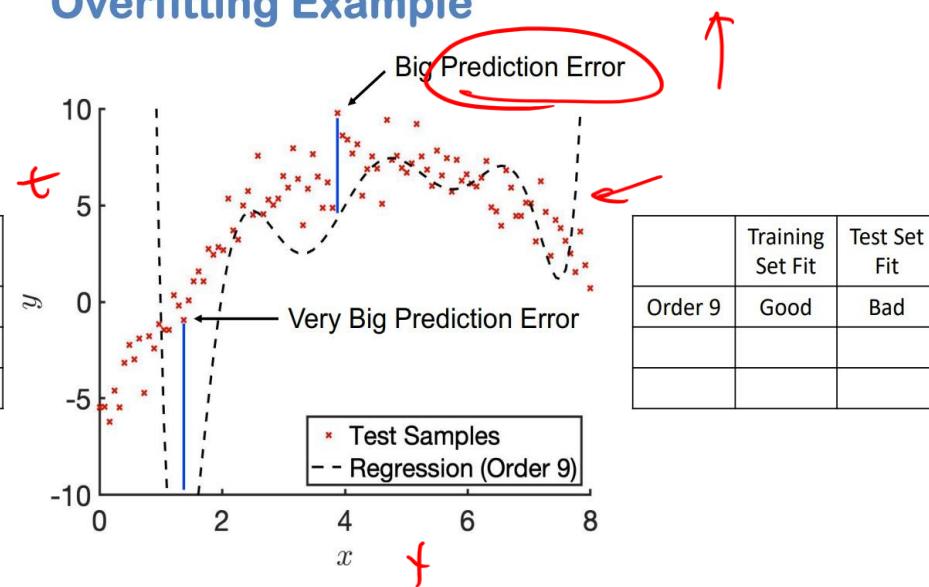


Training

Overfitting Example



Overfitting Example



Example of model bias

overfitting

The CNN Business homepage features a navigation bar with links to Markets, Tech, Media, Calculators, and Videos. Below this, there's a section for 'Markets' with data for DOW, S&P 500, and NASDAQ. To the right is the 'Fear & Greed Index' which shows a reading of 44, indicating 'Fear'. A news summary below the index states: 'Trump is soon expected to announce... Amazon just cut 14,000 jobs, and it's... CNN charts new course with All Acc...'.

Zillow's home-buying debacle shows how hard it is to use AI to value real estate



By Rachel Metz, CNN Business

7 min read · Published 7:32 AM EST, Tue November 9, 2021



A screenshot from a CNN Business video titled 'FORESEEABLE FUTURE: HOUSING MARKET MADNESS'. The video is a panel discussion featuring four participants: Christine Romans (@ChristineRomans), Glenn Kelman (@Redfin), Ryan Williams (@CadreRE), and David Doctorow (@realtordotcom). The video is labeled 'LIVE' in the top right corner. The CNN Business logo is in the top left. A 'Video Ad Feedback' button is visible at the bottom right.

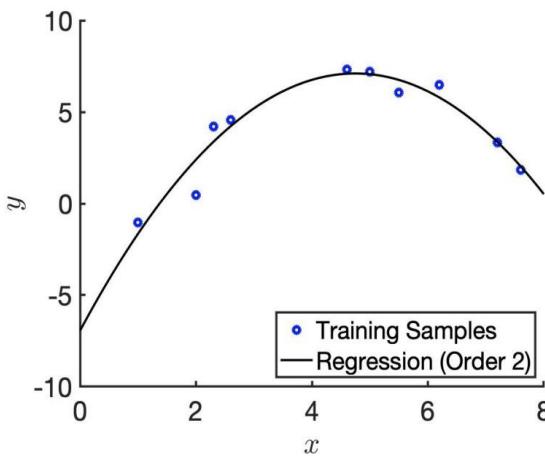
Source: <https://edition.cnn.com/2021/11/09/tech/zillow-is-buying-home-zestimate>

Perfect fitting

The model fits *both training and test* data well.

Training

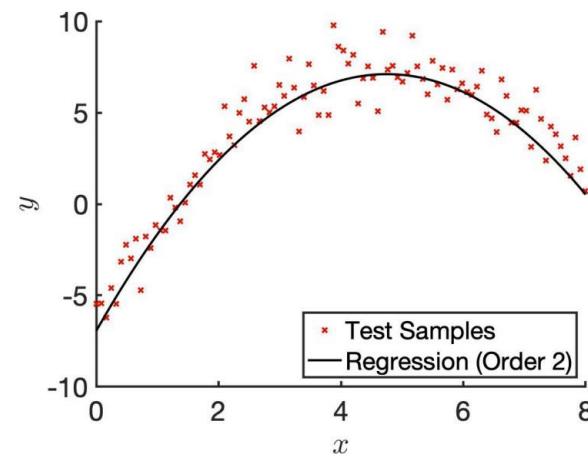
“Just Nice”



	Training Set Fit	Test Set Fit
Order 9	Good	Bad
Order 1	Bad	Bad
Order 2	Good	

Testing

“Just Nice”



	Training Set Fit	Test Set Fit
Order 9	Good	Bad
Order 1	Bad	Bad
Order 2	Good	Good

What causes overfitting (model's perspective)

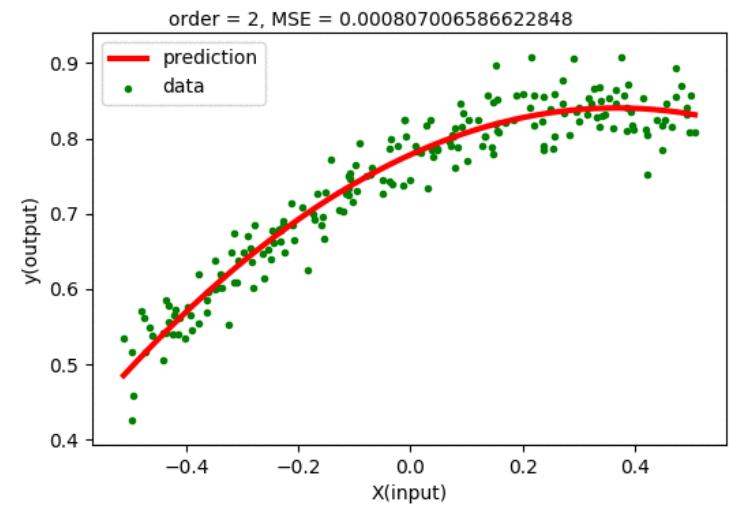
- Possible cause 1: The model is too complex for the given data

An overly complex model tries to capture noise or small fluctuations in the data rather than the underlying pattern.

Solution:

- Use a simpler model (e.g., reduce polynomial order)
- Use regularization (e.g., ridge regression)

Indications of model complexity in our current situation: # of orders; # of parameters; complexity of weight matrix, ...



Source: <https://aiinarabic.com/regression/>

What causes overfitting (data's perspective)

- Possible cause 2: There are too many features and too few training samples.

Application example: Predicting cancer risk based on gene expression

Inputs (Features, X): Gene expression levels of 10,000 genes for each patient.

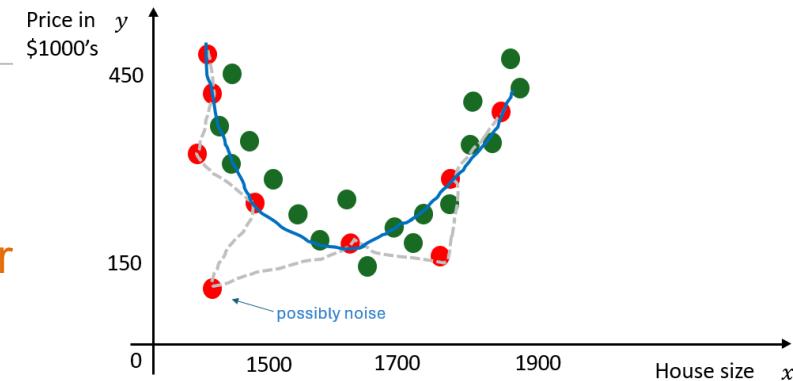
Output (Target, y): A binary variable indicating whether each patient developed cancer (1) or not (0).

$$y = w_0 + w_1x_1 + w_2x_2 + \dots + w_{10000}x_{10000}$$

What do we do if we have an equation with 10,001 parameters and only 500 data points? (under-determined system)

Solution:

- Collect more data
- Reduce the data dimensionality (e.g., feature selection)
- Use regularization

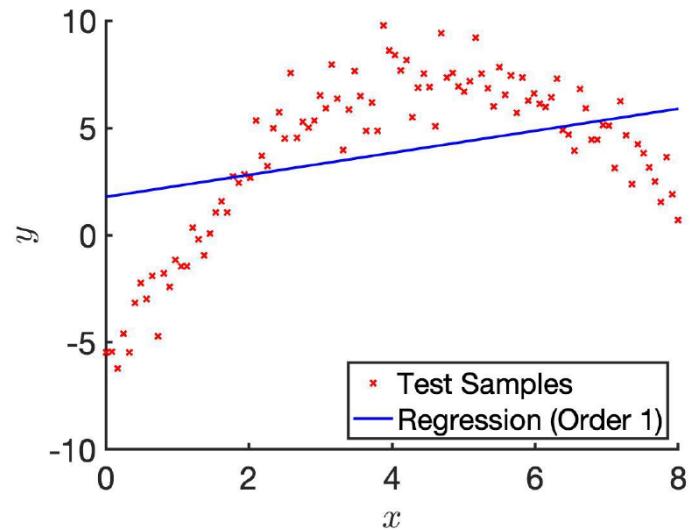


What causes underfitting (model's perspective)

- Possible cause 1: The model is too simple for the given data

Solution:

- Use a more complex model (e.g., increasing the polynomial order, or adopting a more sophisticated algorithm that better captures the data's complexity.)



What causes underfitting (data's perspective)

- Possible cause 2: The features are not informative enough

If the dataset's features lack relevance or meaning, the model may fail to learn useful patterns and instead capture irrelevant information.

Solution:

- Enhance feature engineering, incorporate relevant features, or collect higher-quality data to capture more informative patterns.

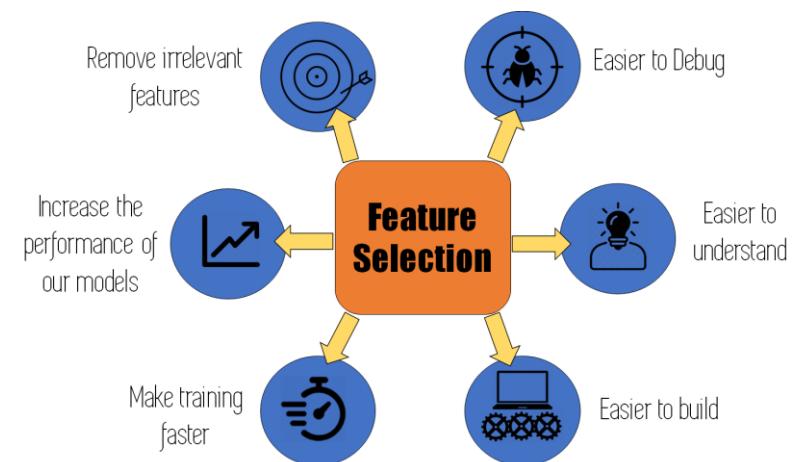


Image credit: <https://medium.com/@dancerworld60/filtering-out-irrelevant-features-a-comprehensive-survey-of-filter-based-techniques-for-feature-9275d86996ab>

Takeaway: Reduce model bias in practice



- **Perform exploratory data analysis**
Detect missing or underrepresented groups and skewed feature distributions early.
- **Apply careful feature selection & engineering**
Include relevant variables, especially those tied to fairness and context.
- **Regularly re-train with updated data**
Refresh the model with diverse, current data to reduce drift and bias over time.

use pearson's
correlation for
better feature
selection

Common sources of bias



- **Data bias**

Non-representative/imbalanced training datasets, biased heuristics

- **Model bias**

Simplifying assumptions that don't reflect real-world complexity.
(Sometimes) overly complex or overfitted assumptions that reduce generalizability.

- **Deployment bias**

Mismatch between model training conditions and real-world usage, such as applying it in the wrong setting or with different data than it was trained on.

- **Evaluation bias**

Flawed or inconsistent evaluation methods, such as relying on a single train/test split or using metrics that don't capture real-world performance

Example of deployment bias: The Challenger Disaster (1986)

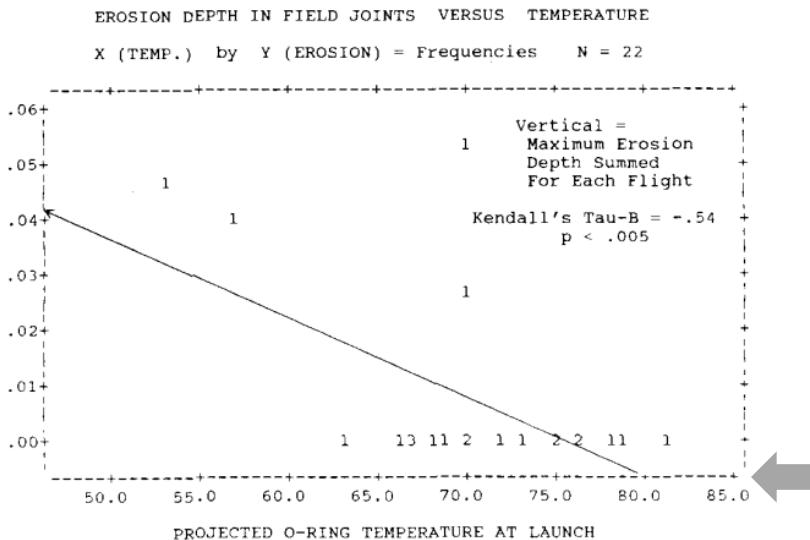
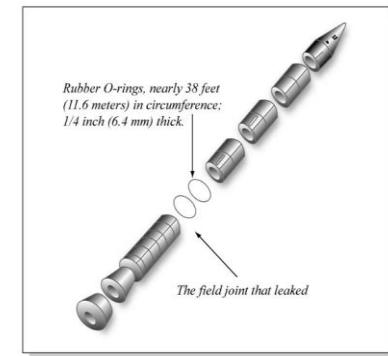
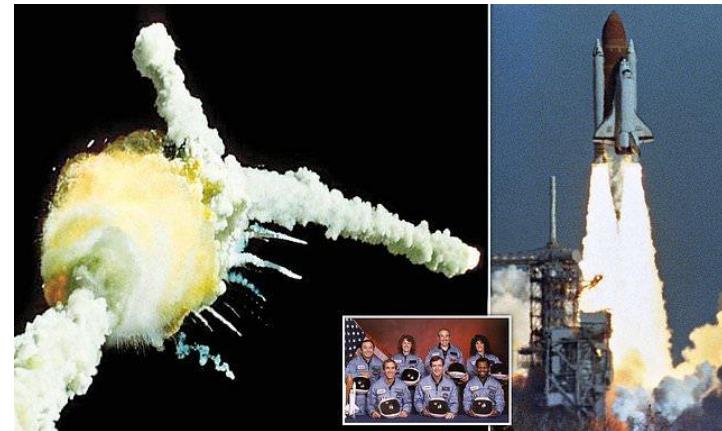


Fig. 3. Linear correlation between O-ring temperature and summed maximum depths (inches) of O-ring erosion for each flight for which complete data were available on January 27, 1986. Frequencies of observations at each bi-variate point in graph are indicated by number.

All the past data came from warmer temperatures (above 53° F or ~12° C).

On launch day, the temperature was unusually low (31° F or ~-1° C).



Courtesy of NASA, Figure by MIT OpenCourseWare

[1] References: Dixon, Robert. "The Challenger Space Shuttle Disaster: A Case Study in the Analysis of Binary Data Using Scatter Diagrams and Logit Regression." *Australian Economic Review* 54.2 (2021): 294-305.

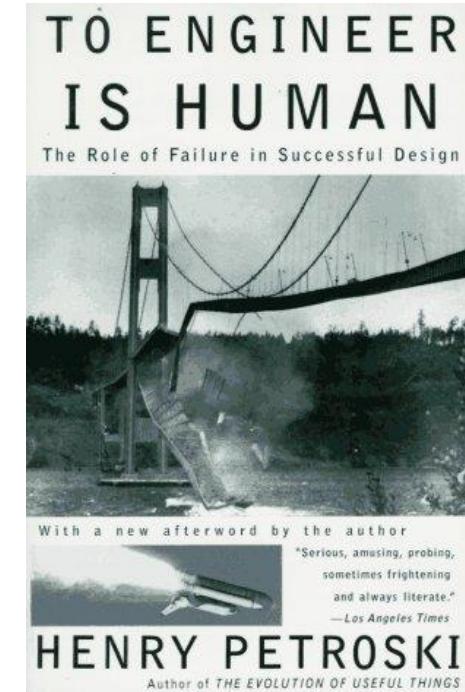
[2] Lighthall, Frederick F. "Launching the space shuttle challenger: disciplinary deficiencies in the analysis of engineering data." *IEEE Transactions on Engineering Management* 38.1 (1991): 63-74.

The Challenger Disaster (1986) : Lessons

Be mindful of your model's **effective range** – predictions made beyond the training data range are **unreliable** and may have serious consequences.

thats why open source models used
HUGE data to train

“Examination of testimony and data from the Challenger launch shows that analysis and reasoning were flawed, not due to personal or organizational failings, but to a professional weakness... **a gap in the education of engineers.**”



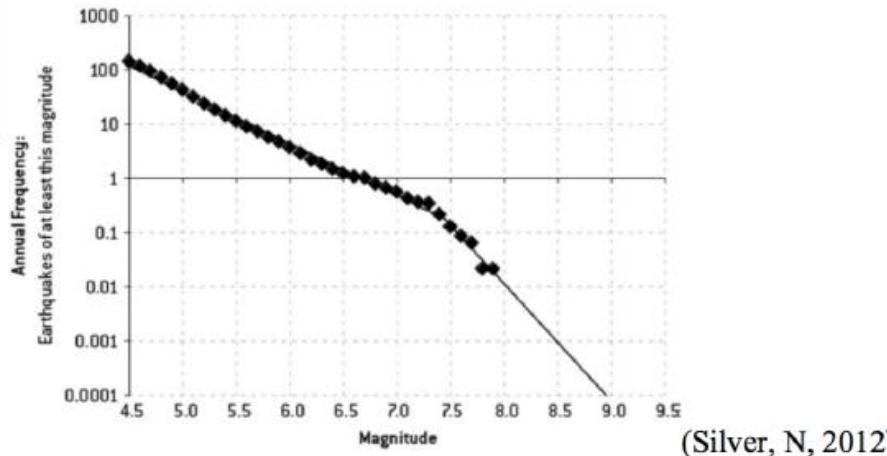
Lighthall, Frederick F. "Launching the space shuttle challenger: disciplinary deficiencies in the analysis of engineering data." *IEEE Transactions on Engineering Management* 38.1 (1991): 63-74.

Petroski, Henry. "To engineer is human: the role of failure in successful design." *Barnes and noble books* (1985).

Example of deployment bias: Fukushima nuclear accident (2011)



FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
CHARACTERISTIC FIT



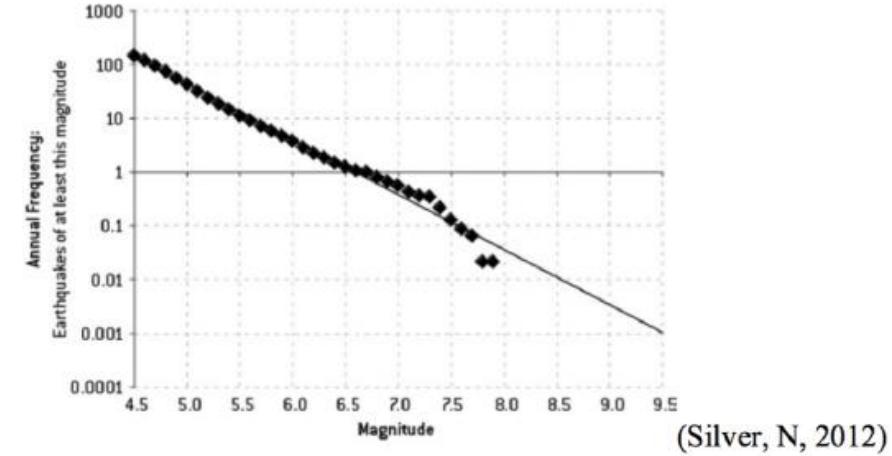
Learning the lessons of Fukushima

Japan's nuclear meltdowns provide valuable lessons for the design of future nuclear powerplants.

David L. Chandler, MIT News Office
August 23, 2011



FIGURE 5-7B: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
GUTENBERG-RICHTER FIT



Reference: N. Silver. The Signal and the Noise: The Art and Science of Prediction. Penguin Books Ltd, 2012.
<https://vyftan.github.io/papers/ee2211book.pdf> P46-47

Takeaway: Reduce deployment bias in practice

- Deploy models only in settings that align closely with their training conditions.
- Regularly track model performance and retrain when data or user behavior changes over time.
- Train models on up-to-date, representative datasets that reflect real-world diversity.

Question to ponder



As engineers and scientists, we often focus on building systems that ‘work’—but what happens when those systems are used in ways we didn’t expect? Should engineers be responsible for how their AI models are used in society, even after deployment?

CNN Business Markets Tech Media Calculators Videos

Uber self-driving car test driver pleads guilty to endangerment in pedestrian death case

By Rebekah Riess and Zoe Sottile, CNN

2 min read · Published 2:27 PM EDT, Sat July 29, 2023



BBC

Uber 'not criminally liable' for self-driving death

6 March 2019

Share  Save 



Reference: <https://www.bbc.com/news/business-43459156>
<https://edition.cnn.com/2023/07/29/business/uber-self-driving-car-death-guilty>

Takeaway

Engineering Responsibility in the Real World

"What happens when systems are used in ways we didn't expect?"

- AI models can cause **harm**, even when technically correct
- Engineers play a role in anticipating risks and misuse

Ask yourself:

- Who could be harmed by this system?
- What assumptions am I making?
- Have I clearly communicated its limits?

Your responsibility does not end at deployment—ethics must guide design, testing, and real-world impact.

Common sources of bias



- **Data bias**

Non-representative/imbalanced training datasets, biased heuristics

- **Model bias**

Simplifying assumptions that don't reflect real-world complexity.
(Sometimes) overly complex or overfitted assumptions that reduce generalizability.

- **Deployment bias**

Mismatch between model training conditions and real-world usage, such as applying it in the wrong setting or with different data than it was trained on.

- **Evaluation bias**

Flawed or inconsistent evaluation methods, such as relying on a single train/test split or using metrics that don't capture real-world performance

Example of evaluation bias

Gender Shades study (by Joy Buolamwini and Timnit Gebru at MIT Media Lab)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



References: Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. PMLR, 2018.

Example of evaluation bias

Suppose we have imbalanced data. Suppose we're building a model to detect diabetes. We've collected data from 1,000 patients, but only 20 are diagnosed with diabetes, while 980 are not.

Use cost matrix
instead of
confusion matrix

Suppose our model gets the following result:

	Predicted P	Predicted N
Actual P	10 (TP)	10 (FN)
Actual N	10 (FP)	970 (TN)

In cases of imbalanced data, accuracy alone can be *misleading*. Instead, we can:

- 1) Evaluate using *precision and recall*;
- 2) Apply cost-sensitive metrics

$$\text{Accuracy: } (10+970)/1000 = 0.98$$

However, despite the high accuracy, the model correctly identified only **half of the actual diabetes patients**.

The model is **good at predicting negatives but struggles with positives**.

Recap: Precision and Recall

When predicting wine variety (red/white), precision and recall (as compared to accuracy) is the most appropriate to evaluate the model's performance?

Confusion Matrix for Binary Classification

	\hat{P} (predicted)	\hat{N} (predicted)	
P (actual)	TP	FN	Recall $TP/(TP+FN)$
N (actual)	FP	TN	
	Precision $TP/(TP+FP)$		Accuracy $(TP+TN)/(TP+TN+FP+FN)$

Recap: Precision and Recall (cont.)

$$Precision = \frac{TP}{TP + FP}$$

Correctly predicted positives

Total predicted positives

		Precision
		Prediction
		+ -
+		TP (True Positive)
-		FN (False Negative)
-		FP (False Positive)
-		TN (True Negative)

$$Recall = \frac{TP}{TP + FN}$$

Correctly predicted positives

Total actual positives

		Recall
		Prediction
		+ -
+		TP (True Positive)
-		FN (False Negative)
-		FP (False Positive)
-		TN (True Negative)

Evaluation Metrics: Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

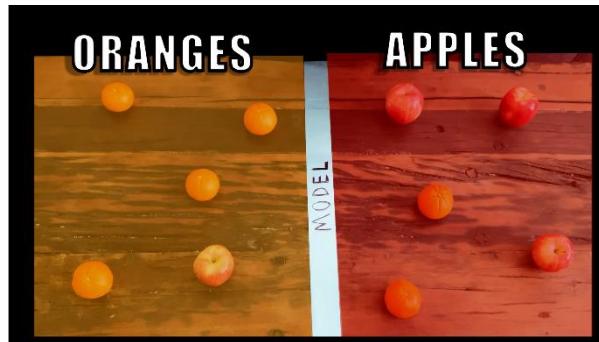


Precision (How accurate the positives are): It measures the quality of model predictions for one particular class. If in a spam detector, precision measures how many of the emails flagged as spam are actually spam. High precision means fewer false alarms.

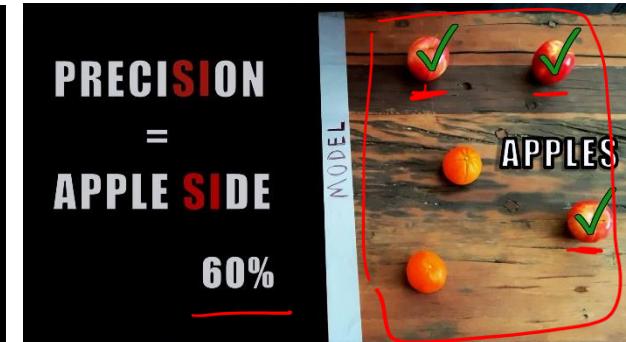
Precision focuses on the **reliability of positive predictions**, important when the cost of false positives is high.



data: 4 apples (positive class),
6 oranges (negative class)



model prediction results



precision (correctly predicted positive/total predicted positive)

$$\frac{3}{5} = 60\%$$

Evaluation Metrics: Recall

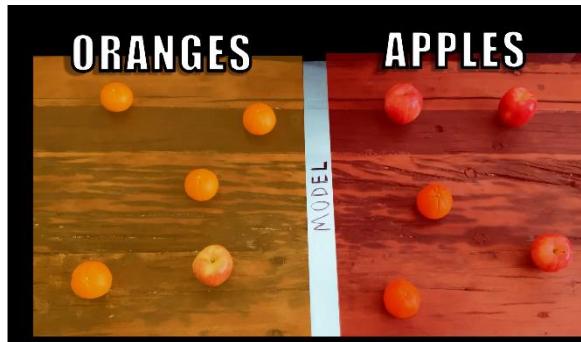
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall (How many actual positives are detected): If in a spam detector, recall measures how many of the actual spam emails are correctly caught. *High recall means fewer missed cases.*

Recall measures the model's **ability to capture actual positives**, crucial when the **cost of false negatives** is high (e.g., cancer screening, terrorist detection).



data: 4 apples (positive class),
6 oranges (negative class)



model prediction results



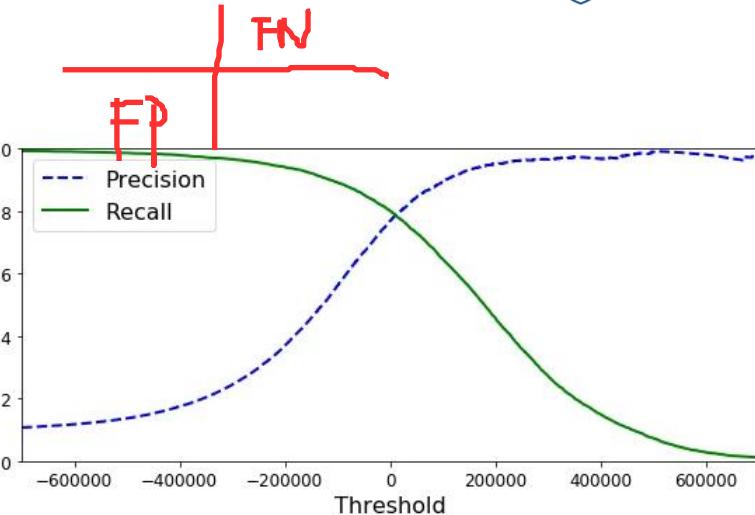
recall (correctly predicted positive/total **actual** positive)

$$\frac{3}{4} = 75\%$$

Evaluation Metrics: Precision and Recall

In practice, there's often a *trade-off*:

- Prioritizing precision reduces false positives (e.g., ensuring only actual spam emails are filtered out). $\text{FP} \downarrow$
- Prioritizing recall reduces false negatives (e.g., catching as much spam as possible, even at the risk of misclassifying some legitimate emails as spam). $\text{FN} \downarrow$



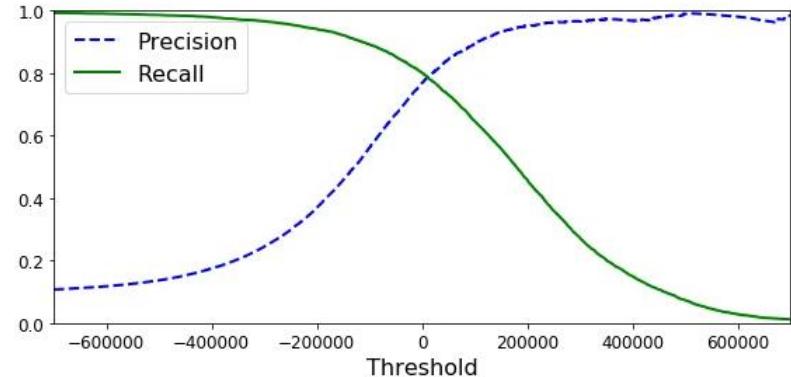
Class practice: Consider these real-life application examples: should we prioritize precision or recall?

- Spam detection Prioritise Precision - dont want real emails to be deleted as spam
- Cancer diagnosis Prioritise Recall - dont want to miss out any detection of cancer
- Seismic risk assessment for buildings Prioritise Recall - dont want to miss out any real risk
- Defect detection in manufacturing Prioritise Recall - dont want to miss out any real defects
- Autonomous vehicle pedestrian detection Prioritise Recall - dont want to miss out any pedestrian

Evaluation Metrics: Precision and Recall

In practice, there's often a *trade-off*:

- **Prioritizing precision** reduces **false positives** (e.g., ensuring only actual spam emails are filtered out).
- **Prioritizing recall** reduces **false negatives** (e.g., catching as much spam as possible, even at the risk of misclassifying some legitimate emails as spam).



Class practice: Consider these real-life application examples: should we prioritize precision or recall?

- Spam detection (prioritize precision)
- Cancer diagnosis (prioritize recall)
- Seismic risk assessment for buildings (prioritize recall)
- Defect detection in manufacturing (prioritize recall)
- Autonomous vehicle pedestrian detection (prioritize recall)

Reduce evaluation bias using precision and recall

Suppose we have imbalanced data. Suppose we're building a model to detect diabetes. We've collected data from 1,000 patients, but only 20 are diagnosed with diabetes, while 980 are not. Suppose our model gets the following result:

	Predicted P	Predicted N
Actual P	10 (TP)	10 (FN)
Actual N	10 (FP)	970 (TN)

Evaluate the model's performance using precision and recall:

$$\text{Precision} = 10/10 = 0.5$$

$$\text{Recall} = 10/10 = 0.5$$

$$\text{Accuracy: } (10+970)/1000 = 0.98$$

F1 score (ranges from 0 to 1): a metric that measures a model's accuracy in terms of both precision and recall.

0: the model completely fails to identify the positive class.

1: the model perfectly predicts all positives.

$$F1 = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 0.5$$

either precision or recall is 0, which only happens when TP is 0

$$\begin{aligned} F1 &= \frac{P \cap N}{P + N} \\ &= \frac{10}{10 + 10} \\ &= 0.5 \end{aligned}$$

recall = 1
precision also = 1

Cost sensitive evaluation metrics

Cost Matrix for Binary Classification

	\hat{P} (predicted)	\hat{N} (predicted)
P (actual)	$C_{p,p} * TP$	$C_{p,n} * FN$
N (actual)	$C_{n,p} * FP$	$C_{n,n} * TN$

Total cost:

$$C_{p,p} * TP + \\ C_{p,n} * FN + \\ C_{n,p} * FP + \\ C_{n,n} * TN$$

Main idea of cost matrix:

It assigns different penalties for different types of outcomes.

Higher penalties for more severe errors.

Usually, $C_{p,p}$ and $C_{n,n}$ are set to 0; $C_{n,p}$ and $C_{p,n}$ may and may not be equal.

Cost sensitive evaluation metrics example

Evaluate the model's performance using cost matrix (conceptual)

	Predicted P	Predicted N
Actual P	10 (TP)	10 (FN)
Actual N	10 (FP)	970 (TN)

Cost Matrix for Binary Classification

	\hat{P} (predicted)	\hat{N} (predicted)
P (actual)	$C_{p,p} * TP$	$C_{p,n} * FN$
N (actual)	$C_{n,p} * FP$	$C_{n,n} * TN$

Total cost:
 $C_{p,p} * TP +$
 $C_{p,n} * FN +$
 $C_{n,p} * FP +$
 $C_{n,n} * TN$

Original accuracy: $(10+970)/1000 = 0.98$

But if we care more about false negatives (FN), we can assign a higher penalty,
e.g., $C_{p,n} = 100$.

The cost-sensitive accuracy drops to $(10+970)/(10 + 10*100+10+970) = 0.49$

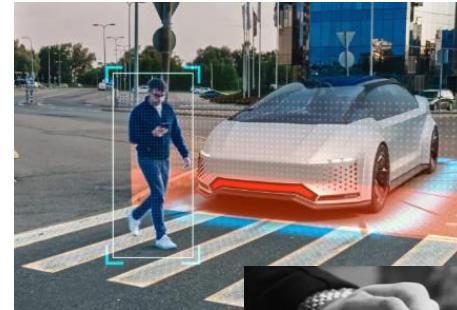
Class practice: $C_{n,p} > C_{p,n}$ or $C_{n,p} < C_{p,n}$

Cost Matrix for Binary Classification

	\hat{P} (predicted)	\hat{N} (predicted)
P (actual)	$C_{p,p} * TP$	$C_{p,n} * FN$
N (actual)	$C_{n,p} * FP$	$C_{n,n} * TN$

Total cost:

$$C_{p,p} * TP + \\ C_{p,n} * FN + \\ C_{n,p} * FP + \\ C_{n,n} * TN$$



Pedestrian detection in self-driving car systems

<

Credit card fraud detection

<

Automated candidate screening systems for universities/companies

>

Security screening system at customs/airports

<

Another example of evaluation bias

Suppose we are building a dog/cat classifier. The training and test sets are as shown below. What potential problems could affect model evaluation in this scenario?



Train set



Test set

This could potentially cause **evaluation bias due to an unrepresentative test set**
Relying on a single train/test split can lead to evaluation bias

Reduce evaluation bias using K-fold cross-validation

K-Fold Cross-Validation

Problem: One train–test split may give misleading results, as the chose test set may not represent the whole dataset

Solution: K-Fold Cross-Validation

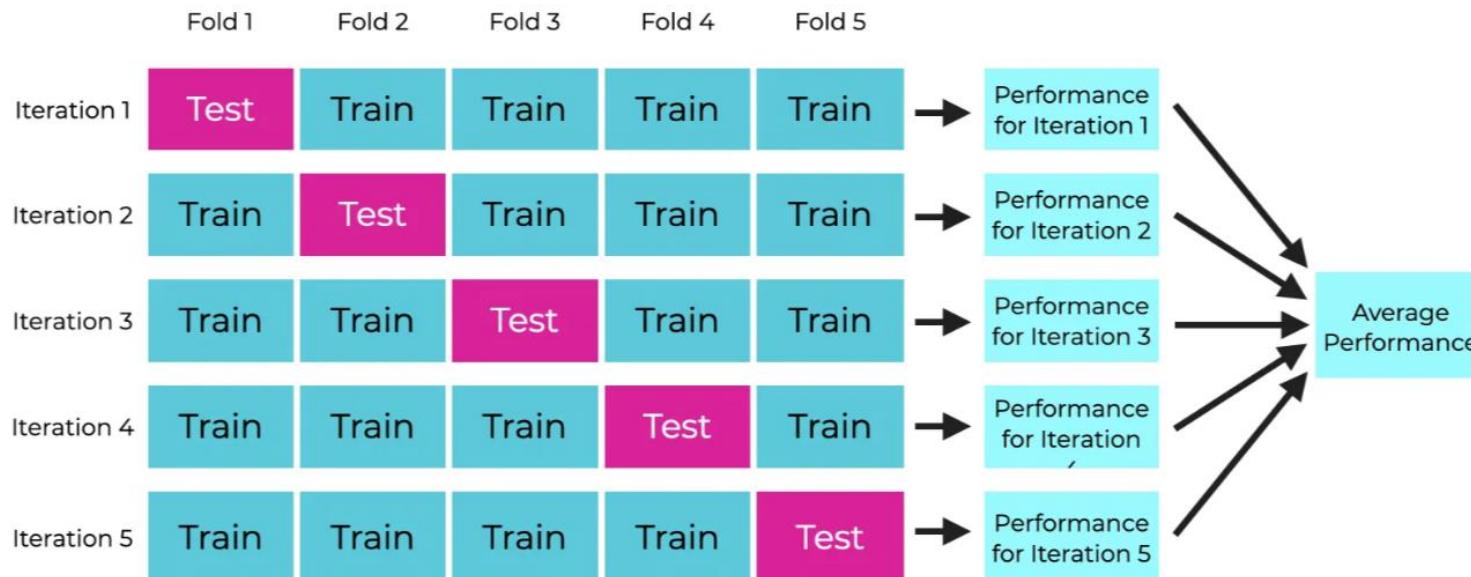
Split the dataset into k equal folds.

Train on $(k-1)$ folds, test on the remaining fold.

Repeat k times, each fold serving once as the test set.

Average the results for a more reliable performance estimate.

Step 3.1: Within each fold, if we have n parameter/model candidates, we will train n models, and we check their validation performance.



Takeaway: Engineering practices to reduce AI bias



1. Examine your data thoroughly

Use exploratory data analysis to uncover imbalances, missing groups, or skewed distributions that may introduce bias early on.

2. Design models with real-world complexity in mind

Avoid overly simplistic assumptions—include relevant features like context, fairness factors, and user diversity.

3. Perform robust validation and lifecycle monitoring

Conduct stratified cross-validation across diverse subpopulations and deployment conditions. Implement continuous monitoring to detect distribution shifts and feedback loops that may amplify bias over time.

4. Know where your model works (and doesn't)

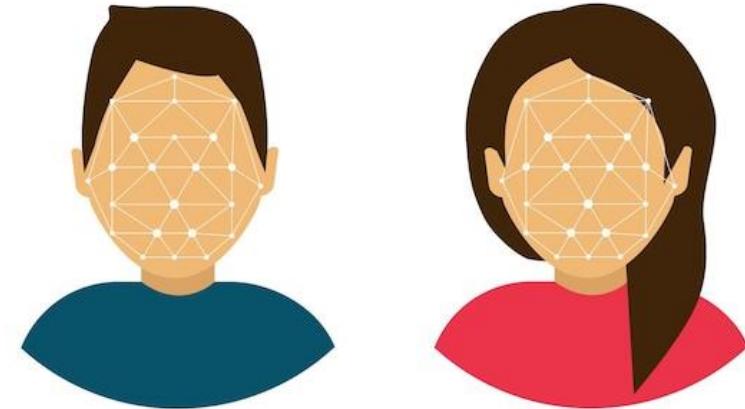
Be clear about the model's effective range; avoid deploying it beyond its design scope.

5. Evaluate your model using k-fold cross validation, or cost sensitive metrics whenever necessary.

Your responsibility does not end at deployment—ethics must guide design, testing, and real-world impact.

Agenda

- Generative AI & its social impact
- Bias in AI
 - Type of common biases
 - Ways to reduce bias
- Privacy concerns



Data privacy issues in LLMs



Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History

A publicly accessible database belonging to DeepSeek allowed full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams with highly sensitive information.

Gal Nagli
January 29, 2025
3 minute read



Forbes

DeepSeek Data Leak Exposes 1 Million Sensitive Records

By Lars Daniel, Contributor. Lars Daniel covers digital evidence and forensics... [Follow Author](#)

Published Feb 01, 2025, 08:27pm EST, Updated Feb 02, 2025, 11:15am EST

Share Save Comment 0

Source: <https://www.forbes.com/sites/larsdaniel/2025/02/01/deepseek-data-leak-exposes--1000000-sensitive-records/>
<https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak>

Efforts to address data privacy issues in LLMs



- **Anonymized training data**
Remove personal identifiers before model training.
- **Differential privacy**
Add noise to protect individual data in learning/output.
- **Access controls & filters**
Restrict sensitive content during model use.
- **On-device / federated learning**
Keep data local; train without direct access.

Data protection against deepfakes



Source: <https://www.youtube.com/shorts/0dTmRCHdXT8>

Privacy issues on social media

Forbes

Clearview AI—Controversial Facial Recognition Firm—Fined \$33 Million For ‘Illegal Database’

By [Robert Hart](#), Former Staff. Robert Hart is a former London-based Forbes senior reporter.

Published Sep 03, 2024, 07:54am EDT

[Share](#) [Save](#) [Comment 0](#)

TOPLINE

Controversial U.S. facial recognition company Clearview AI, reportedly embraced U.S. government and law enforcement agencies, has been [fined](#) more than \$30 million by the Netherlands' data protection watchdog on Tuesday for building “an illegal database” containing billions of faces taken from social media and the internet.



FORBES' VIDEO WILL PLAY AFTER THIS AD

AD

RECEIVE VOUCHERS

- \$800 CDC Vouchers
- \$400 Climate Vouchers
- \$600 - \$800 SG60 Vouchers

*Subject to individual's or household's elig.

Continue watching [Miley Cyrus Explains Why Growing...](#) after the ad

FORBES' FEATURED VIDEO



Source: <https://www.forbes.com/sites/roberthart/2024/09/03/clearview-ai-controversial-facial-recognition-firm-fined-33-million-for-illegal-database/>

Privacy protection on social media

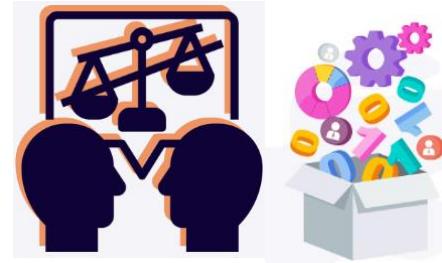
	Original image		Perturbed image	
	Human judgment	Machine judgment	Human judgment	Machine judgment
People present?	✓	✓	✓	✗
Text present?	✓	✓	✓	✗
Contain focused object?	✓	✓	✓	✓
Aesthetic?	✓	✓	✓	✓
Pleasant?	✓	✓	✓	✓

Shen, Zhiqi, Shaojing Fan, Yongkang Wong, Tian-Tsong Ng, and Mohan Kankanhalli. "Human-imperceptible privacy protection against machines." In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1119-1128. 2019.

Summary: Societal impact and more



Enhancing content moderation
for information integrity



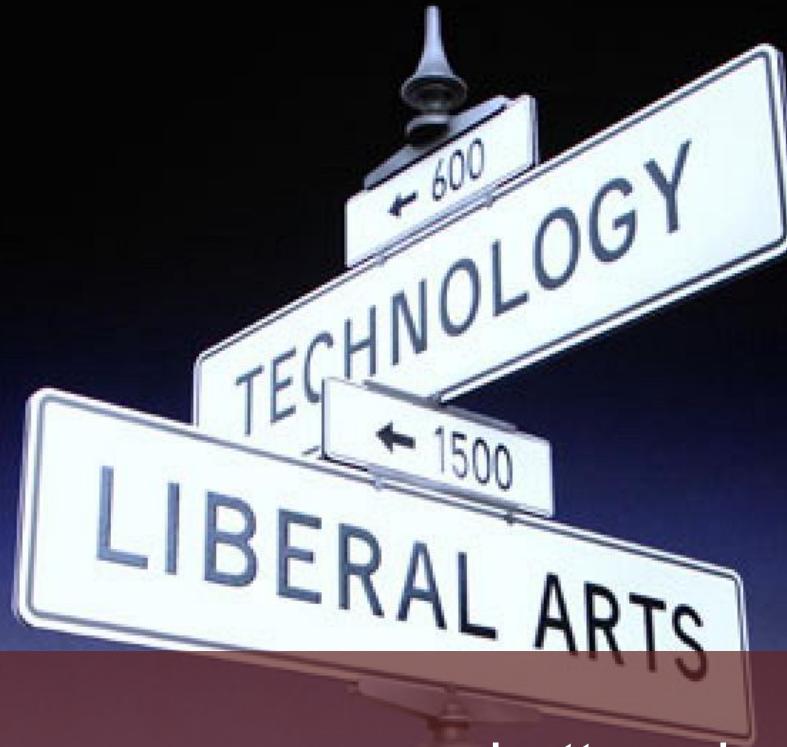
Bias mitigation via diverse data
and fairness-aware models



Minimize risks and
maximize AI benefits



Privacy-preserving methods in
data generation and transmission



AI, created by humans, serves us better when guided by human insights.

**NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING**

**ACADEMIC YEAR 2025-2026
SEMESTER 1**

EE2213: Introduction to AI

TUTORIAL 8: L12-L13

Question 1 (Lecture 12)

“If a car won’t start and battery is old, then replacing the battery might fix the issue”. What type of knowledge does this statement belong to?

- (a) Declarative
- (b) Procedural
- (c) Structural
- (d) Heuristic
- (e) Meta

Answer: (d). Experience-based guideline: replacing the battery might fix the issue. The solution is not guaranteed to be correct/optimal.

Question 2 (Lecture 12)

Which of the following statements represent meta knowledge?

- (a) “The Eiffel Tower is in Paris.” (**Declarative**)
- (b) “The steps to solve a quadratic equation are: 1) rearrange terms, 2) apply quadratic formula.” (**Procedural**)
- (c) “If the reasoning process takes too long, switch from model checking to resolution.”
- (d) “The ground is wet. It possibly rained last night.” (**Heuristic**)

Answer: (c). It’s about choosing a reasoning strategy.

Question 3 (Lecture 12)

In which scenario is probabilistic knowledge representation most useful?

- (a) When all facts are certain and must be represented with strict rules. (**Logical**)
- (b) When handling noisy sensor data with uncertain outcomes.
- (c) When representing fixed hierarchical relationships, like a taxonomy of animals. (**Structured**)
- (d) When encoding concepts as high-dimensional vectors for natural language processing. (**Distributed**)

Answer: (b)

Question 4 (Lecture 12)

What type of reasoning does propositional logic use?

- (a) Deductive
- (b) Inductive

(c) Abductive

Answer: (a). Deterministic in nature.

Question 5 (Lecture 13)

A single proposition symbol can represent a compound proposition.

(a) True

(b) False

Answer: (b). A proposition symbol can only represent an atomic proposition.

Question 6 (Lecture 13)

Which of the following is/are a well-formed formula in propositional logic?

(a) $\alpha = \gamma \Leftrightarrow \delta$

(b) $\neg\neg\alpha$, provided that α is a well-formed formula

(c) $\neg\alpha \vee \beta \Rightarrow \beta$, provided that α and β are both well-formed formulas.

(d) $\alpha \wedge \sigma \wedge \neg\beta$

(e) $\beta \vDash \alpha$, provided that α and β are both well-formed formulas.

Answer: (b)(c)

Question 7 (Lecture 13)

Which of the following is/are correct?

(a) $(A \wedge B) \vDash (A \vee B)$

(b) $\alpha \vDash \beta$ if and only if $(\alpha \Rightarrow \beta)$ is a tautology

(c) $\alpha \equiv \beta$ if and only if $(\alpha \Leftrightarrow \beta)$ is a tautology

(d) $(A \wedge B) \wedge (A \Leftrightarrow B)$ is satisfiable

Answer: (a)(b)(c)(d)

(a) Since “if $(A \wedge B)$ is true, $(A \vee B)$ will always be true”, $(A \wedge B) \vDash (A \vee B)$

(b) Need to check: ① if $\alpha \vDash \beta$, $(\alpha \Rightarrow \beta)$ is always true; and ② if $(\alpha \Rightarrow \beta)$ is always true, $\alpha \vDash \beta$.

① If $\alpha \vDash \beta$, that said, whenever α is true, β is true. If α is true, β is true. Then $(\alpha \Rightarrow \beta)$ is true; if α is false, no matter what truth-value β takes, $(\alpha \Rightarrow \beta)$ is true. Therefore, $(\alpha \Rightarrow \beta)$ is always true if $\alpha \vDash \beta$.

② if $(\alpha \Rightarrow \beta)$ is always true, that said, only 3 pairs of truth value assignments

for (α, β) are allowed: (true, true), (false, true), (false, false). In these 3 models, β is always true if α is true, which means $\alpha \models \beta$.

(c) Similar to (b). Need to check: ① if $\alpha \equiv \beta$, $(\alpha \Leftrightarrow \beta)$ is always true; and ② if $(\alpha \Leftrightarrow \beta)$ is always true, $\alpha \equiv \beta$.

(d) Need to check that $(A \wedge B) \wedge (A \Leftrightarrow B)$ is true in at least one model.

$(A \wedge B)$ and $(A \Leftrightarrow B)$ should both be true if we want $(A \wedge B) \wedge (A \Leftrightarrow B)$ to be true. Then, A and B should both be true if we want $(A \wedge B)$ to be true; A and B should both be true or false if we want $(A \Leftrightarrow B)$ to be true. Then, we find out one model which is $\{A=\text{true}, B=\text{true}\}$ that satisfies it.

Question 8 (Lecture 13)

On an island, Knights always tell the truth and Knaves always lie.

Person A says: “Person B is a Knave.”

Person B says: “We are both Knights.”

- (i) Construct the knowledge base using propositional logic. Hint: you may use “sympy” library.
- (ii) Determine who is Knight and who is Knave.

Answer: See TUT8_Q8.ipynb

(i) Proposition Symbols: A: “Person A is a Knight”; B: “Person B is a Knight”

$$KB = \{A \Leftrightarrow \neg B, B \Leftrightarrow (A \wedge B)\}$$

(ii) Check if $KB \models A$, $KB \models B$. Since $KB \models A$, A is a Knight. Since KB does not entail B, we further check if $KB \models \neg B$. Since $KB \models \neg B$ (i.e., B is always false whenever KB is true), B is a Knave.

Question 9 (Lecture 13)

Consider a simple 3x3 Wumpus World, where at least one Wumpus and at least one pit are located in some rooms. In the rooms directly (not diagonally) adjacent to the Wumpus, the agent will perceive a Stench. In the squares directly adjacent to a pit, the agent will perceive a Breeze.

(i) Construct a simple knowledge base using propositional logic for this 3x3 Wumpus World. You may use “sympy” library. Use the following proposition symbol for each $[x,y]$ location:

$$P_{x,y}: \text{There is a pit in } [x, y]$$

$$W_{x,y}: \text{There is a Wumpus in } [x, y]$$

$$B_{x,y}: \text{The agent perceives a breeze in } [x, y]$$

$S_{x,y}$: The agent perceives a stench in $[x,y]$

- (ii) Suppose the agent is now in location [1,1], where neither Breeze nor Stench is perceived.
Assume there is neither pit nor Wumpus in location [1,1]. Is there a Wumpus in location [2, 2]? Is there a pit in location [2, 2]?
- (iii) Continuing from part (ii), are adjacent cells of location [1,1] safe? Suppose the agent moves to location [2,1] and perceive a Breeze but no Stench. Are adjacent cells of location [2,1] safe?

Answer: See [TUT8_Q9.ipynb](#)

- (ii) Not sure.
- (iii) Adjacent cells of location [1,1] are safe;
Adjacent cells of location [2,1]: [1,1] safe; [3,1] no Wumpus, not sure about pit; [2,2] no Wumpus, not sure about pit.

**NATIONAL UNIVERSITY OF SINGAPORE
DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING**

**ACADEMIC YEAR 2025-2026
SEMESTER 1**

EE2213: Introduction to AI

TUTORIAL 9: L14

Question 1 (Lecture 14)

True or false: While the underrepresentation of specific groups in training data is a common contributor to data bias, addressing this imbalance alone is always sufficient to ensure equitable and unbiased performance of an AI model in real-world deployment scenarios.

- (a) True
- (b) False

Ans: (b)

Question 2 (Lecture 14)

Which of the following best describes model bias?

- a) A biased dataset that leads to skewed predictions
- b) A mismatch between training and deployment settings
- c) A model that makes incorrect or overly complex assumptions, either oversimplifying or overfitting the data
- d) A model trained using synthetic data

Ans: (c)

Question 3 (Lecture 14)

True or false: Model bias can occur even if the training data is perfectly balanced.

- (a) True
- (b) False

Ans: (a)

Question 4 (Lecture 14)

Which of the following practices most effectively mitigates model bias at its source?

- a) Employing rigorous data standardization techniques to ensure all input features are scaled uniformly across the dataset
- b) Maximizing the complexity of the chosen model architecture to allow for the learning of highly intricate and nuanced patterns within the training data
- c) Consistently enriching the training dataset with new, contextually diverse, and representative examples, followed by iterative model retraining
- d) Implementing post-hoc fairness algorithms to adjust the model's outputs only after initial predictions are generated, without altering the training process.

Ans: (c)

Question 5 (Lecture 14)

Which of the following is a key cause of deployment bias?

- a) Using outdated training data
- b) Selecting an overly simple model
- c) Applying the model in a context very different from its training conditions
- d) Ignoring missing values during data preprocessing

Ans: (c)

Question 6 (Optional – For Exploration)

An engineering firm is developing a new Large Language Model (LLM) application for use in confidential industrial design and simulation workflows. This LLM will process sensitive proprietary data and potentially intellectual property during its operation and refinement. The firm is acutely aware of the data privacy challenges inherent in LLMs.

Consider the following common privacy-enhancing techniques employed in LLM development and deployment:

- Data Anonymization
- Differential Privacy
- Access Controls and Filters
- Federated Learning / On-device processing

Select any two of these techniques. For each chosen technique, provide a detailed discussion of the key engineering **trade-offs** that must be evaluated during its design and implementation. Your analysis should encompass considerations such as:

- The impact on the LLM's performance (e.g., accuracy, model utility, inference speed).
- The user experience for end-users interacting with the LLM (e.g., ease of use, perceived

privacy, functionality limitations).

- The system complexity introduced for development, integration, and maintenance.
- The guaranteed level of privacy assurance versus the feasibility and implementation cost.

Sample Answer:

Selected Techniques: Data anonymization & Federated Learning

1. Data Anonymization

Trade-offs:

Data anonymization removes or masks identifiable information before model training, protecting individual and organizational privacy. However, excessive anonymization can strip valuable context, reducing the LLM's accuracy and limiting its ability to learn domain-specific nuances. From a user's standpoint, anonymization strengthens confidence in data protection but may result in less precise or relevant responses. Engineering-wise, implementing robust anonymization pipelines adds preprocessing complexity and requires continuous updates to address re-identification risks as data and models evolve.

2. Federated Learning / On-device Processing

Trade-offs:

Federated learning allows model training or fine-tuning to occur locally on user devices or secure company servers, ensuring sensitive data never leaves its source. This approach significantly enhances privacy but increases system complexity, as it requires distributed coordination, version control, and secure communication protocols. Performance may also be affected by heterogeneous device capabilities or slower update cycles. For users, it offers high perceived privacy and ownership of data but may come with reduced responsiveness or higher latency in some workflows.

Takeaway

When deploying privacy-enhancing techniques for LLMs, engineers must carefully balance trade-offs. No free lunch in data privacy: improving privacy often impacts model performance, user experience, and system complexity. Selecting the right approach depends on the context, desired privacy guarantees, and practical feasibility.