Artificial Intelligence

Topic 11

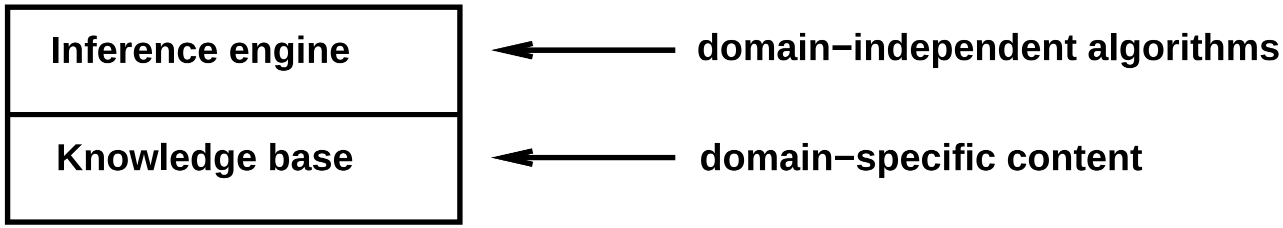
Logical Agents

Reading: Russell and Norvig, Chapter 7, Sections 1–4

Outline

- ♢ Knowledge-based agents
- ♢ Wumpus world
- ♢ Logics
- ♢ Propositional (Boolean) logic
  - – syntax
  - – semantics
- ♢ Models
- ♢ Entailment
- ♢ Equivalence, validity, satisfiability

Knowledge bases



Knowledge base = set of sentences in a **formal** language

Declarative approach to building an agent (or other system):
TELL it what it needs to know

Then it can ASK itself what to do—answers should follow from the KB

Agents can be viewed at the knowledge level
i.e., **what they know**, regardless of how implemented

Or at the implementation level
i.e., data structures in KB and algorithms that manipulate them

A simple knowledge-based agent

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

   TELL(KB, MAKE-PERCEPT-SENTENCE( percept, t))
   action ← ASK(KB, MAKE-ACTION-QUERY(t))
   TELL(KB, MAKE-ACTION-SENTENCE(action, t))
   t ← t + 1
   return action
```

The agent must be able to:

Represent states, actions, etc.

Incorporate new percepts

Update internal representations of the world

Deduce hidden properties of the world

Deduce appropriate actions

Wumpus World PEAS description

Performance measure

gold +1000, death -1000

-1 per step, -10 for using the arrow

Environment

Squares adjacent to wumpus are smelly

Squares adjacent to pit are breezy

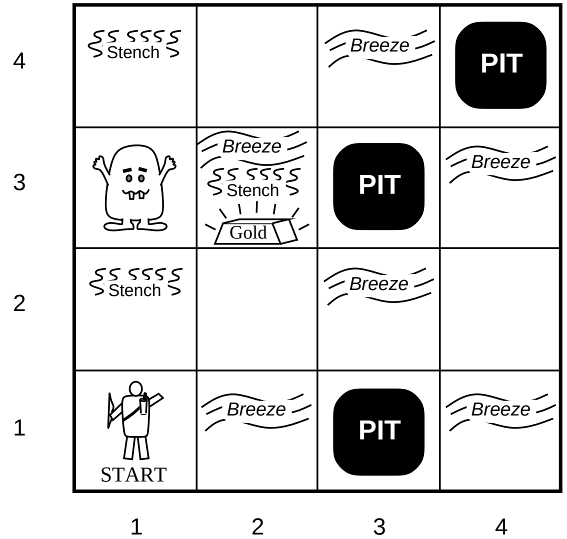
Glitter iff gold is in the same square

Shooting kills wumpus if you are facing it

Shooting uses up the only arrow

Grabbing picks up gold if in same square

Releasing drops the gold in same square



Actuators Left turn, Right turn,
Forward, Grab, Release, Shoot

Sensors Breeze, Glitter, Smell

Wumpus world characterization

Observable?? No—only local perception

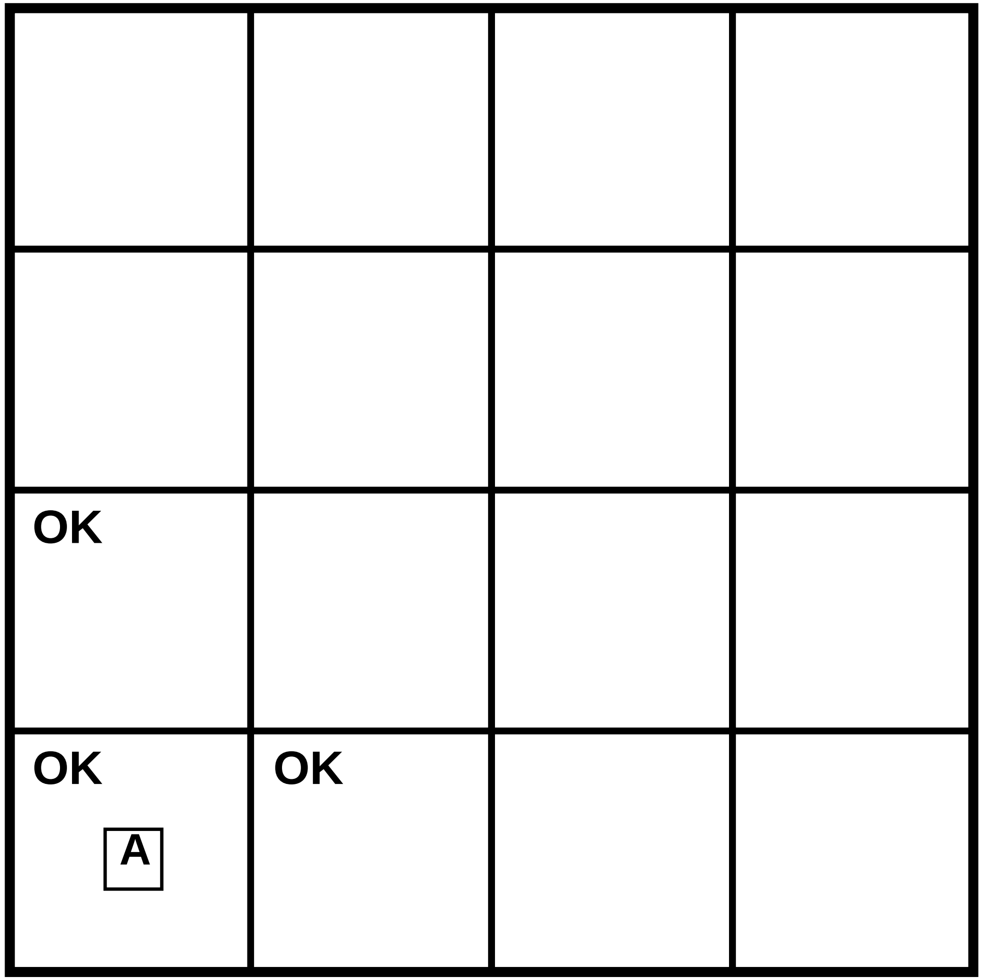
Deterministic?? Yes—outcomes exactly specified

Static?? Yes—Wumpus and Pits do not move

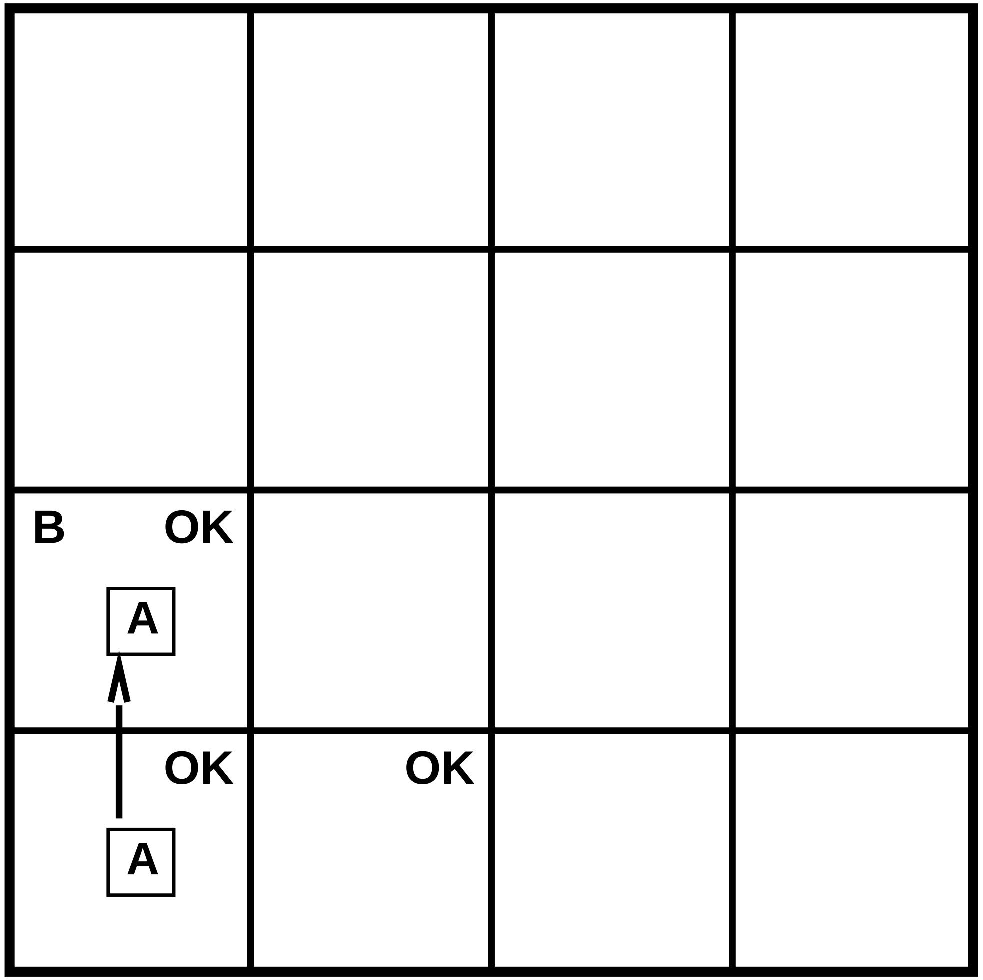
Discrete?? Yes

Single-agent?? Yes—Wumpus is essentially a natural feature

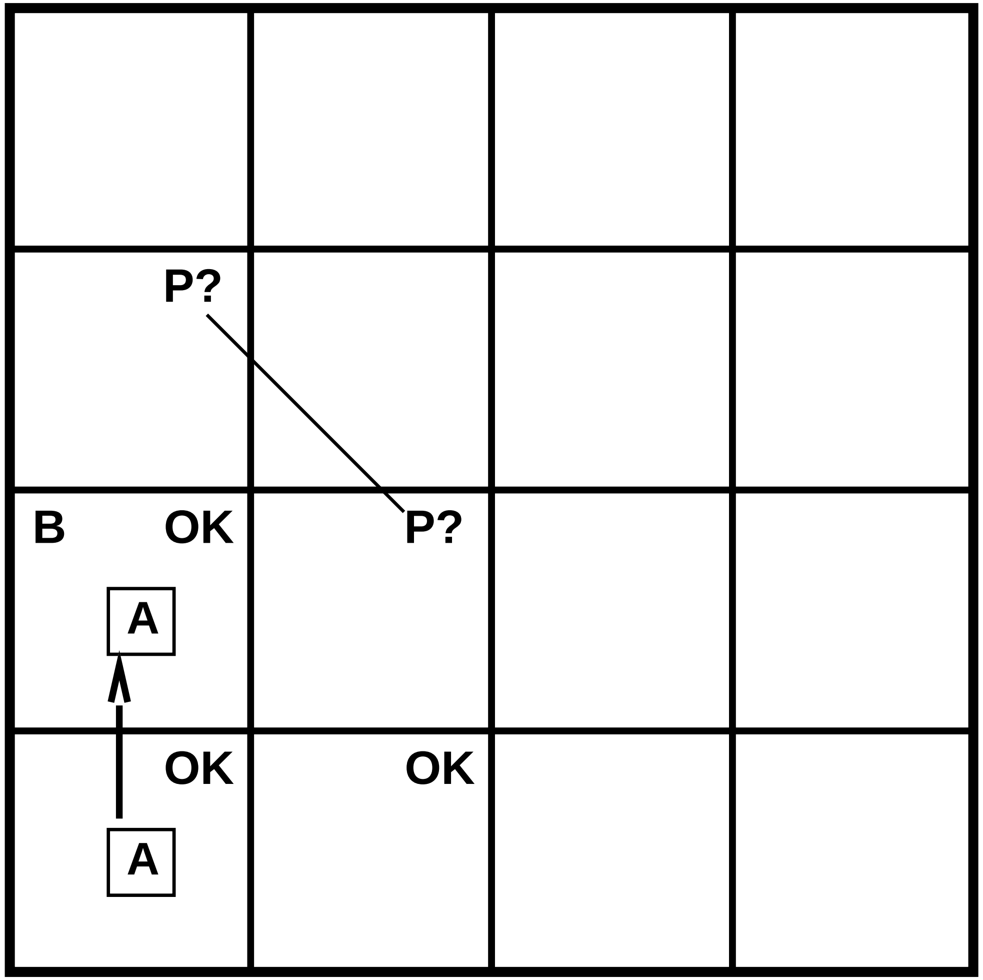
Exploring a wumpus world (different world to previous picture)



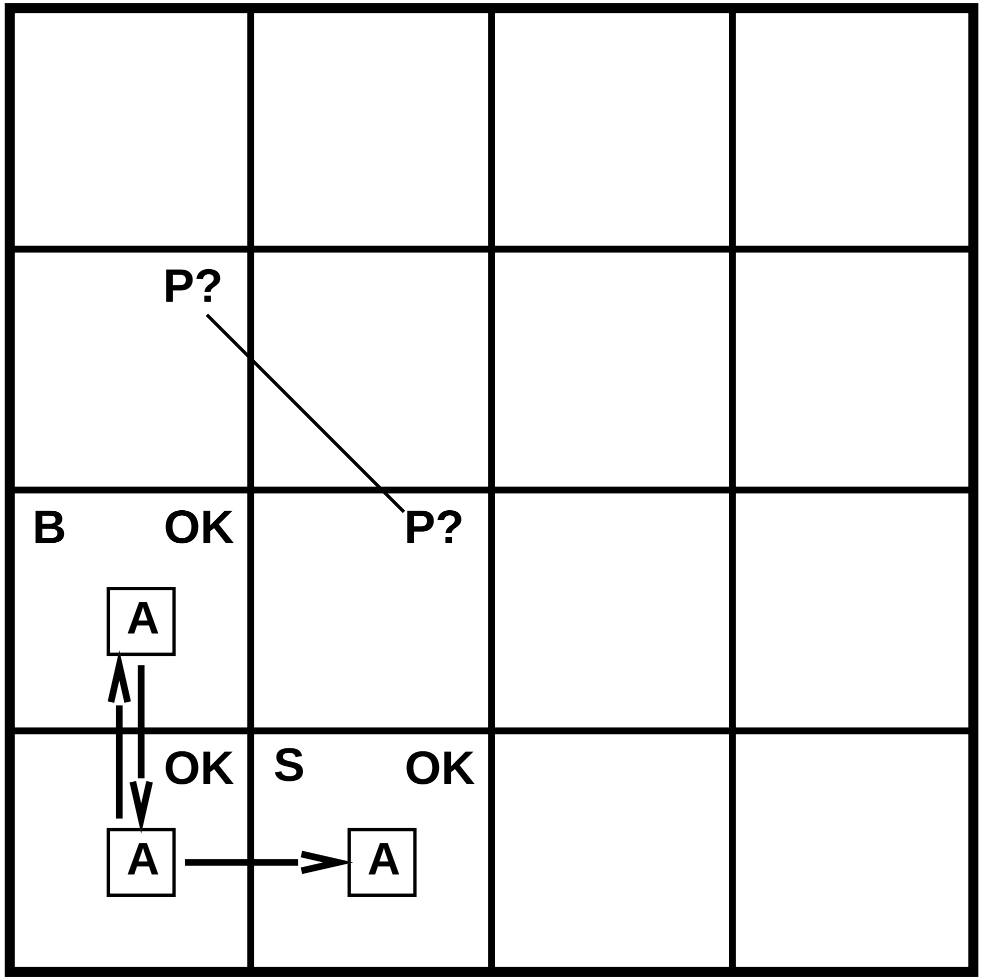
Exploring a wumpus world



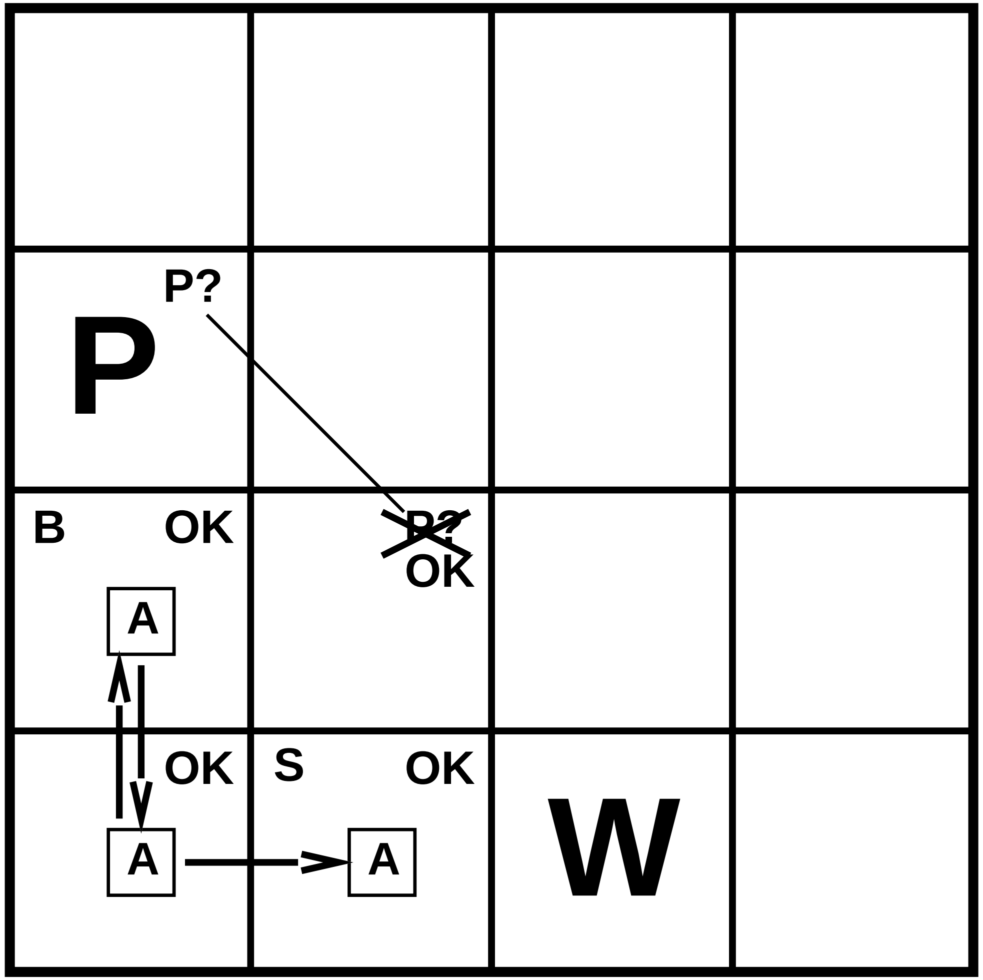
Exploring a wumpus world



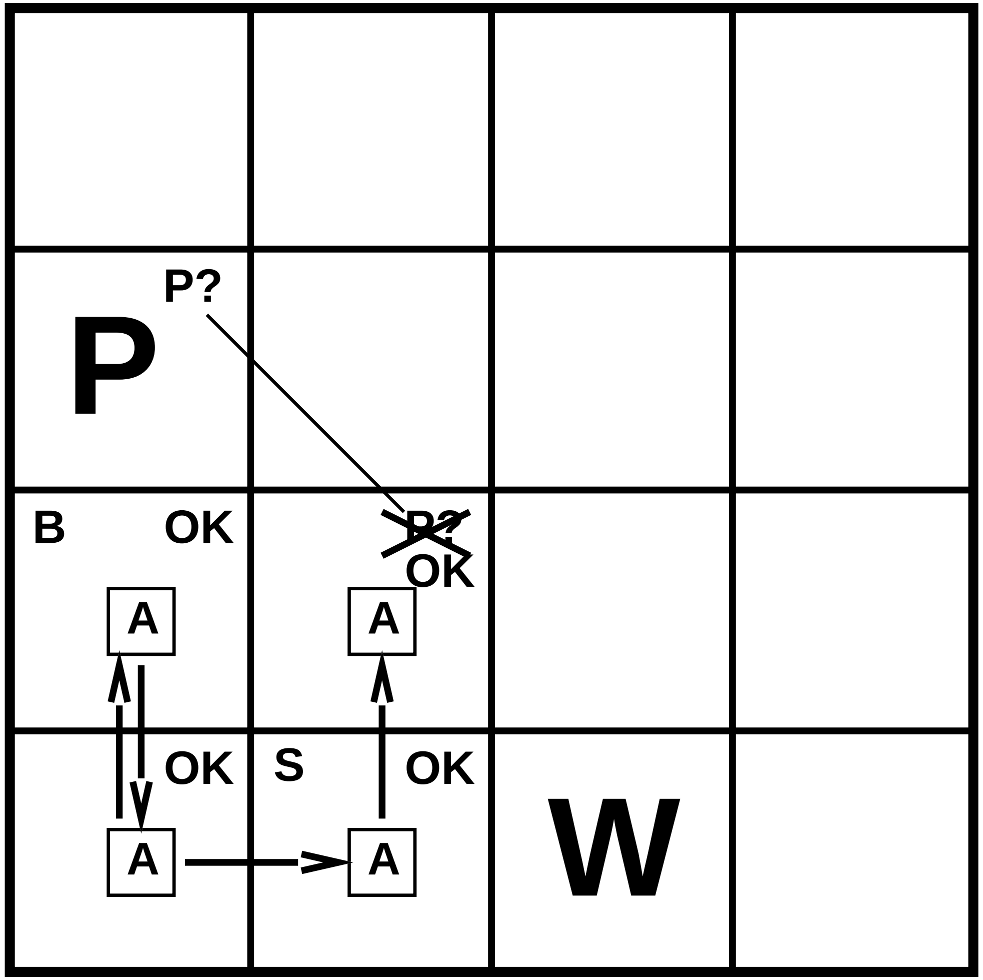
Exploring a wumpus world



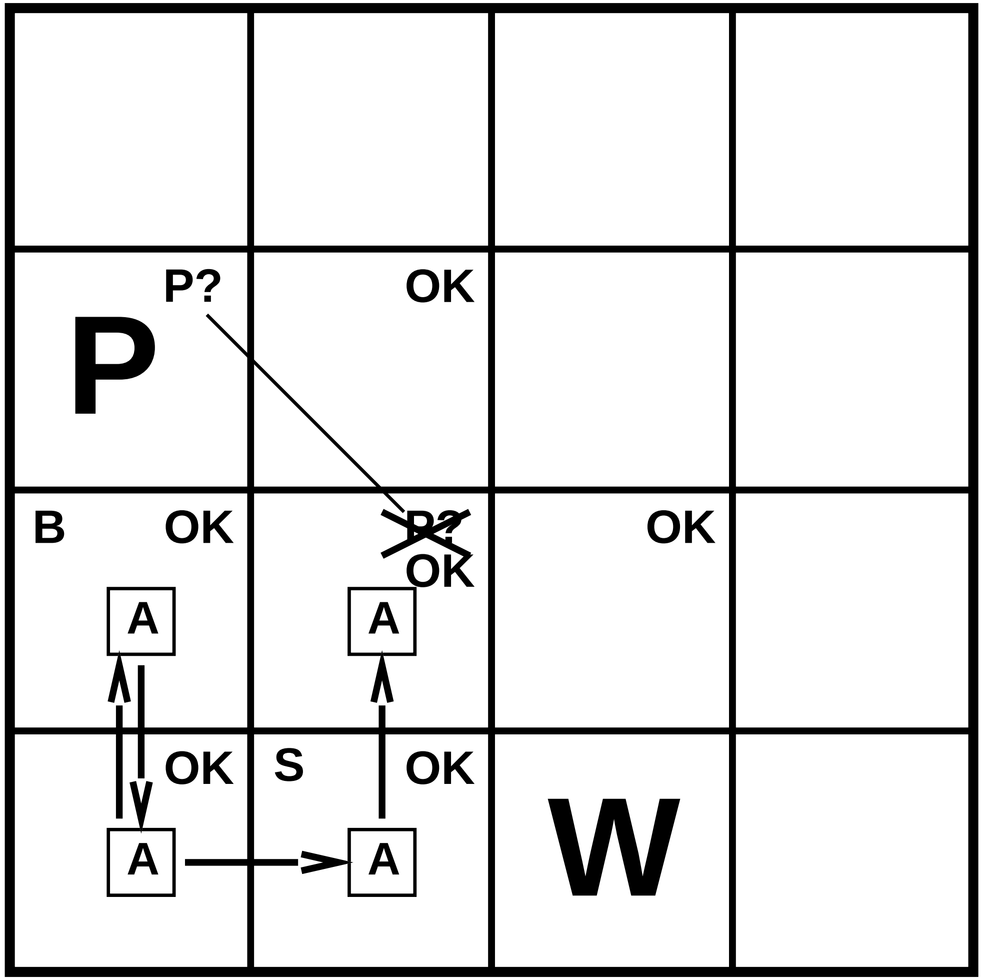
Exploring a wumpus world



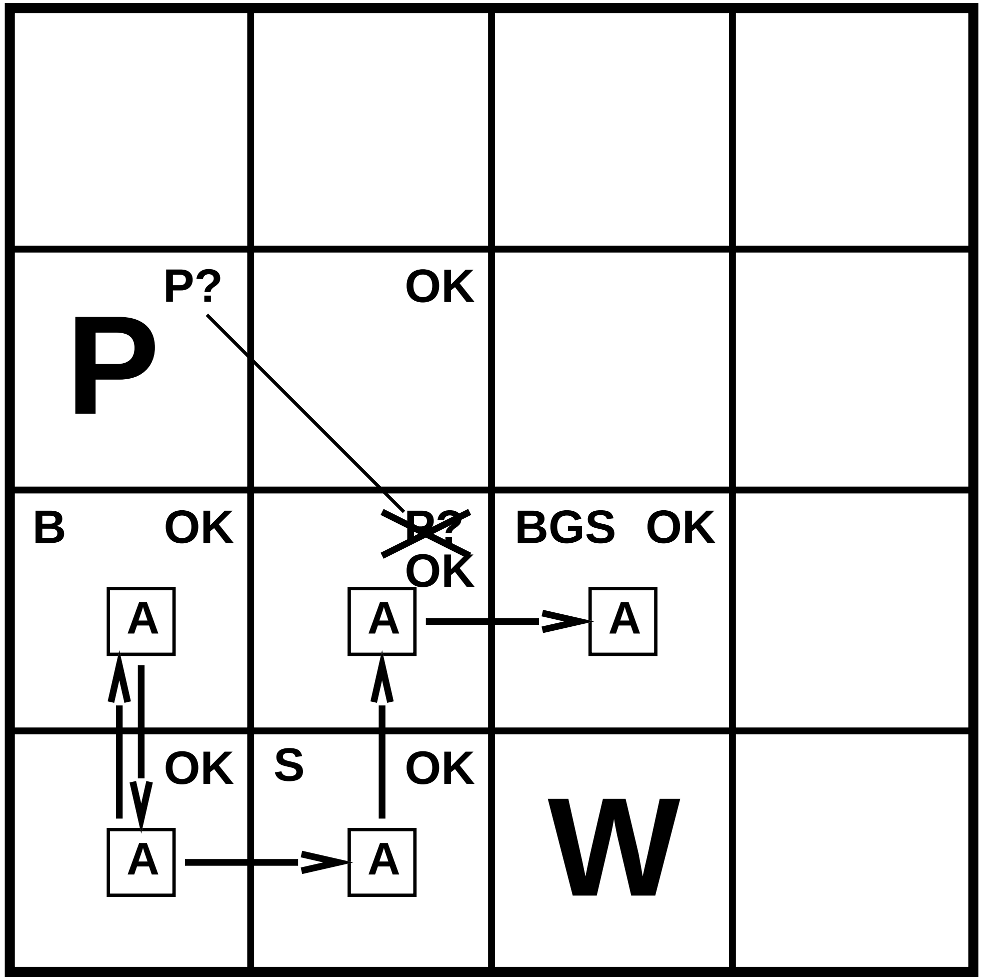
Exploring a wumpus world



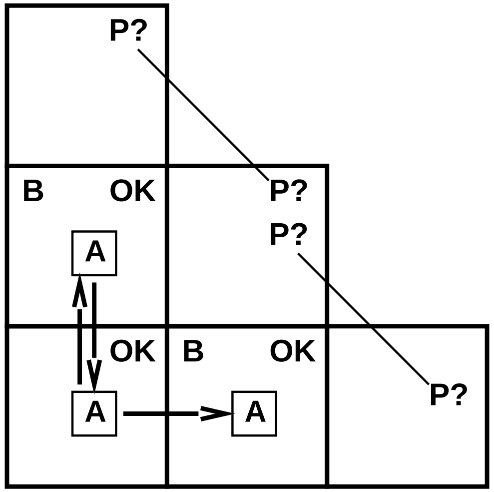
Exploring a wumpus world



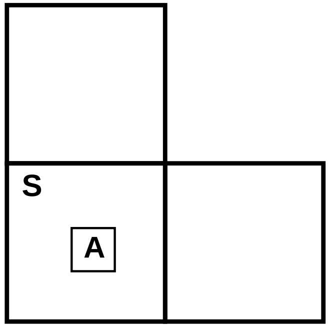
Exploring a wumpus world



Other tight spots



Breeze in (1,2) and (2,1)
$\Rightarrow$ no safe actions



Smell in (1,1)
$\Rightarrow$ cannot move

Can use a strategy of coercion:

shoot straight ahead

wumpus was there $\Rightarrow$ dead $\Rightarrow$ safe

wumpus wasn't there $\Rightarrow$ safe

Logics

Logics are formal languages for representing information
such that conclusions can be drawn

Syntax defines the sentences in the language

Semantics define the "meaning" of sentences;
i.e., define truth of a sentence in a world

E.g., the language of arithmetic

$x + 2 \geq y$ is a sentence; $x2 + y >$ is not a sentence

$x + 2 \geq y$ is true iff the number $x + 2$ is no less than the number y

$x + 2 \geq y$ is true in a world where $x = 7, \; y = 1$
$x + 2 \geq y$ is false in a world where $x = 0, \; y = 6$

Propositional logic: Syntax

Propositional logic is the simplest logic—illustrates basic ideas

The proposition symbols P_1, P_2 etc are sentences

If S is a sentence, $\neg S$ is a sentence (negation ["not"])

If S_1 and S_2 are sentences, $S_1 \wedge S_2$ is a sentence (conjunction ["and"])

If S_1 and S_2 are sentences, $S_1 \vee S_2$ is a sentence (disjunction ["or"])

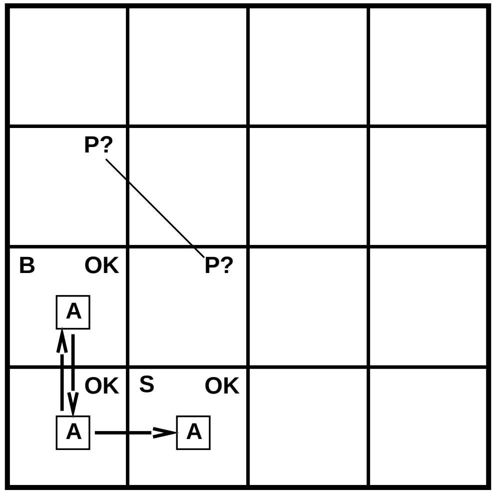
If S_1 and S_2 are sentences, $S_1 \;\Rightarrow\; S_2$ is a sentence (implication ["if-then"])

If S_1 and S_2 are sentences, $S_1 \;\Leftrightarrow\; S_2$ is a sentence (biconditional ["if-and-only-if"])

Wumpus world sentences

Let $P_{i,j}$ be true if there is a pit in $[i, j]$.
Let $B_{i,j}$ be true if there is a breeze in $[i, j]$.

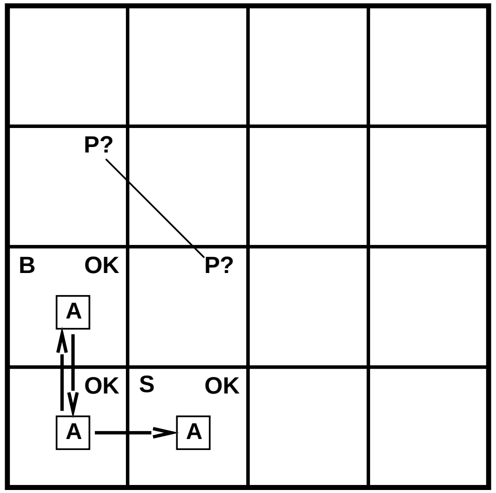
$$\neg P_{1,1}, \neg P_{1,2}, \neg P_{2,1}$$
$$\neg B_{1,1}, B_{1,2}, \neg B_{2,1}$$



Wumpus world sentences

Let $P_{i,j}$ be true if there is a pit in $[i, j]$.
Let $B_{i,j}$ be true if there is a breeze in $[i, j]$.

$$\neg P_{1,1}, \neg P_{1,2}, \neg P_{2,1}$$
$$\neg B_{1,1}, B_{1,2}, \neg B_{2,1}$$

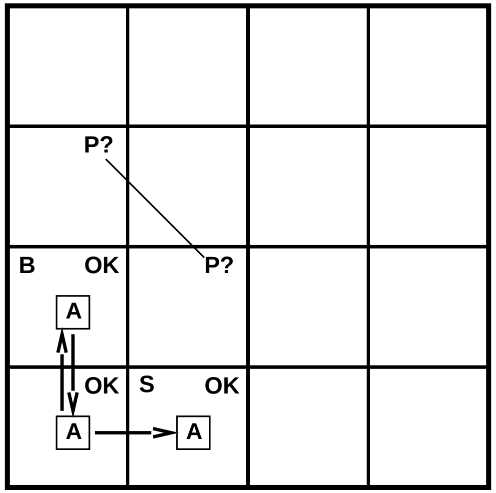


"Pits cause breezes in adjacent squares"

$$P_{1,3} \Rightarrow (B_{1,2} \land B_{2,3} \land B_{1,4})$$
$$P_{2,2} \Rightarrow (B_{1,2} \land B_{2,3} \land B_{3,2} \land B_{2,1})$$

Can we conclude $\neg P_{2,2}$?, $P_{1,3}$?

Wumpus world sentences



Let $P_{i,j}$ be true if there is a pit in $[i,j]$.
Let $B_{i,j}$ be true if there is a breeze in $[i,j]$.

$$\neg P_{1,1}, \neg P_{1,2}, \neg P_{2,1}$$
$$\neg B_{1,1}, B_{1,2}, \neg B_{2,1}$$

"A square is breezy **if and only if** there is an adjacent pit"

$$B_{1,2} \quad \Leftrightarrow \quad (P_{1,1} \vee P_{2,2} \vee P_{1,3})$$
$$B_{2,1} \quad \Leftrightarrow \quad (P_{1,1} \vee P_{2,2} \vee P_{3,1})$$

Can we conclude $\neg P_{2,2}$?, $P_{1,3}$?

What do we mean by "conclude"?

Propositional logic: Semantics

Specifies true/false for each proposition symbol

E.g. $\begin{array}{ccc} P_{1,2} & P_{2,2} & P_{3,1} \\ true & true & false \end{array}$

With n symbols, 2^n possible truth assignments can be enumerated automatically.

Rules for evaluating truth of compound sentences:

$\neg S$	is true iff	S	is false		
$S_1 \wedge S_2$	is true iff	S_1	is true **and**	S_2	is true
$S_1 \vee S_2$	is true iff	S_1	is true **or**	S_2	is true
$S_1 \Rightarrow S_2$	is true iff	S_1	is false **or**	S_2	is true
i.e.,	is false iff	S_1	is true **and**	S_2	is false
$S_1 \Leftrightarrow S_2$	is true iff	$S_1 \Rightarrow S_2$	is true **and**	$S_2 \Rightarrow S_1$	is true

Truth tables for connectives

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

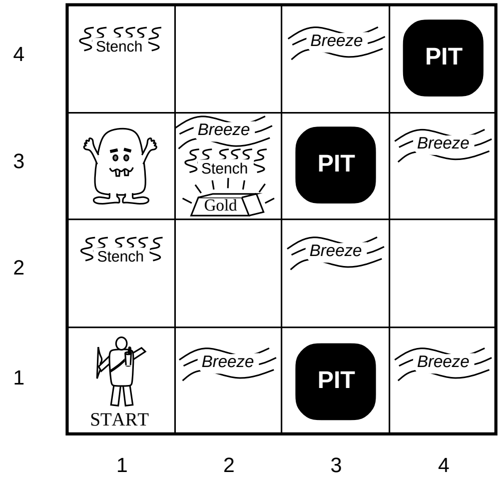
Simple recursive process evaluates an arbitrary sentence, e.g.,
$\neg P_{1,2} \wedge (P_{2,2} \vee P_{3,1}) = true \wedge (false \vee true) = true \wedge true = true$

Models

Logicians typically think in terms of models, which are formally structured worlds with respect to which truth can be evaluated

We say m is a model of a sentence α, or m *satisfies* α, if α is true in m

For example, the model illustrated here satisfies the sentences:
$\alpha_1 =$ there is a breeze in square [4,1]
and:
$\alpha_2 =$ every square adjacent to a pit has a breeze
but not the sentence:
$\alpha_3 =$ every square adjacent to a breeze has a pit



We say a model m satisfies a set of sentences KB iff m satisfies *all* sentences $\alpha \in KB$.

Entailment

Entailment means that one thing **follows from** or is a *logical consequence* of another:

$$KB \models \alpha$$

Knowledge base KB entails sentence α
if and only if
α is true whenever KB is true

E.g., the KB containing "the Dockers won" and "the Eagles lost" entails "Either the Dockers won or the Eagles won" (irrespective of the other teams)

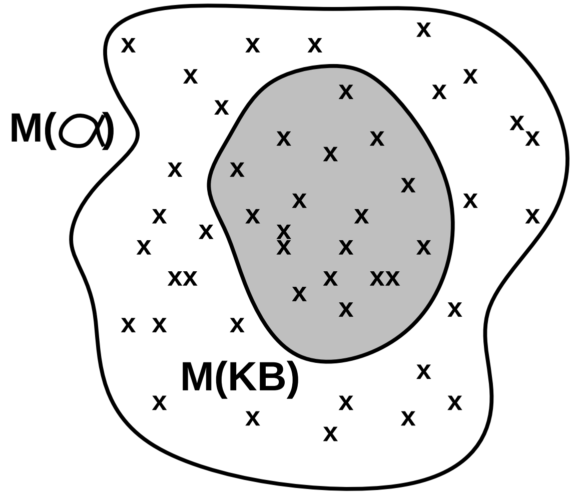
E.g., $x + y = 4$ entails $4 = x + y$

Entailment is a relationship between sentences (i.e., **syntax**) that is based on **semantics**

Models and Entailment

Let $M(\alpha)$ be the set of all models satisfying α, and $M(KB)$ be the set of all models satisfying KB.

Then $KB \models \alpha$ if and only if $M(KB) \subseteq M(\alpha)$



E.g. KB = Dockers won and Eagles lost
α = Dockers won

Assume there are 4 teams in the league.

How many possible outcomes (assuming no draws)?

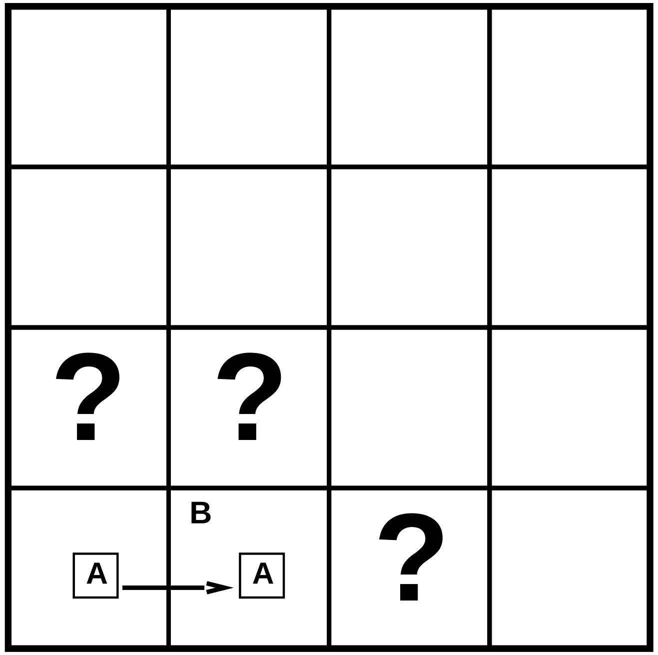
Which results satisfy KB? Which results satisfy α?

Does $KB \models \alpha$?

Entailment in the wumpus world

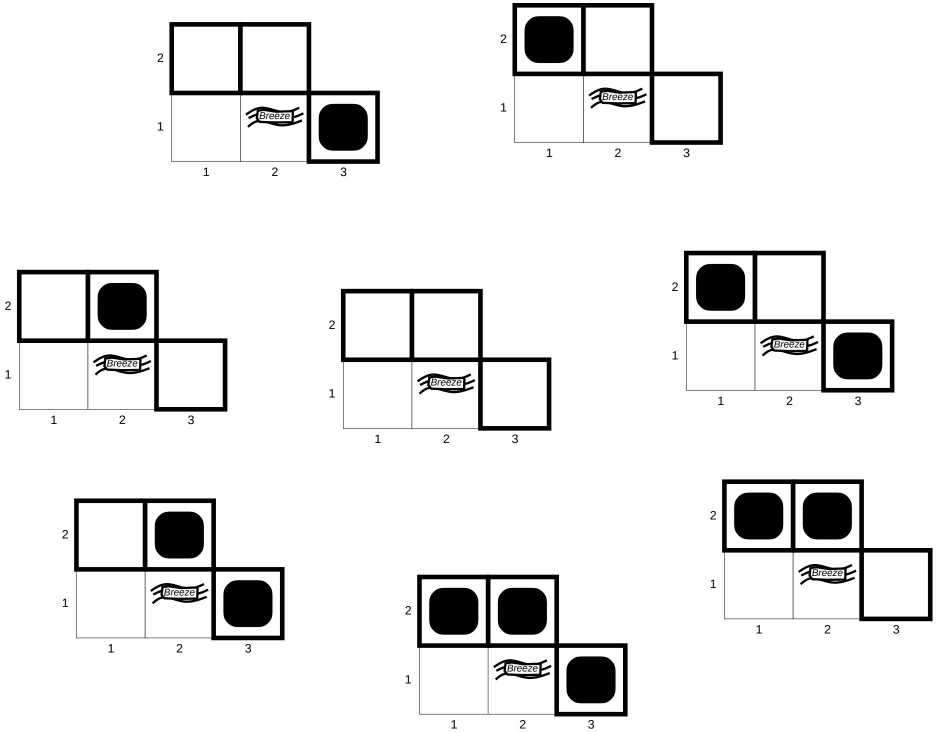
Situation after detecting nothing in [1,1], moving right, breeze in [2,1]

Consider possible models for ?s
assuming only pits

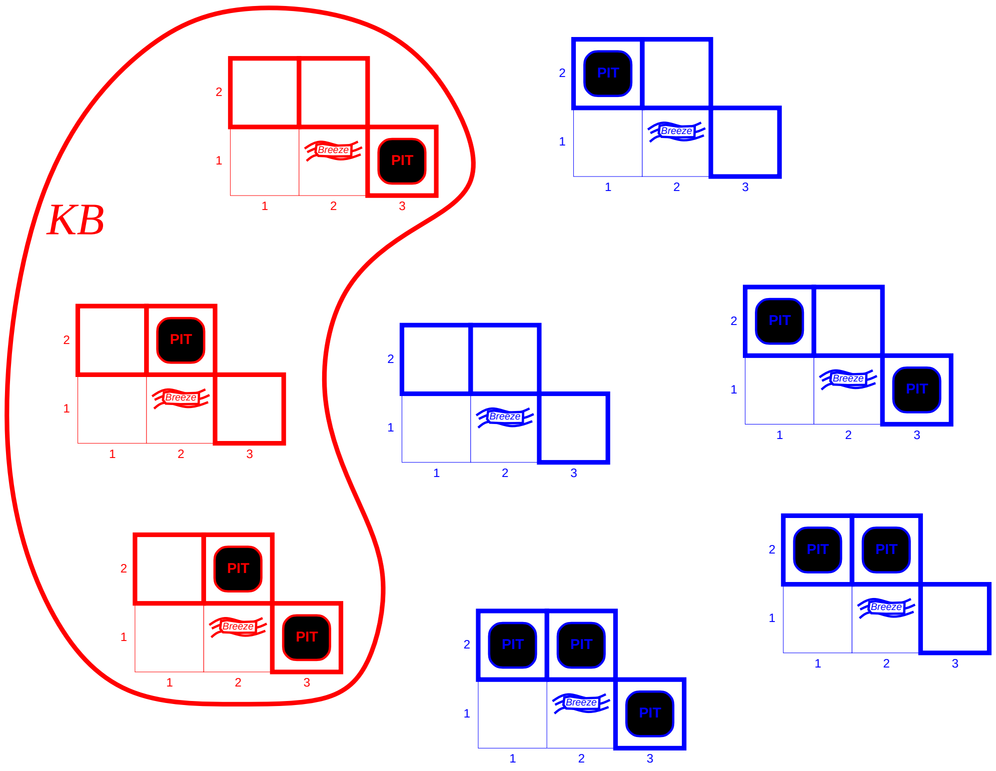


3 Boolean choices $\Rightarrow$ 8 possible models

Wumpus models

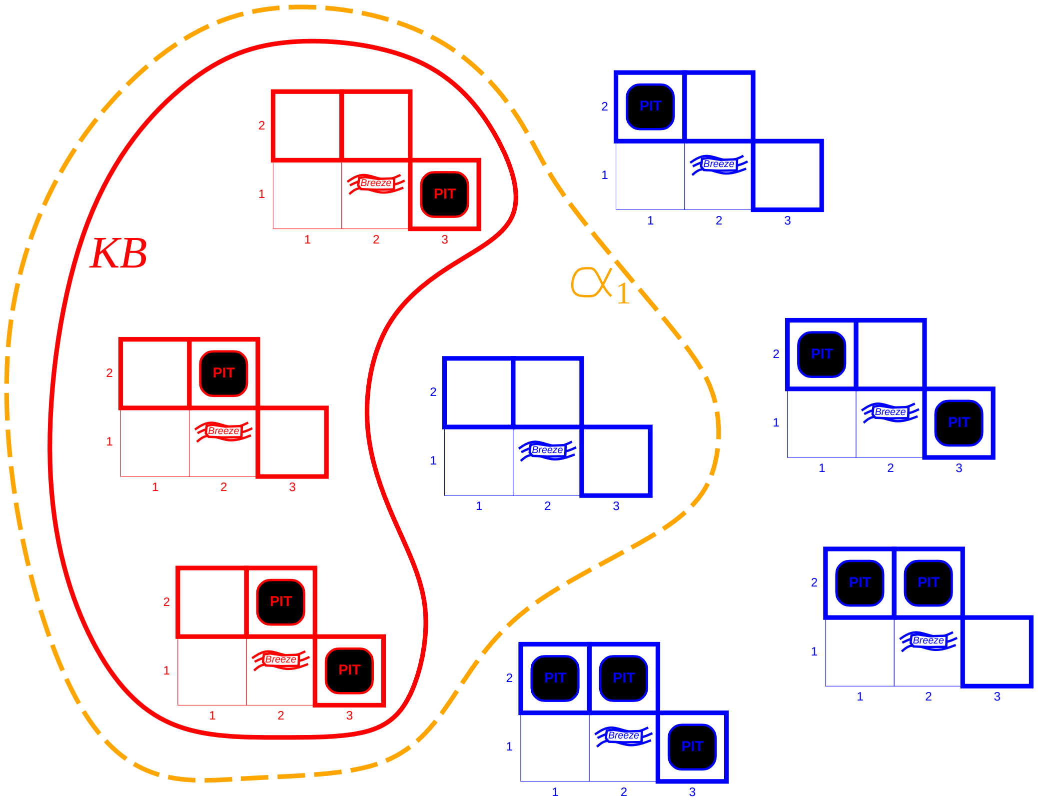


Wumpus models



KB = wumpus-world rules + observations

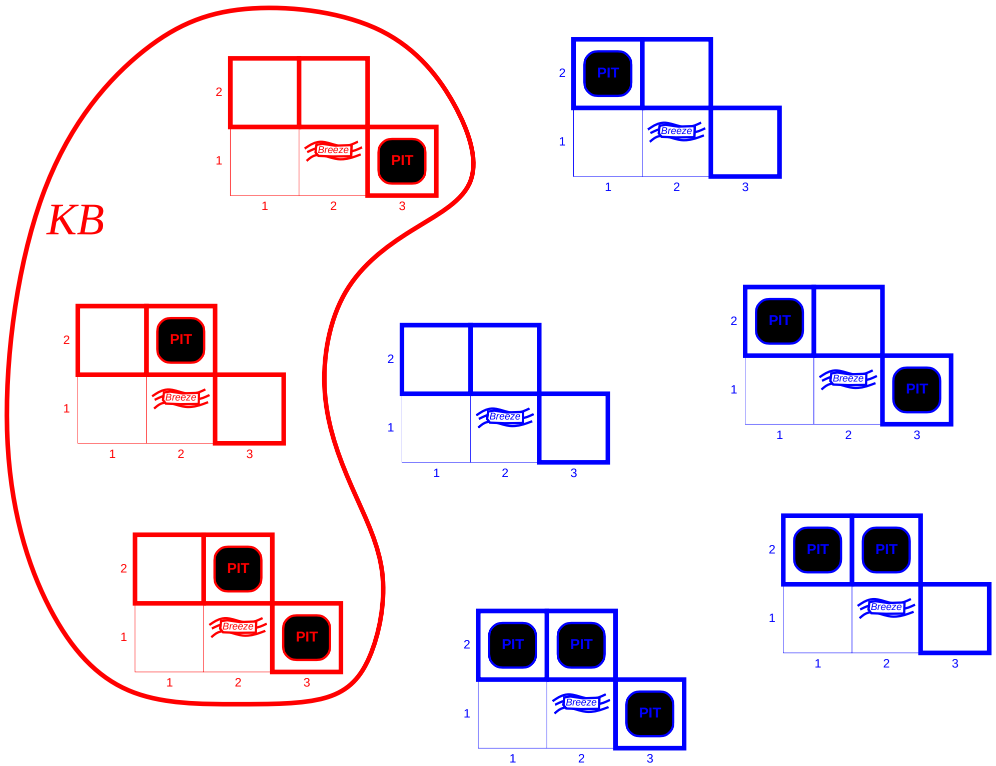
Wumpus models



KB = wumpus-world rules + observations

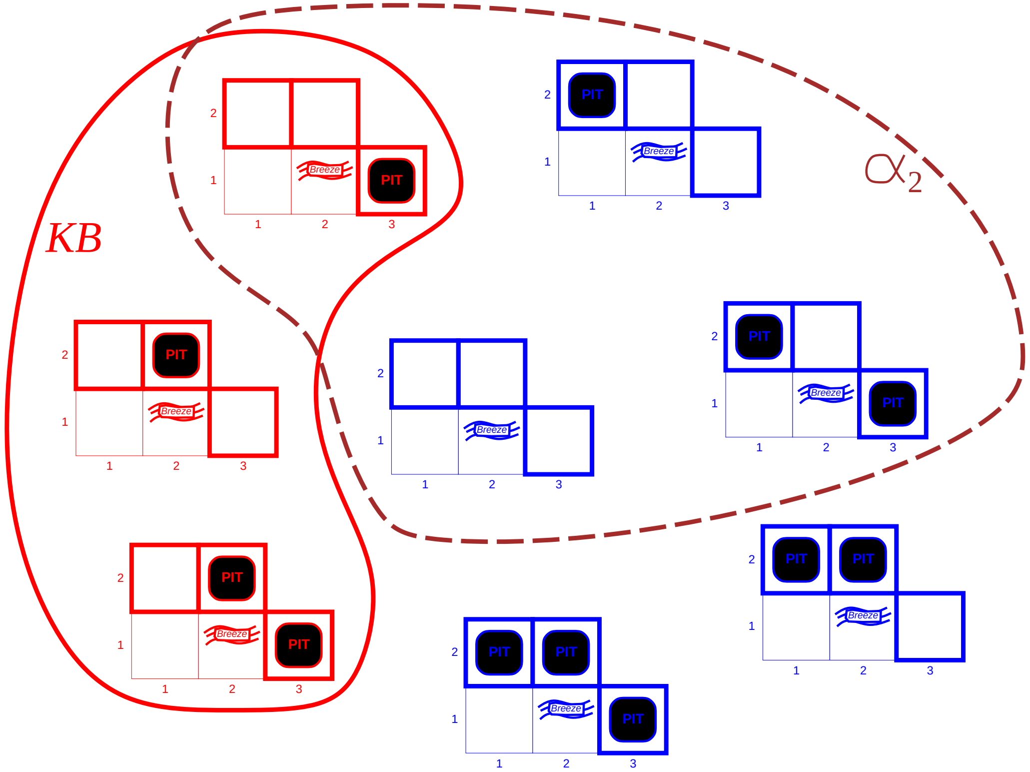
α_1 = "[1,2] is safe", $KB \models \alpha_1$, proved by model checking

Wumpus models



KB = wumpus-world rules + observations

Wumpus models



KB = wumpus-world rules + observations

α_2 = "[2,2] is safe", $KB \not\models \alpha_2$

Determining Entailment by Enumerating Truth Tables

Show that $\{\neg B_{2,1},\ P_{2,2} \Rightarrow B_{1,2} \wedge B_{2,1}\} \models \neg P_{2,2}$

$B_{1,2}$	$B_{2,1}$	$P_{2,2}$	$\neg B_{2,1}$	$P_{2,2} \Rightarrow B_{1,2} \wedge B_{2,1}$	$\neg P_{2,2}$
F	F	F	T	T	T
F	F	T	T	F	F
F	T	F	F	T	T
F	T	T	F	F	F
T	F	F	T	T	T
T	F	T	T	F	F
T	T	F	F	T	T
T	T	T	F	T	F

Enumerate rows (different assignments to symbols), if KB is true in row, check that α is too

Logical equivalence

Two sentences are logically equivalent iff true in same models:
$\alpha \equiv \beta$ if and only if $\alpha \models \beta$ and $\beta \models \alpha$

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

Validity and satisfiability

A sentence is valid if it is true in **all** models,

e.g., $True$, $A \lor \neg A$, $A \Rightarrow A$, $(A \land (A \Rightarrow B)) \Rightarrow B$

A sentence is satisfiable if it is true in **some** model

e.g., $A \lor B$, C

A sentence is unsatisfiable if it is true in **no** models

e.g., $A \land \neg A$

Next — finding logical consequences...