Agent-based Systems

Paolo Turrini

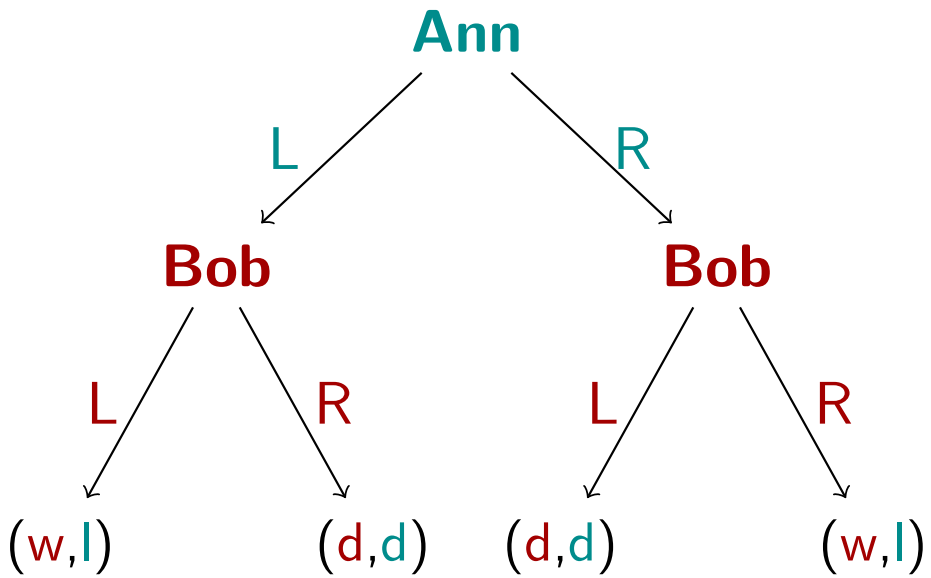
www.dcs.warwick.ac.uk/~pturrini p.turrini@warwick.ac.uk

Knowledge

The plan for today

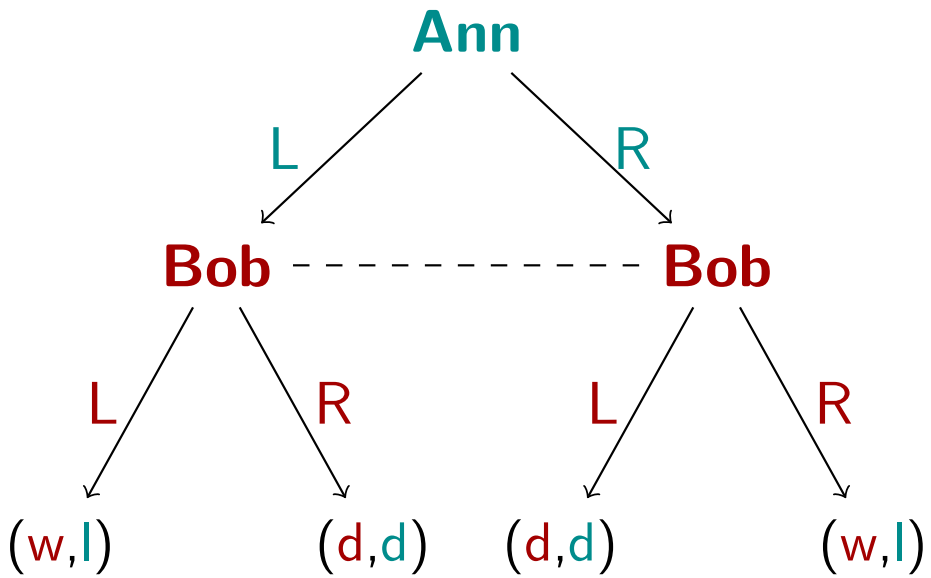
- Knowledge and possible worlds:
  - the knowledge relation
  - one agent versus many
  - various forms of group knowledge
- Knowing how to play
  - combining knowledge and strategy
  - intuitions from extensive games

Why knowledge?



Bob has a strategy to win...

Why knowledge?



Bob has a strategy to win... but does he know which one?

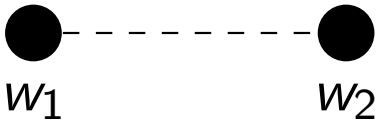
Knowledge vs Ignorance

The main underlying assumption behind the theory of knowledge in multi-agent systems (and in game theory) is that we live in one world and this world collects all the relevant facts (for instance, those that are true now, were true in the past and will be true in the future).

However, we are typically not able to fully determine it:

- is my mom thinking of me now?
- was there a penalty on Cristiano Ronaldo?
- will Berlusconi be president again?

Knowledge vs Ignorance



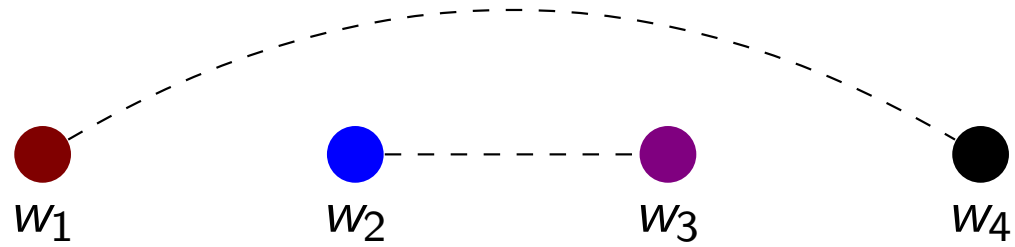
Example (Cristiano Ronaldo)

I'm watching a ~~Real Madrid~~ Juventus game and my stream pauses.
I've missed a penalty on Cristiano Ronaldo.
If Ronaldo is all that matters to me, there are two possible worlds:

- A world in which Ronaldo dives
- A world in which Ronaldo does not dive

I cannot say which one of these two worlds I'm living at.

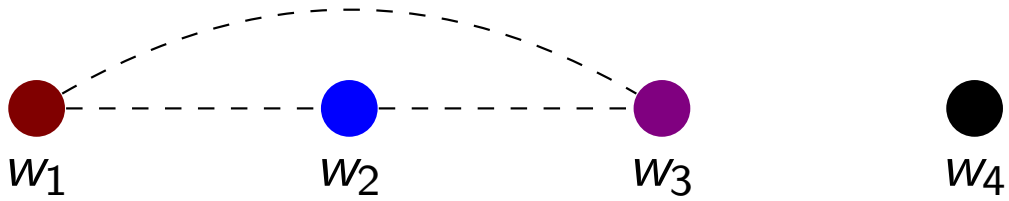
Knowledge vs Ignorance



Red is dive, blue is mum, purple is both, black is none.

- Typically there are other matters that I deem relevant.
- I might be able to distinguish some of their combination

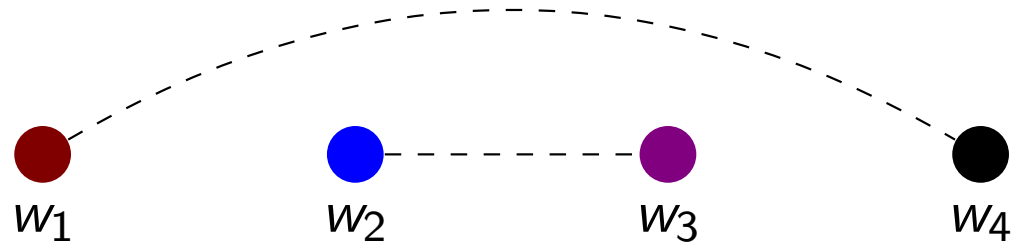
Knowledge vs Ignorance



Red is dive, blue is mum, purple is both, black is none.

- Typically there are other matters that I deem relevant.
- I might be able to distinguish some of their combination

Knowledge vs Ignorance



Red is dive, blue is mum, purple is both, black is none.

Definition (Facts)

Call any $E \subseteq W$ a **fact**, i.e., a fact is a subset of the possible worlds.

Name two facts in the example. What can you say about what I know?

Knowledge, formally

We model knowledge as a relation over the set of worlds.
Let us start with one agent, i, and their knowledge relation:

$$\sim_i \subseteq W \times W$$

Intuitively $(w_1, w_2) \in \sim_i$ means that agent i cannot distinguish between worlds w_1 and w_2.

So if i lives at w_1 they consider it possible that real world is actually w_2.
We call this relation the **indistinguishability** relation.

Properties

Again: $(w_1, w_2) \in \sim_i$ means i cannot distinguish between w_1 and w_2.

What are the intuitive properties of the indistinguishability relation?

Is it reflexive?

Reflexivity: for all $w \in W$, $(w, w) \in \sim_i$

Does this make sense to you?

Properties

Again: $(w_1, w_2) \in \sim_i$ means i cannot distinguish between w_1 and w_2.

What are the intuitive properties of the indistinguishability relation?

Is it symmetric?

Symmetry: for all $w_1, w_2 \in W$, IF $(w_1, w_2) \in \sim_i$ THEN $(w_2, w_1) \in \sim_i$

Does this make sense to you?

Properties

Again: $(w_1, w_2) \in \sim_i$ means i cannot distinguish between w_1 and w_2.

What are the intuitive properties of the indistinguishability relation?

Is it transitive?

Transitivity: for all $w_1, w_2, w_3 \in W$, IF $(w_1, w_2) \in \sim_i$ and $(w_2, w_3) \in \sim_i$ THEN $(w_1, w_3) \in \sim_i$

Does this make sense to you?

Properties

Again: $(w_1, w_2) \in \sim_i$ means i cannot distinguish between w_1 and w_2.

What are the intuitive properties of the indistinguishability relation?

Is it serial?

Seriality: for all $w_1 \in W$, there exists $w_2 \in W$ such that $(w_1, w_2) \in \sim_i$.

Does this make sense to you?

Can you possibly accept seriality without accepting the previous properties?

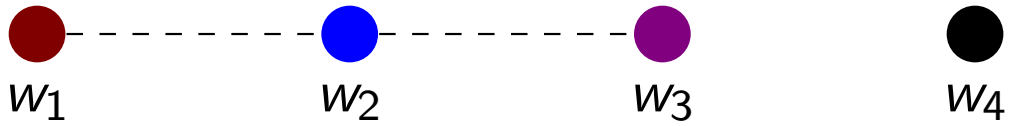
Knowledge

There is no ultimate model of knowledge, it depends on what *kind* of knowledge we want to capture.

For the time being, we are going to assume that it is an **equivalence relation** over the set of all states: reflexive, transitive, symmetric.

It is therefore a **partition** of the set of all states
(if you dont know why, prove it!)

Knowledge and Ignorance

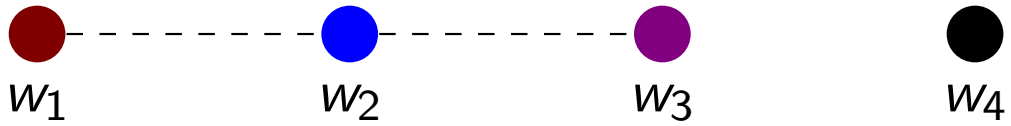


Red is dive, blue is mum, purple is both, black is none.

I'm going to compactly represent the relation with undirected arcs between indistinguishable worlds, omitting reflexive and transitive links.

Notice how this visually partitions the set of all worlds.

Knowledge and Ignorance

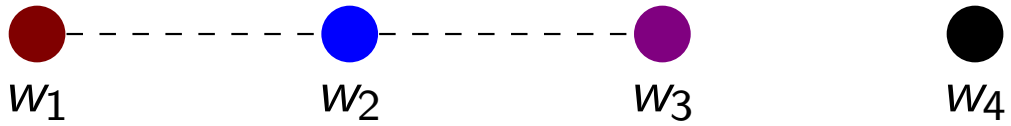


Red is dive, blue is mum, purple is both, black is none.

Now knowledge...

I know that something is true if and only if that something is true at all the worlds that I cannot distinguish from the one I'm at.

Knowledge and Ignorance



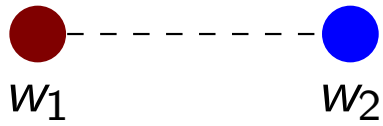
Red is dive, blue is mum, purple is both, black is none.

Let $\sim_i [w] = \{w' \mid (w, w') \in \sim_i\}$ be what i cannot distinguish from w.

We say that i **knows** fact E at w whenever $\sim_i [w] \subseteq E$

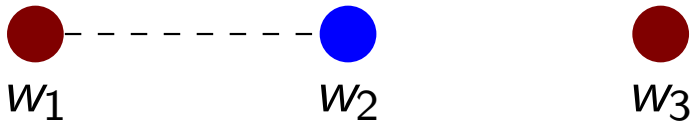
We denote $K_i E = \{w \mid \sim_i [w] \subseteq E\}$ the worlds at which i knows E.

Knowledge and Ignorance



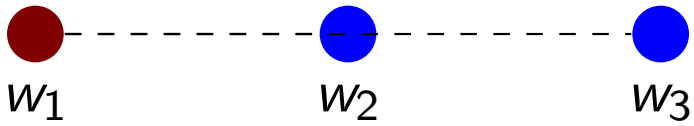
What does the agent know?

Knowledge and Ignorance



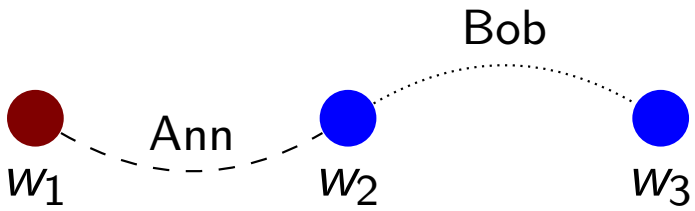
What does the agent know?

Knowledge and Ignorance



What does the agent know?

Knowledge and Ignorance

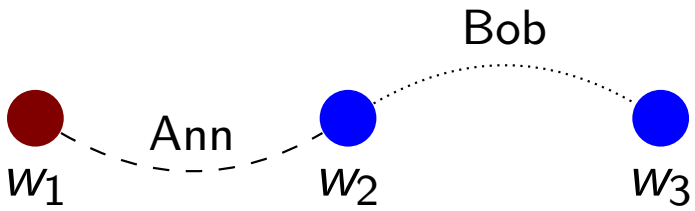


Now let us populate the world with other agents, each with their indistinguishability relation.

Notice the relations don't need to be related in any way.

What do these agents know?

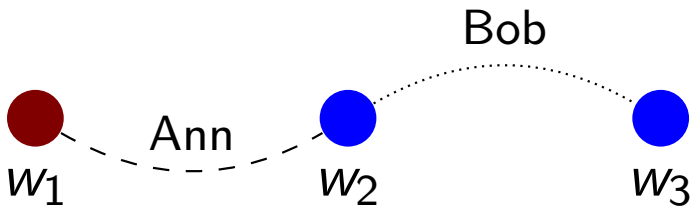
Multi-agent knowledge



At w_1:

- Does Ann know red?
- Does Bob know Ann knows red?
- Does Bob know Ann does not know Bob knows red?
- Does Bob know Bob knows Ann does not know Bob knows red?

Group knowledge

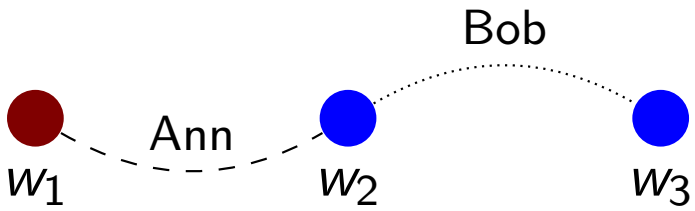


We can talk about the knowledge of many agents.

But we can also talk about what they know **together**.

What does this mean? Any intuitions?

Group knowledge



- What everyone knows already?
- Or with communication?
- Or joint observation?

As always, we need to make choices and restrict our study to some interesting forms of group knowledge.

General Knowledge

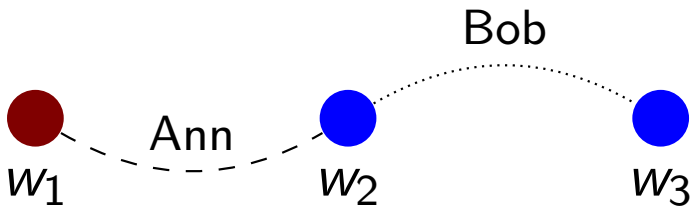
Let N be the set of agents and W the set of worlds.
Let E be a fact.

Definition (General Knowledge)

It is **general knowledge** that E at w if everyone knows E at w.

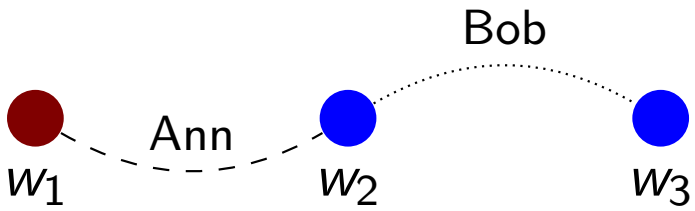
General knowledge is what everyone knows.
For example, Ann knows E **and** Bob knows E.

General Knowledge



Tell me two facts that are general knowledge at w_3.

General Knowledge



Tell me two facts that are general knowledge at w_3.

At w_2?

General Knowledge

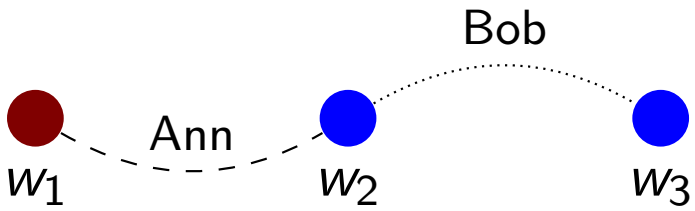
Let N be the set of agents and W the set of worlds.
Let E be a fact.

$$KE = \bigcap_{i \in N} K_i E = \bigcap_{i \in N} \{w \mid \sim_i [w] \subseteq E\}$$

is the set of worlds where everyone knows E.

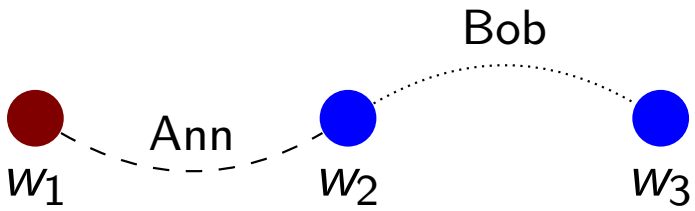
Notice: $w \in KE$ means that everyone knows E at w.

General Knowledge



Tell me one fact that is general knowledge everywhere.

General Knowledge



Tell me one fact that is general knowledge everywhere.

Can it be that (= can you come up with a model such that) nothing is general knowledge?

Distributed Knowledge

Let N be the set of agents and W the set of worlds.
Let E be a fact.

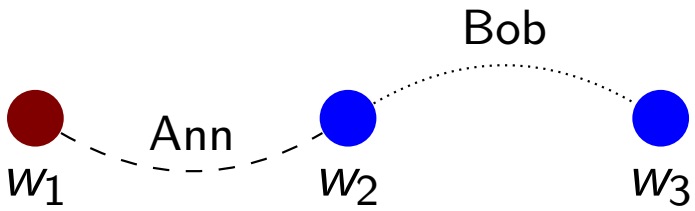
Definition (Distributed Knowledge)

We say that it is **distributed knowledge** that E at w if every agent knew E by intersecting their indistinguishability relation.

Distributed knowledge is the implicit knowledge the agents have, what they would know if they could communicate.
In other words, what they would know as a group.

For example: if Ann cannot distinguish w_1 from w_2 and Bob cannot distinguish w_2 from w_3 then, talking, they would know they are at w_2.

Distributed Knowledge



Tell me one fact that is distributed knowledge at w_2.

Tell me one fact that is not.

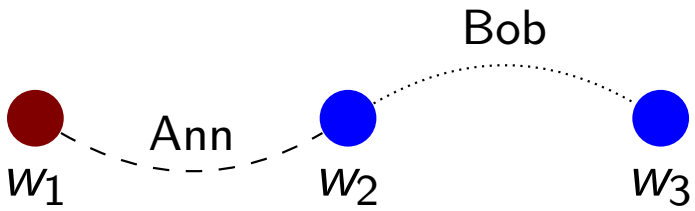
Distributed Knowledge

Let N be the set of agents and W the set of worlds.
Let E be a fact.

$$DKE = \{w \mid (\bigcap_{i \in N} \sim_i [w]) \subseteq E\}$$

is the set of worlds where it is distributed knowledge that E.

Distributed Knowledge



Tell me one fact that is distributed knowledge everywhere.

Tell me one fact that is not.

Common Knowledge

Let N be the set of agents and W the set of worlds.
Let E be a fact.

Definition

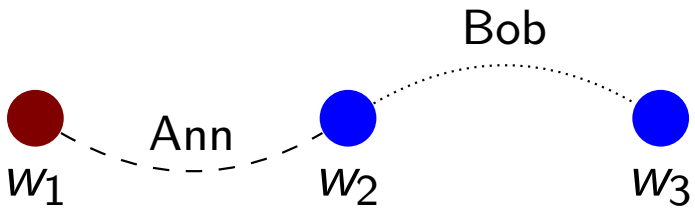
It is **common knowledge** that E at w if everyone knows E at w and knows that everyone knows, and knows that everyone knows that everyone knows, and kn... (too long to write)

Common knowledge is what is experienced by everyone, e.g., a result of a joint public observation.

It is an idealised setting, but - with some lenience - it can reasonably be assumed in certain circumstances.

For instance: look at the clock now. Its time is common knowledge.

Common Knowledge



What is common knowledge at w_3?

What is not?

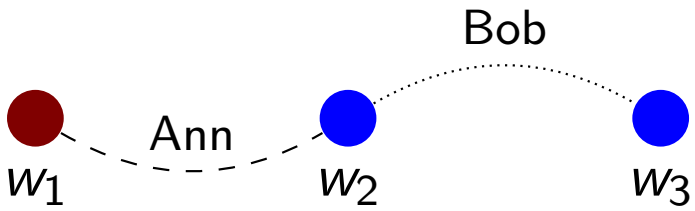
Common Knowledge

Let N be the set of agents and W the set of worlds.
Let E be a fact.

$$CE = \bigcap_{k=1}^{\infty} K^k E = KE \cap KKE \cap KKKE \ldots$$

is the set of worlds where it is common knowledge that E.

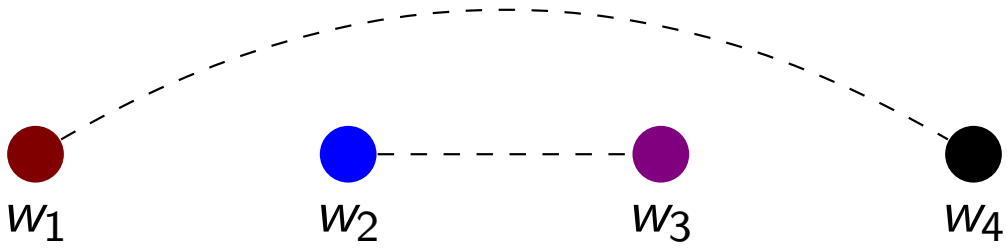
Common Knowledge



What is common knowledge everywhere?

What is not?

Learning



Red is dive, blue is mum, purple is both, black is none.

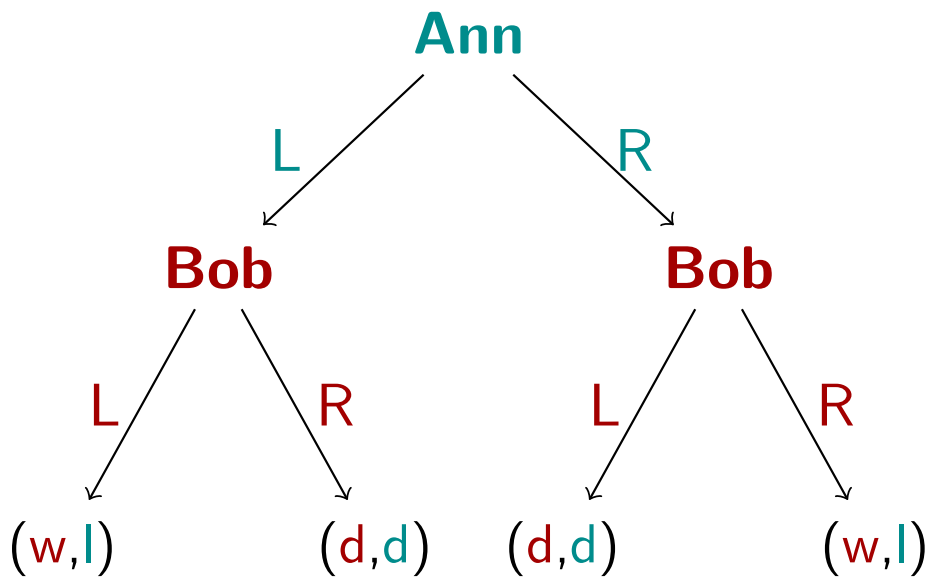
Suppose I'm told Ronaldo did dive.
How does the model change if I learn a new fact?

Learning

Red is dive, blue is mum, purple is both, black is none.

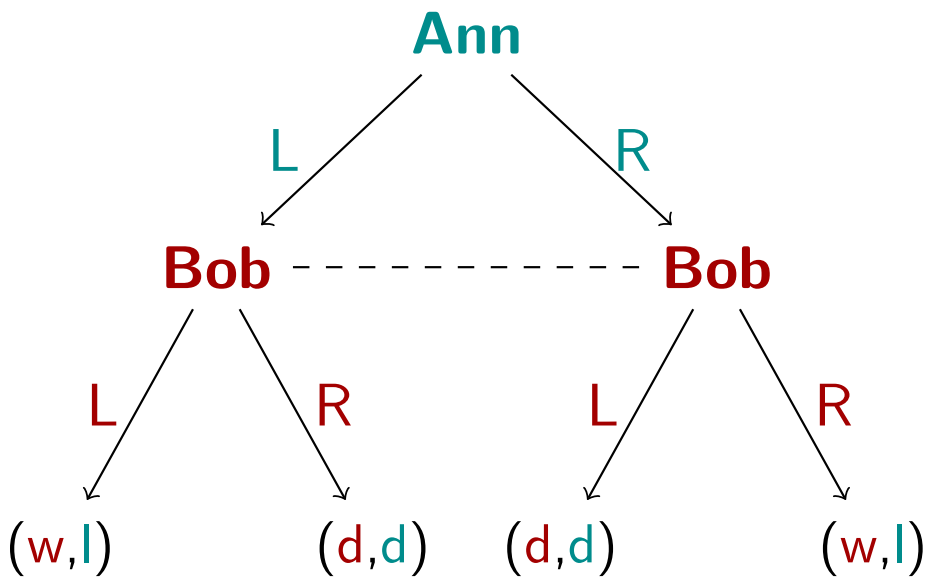
Suppose I'm told Ronaldo did dive.
How does the model change if I learn a new fact?

Chess vs Poker



Here Bob knows what Ann has chosen.

Chess vs Poker

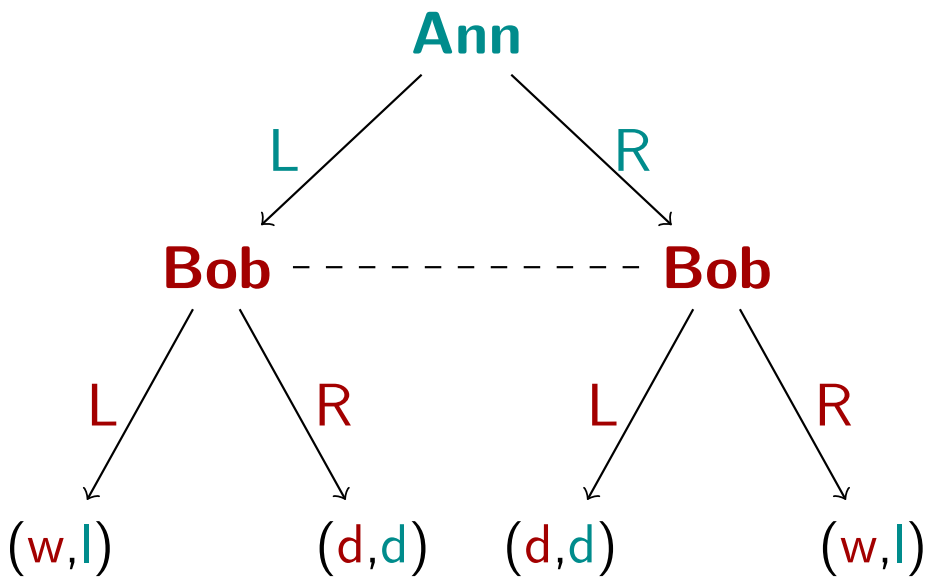


Here he doesn't.

In other words, there are two worlds that Bob cannot distinguish:

- the one in which Ann has chosen to go left
- and the other in which Ann has chosen to go right.

Chess vs Poker



There is a difference between:

- knowing that I have a winning strategy
- and knowing which one it is.

Bob knows he has a winning strategy, but does not know which one.

Knowing that versus knowing how

It makes sense to assume that players will choose the same actions in situations that they cannot distinguish.

After all, how can I say "if Ann has chosen left, then I go right", if I cannot observe any sign of this?

Knowing that versus knowing how

Definition (Uniform Strategies)

Let H be the set of histories of a game and let $\sim_i$ be a relation over these histories. [a]

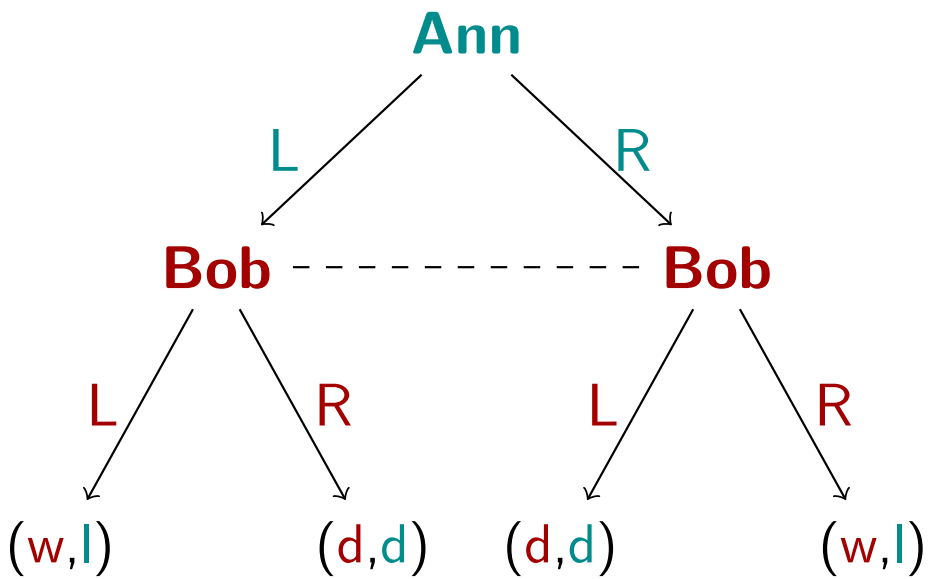
A strategy σ_i is said to be **uniform** if $\sigma_i(h) = \sigma_i(h')$ whenever $h \sim_i h'$.

In words, If I can't distinguish two situations, I'm going to have to play the same action in both.

[a] This is not a rigorous definition at this stage, but it's easy to see that it can fall under our case if we interpret a history as a possible world. We will make it precise later on in the course.

If I have a uniform strategy to win, then I know how to win.
Otherwise, I don't.
Bob, in the previous example, knows he can win, but does not know how.

Chess vs Poker



Which strategies are uniform?
Which ones aren't?

What we have seen

- A mathematical model of knowledge and ignorance
- Multi-agent knowledge
- Group knowledge: general, distributed, common
- Knowing that versus knowing know

What next?

- We are going to redo everything in logic ;)
- Logic and action
- Logic and knowledge
- Logic and preferences