Dealing with Other Agents: Modal Logics

Alan Bundy
School of Informatics
University of Edinburgh

Modalities

- Introduced to formalise *modalities*, *e.g.* necessity and possibility.
- Syntax:

$$\Box\varphi \text{ means, } \varphi \textit{ is necessarily true}$$

$$\Diamond\varphi \text{ means, } \varphi \textit{ is possibly true}$$

- Interdefinable: $\Diamond\varphi \Leftrightarrow \neg\Box\neg\varphi$ and $\Box\varphi \equiv \neg\Diamond\neg\varphi$

Applications of Modalities

Time:

$$\Box\varphi \text{ means, } \varphi \textit{ will be true from now on.}$$

$$\Diamond\varphi \text{ means, } \varphi \textit{ will eventually be true.}$$

Deontic:

$$\Box\varphi \text{ means, } \varphi \textit{ ought to be true.}$$

$$\Diamond\varphi \text{ means, } \varphi \textit{ is permissible.}$$

Knowledge: $\boxed{\mathsf{K}_A}\ \varphi$ means, A *knows that* φ.

Example Modal Formulae

$$\boxed{\mathsf{K}_A}\ \boxed{\mathsf{K}_B}\ \varphi \text{ means, } A \textit{ knows that } B \textit{ knows that } \varphi$$

$$\exists x.\ \boxed{\mathsf{K}_A}\ \varphi(x) \text{ means, } \textit{for some } x,\ A \textit{ knows that } \varphi(x)$$

$$\boxed{\mathsf{K}_A}\ \exists x.\ \varphi(x) \text{ means, } A \textit{ knows that, for some } x, \varphi(x)$$

Suppose $\varphi(x)$ means, *x is the name of the oldest person in Edinburgh*, and you are A.

Possible World Semantics

- There are many possible worlds,
  with different facts true in each: $w \models \varphi$.
  There is a distinguished, current world, *e.g.* w_0.
  Some worlds are *accessible* $(w_1 \equiv w_2)$) from other worlds, some are not.
- $w_0 \models \Box\varphi$ iff $\forall w.\ w_0 \equiv w \Rightarrow w \models \varphi$.
- $w_0 \models \Diamond\varphi$ iff $\exists w.\ w_0 \equiv w \wedge w \models \varphi$.
- $w_0 \models \boxed{\mathsf{K}_A}\,\varphi$ iff $\forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$.

Example of Possible Worlds

- There are 3 cards: King, Queen and Jack.
- There are two agents: A and B.
- Each agent has one card and there is one face down on the table.
- Agent A has the King.
- Agent A considers two possible worlds:
  Agent B has the Queen: w_Q.
  Agent B has the Jack: w_J.
- One of these is the actual world.

Establishing Formulae via Semantics

Suppose:	$w_0 \models \boxed{\mathsf{K}_A}\,\varphi$ and $\varphi \models \psi$
by meaning $\boxed{\mathsf{K}_A}$:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$
by meaning $\models$:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \psi$
by meaning $\boxed{\mathsf{K}_A}$:	$w_0 \models \boxed{\mathsf{K}_A}\,\psi$
discharging assumption:	if $\boxed{\mathsf{K}_A}\,\varphi$ and $\varphi \models \psi$ then $\boxed{\mathsf{K}_A}\,\psi$

Mid-Lecture Exercise

- Represent each of the following statements as a modal logic formula.
  1. Agent X knows that everyone has a name.
  2. Agent X knows what everyone's name is.

  where $Name(p, n)$ means that n is the name of p.
- In what way do these two formulae differ?
- Does either of them imply the other?

Solution to Exercise

- 1. $\boxed{\mathsf{K}_X}\ \forall p.\exists n.\ Name(p,n)$

  2. $\forall p.\exists n.\ \boxed{\mathsf{K}_X}\ Name(p,n)$
- They differ only in whether the modal operator appears before or after the quantifiers.
- 2 implies 1, but not vice versa.

Solution Continued

	Current World	**Accessible World**
1	$\boxed{\mathsf{K}_X}\ \forall p.\exists n.\ Name(p,n)$	$\forall p.\exists n.\ Name(p,n)$ $Name(p_1,n_1)$, $Name(p_2,n_2)$, $\ldots$
2	$\forall p.\exists n.\ \boxed{\mathsf{K}_X}\ Name(p,n)$ $\boxed{\mathsf{K}_X}\ Name(p_1,n_1')$, $\boxed{\mathsf{K}_X}\ Name(p_2,n_2')$, $\ldots$	 $Name(p_1,n_1')$, $Name(p_2,n_2')$, $\ldots$

Property K: What An Agent Infers It Knows

Suppose:	$w_0 \models \boxed{\mathsf{K}_A}\ (\varphi \rightarrow \psi)$
by meaning $\boxed{\mathsf{K}_A}$:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models (\varphi \rightarrow \psi)$
Suppose:	$w_0 \models \boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$
by modus ponens:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \psi$
by meaning $\boxed{\mathsf{K}_A}$:	$w_0 \models \boxed{\mathsf{K}_A}\ \psi$
discharging assumptions:	$\boxed{\mathsf{K}_A}\ (\varphi \rightarrow \psi) \rightarrow (\boxed{\mathsf{K}_A}\ \varphi \rightarrow \boxed{\mathsf{K}_A}\ \psi)$

Property K and Omniscience

Property K: An agent knows it can infer.

Infallible: Agent will never make mistakes during reasoning.

Exhaustive: Agent will draw all possible inferences.

Neither of these is realistic in real agents.

However, adopt as first approximation.

Properties of $\equiv_A$

Reflexive: $\forall w.\ w \equiv_A w$

Symmetric: $\forall w_1.\forall w_2.\ w_1 \equiv_A w_2 \Rightarrow w_2 \equiv_A w_1$

Transitive:

$$\forall w_1.\forall w_2.\forall w_3.\ w_1 \equiv_A w_2 \wedge w_2 \equiv_A w_3 \Rightarrow w_1 \equiv_A w_3$$

Property T: Anything An Agent Knows is True

Suppose:	$w_0 \models \boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$
since $\equiv_A$ is reflexive:	$w_0 \models \varphi$
discharging assumption:	$\boxed{\mathsf{K}_A}\ \varphi \rightarrow \varphi$

Speak of *knowledge* when property **T** holds and *belief* when it fails.

Property 4: An Agent Knows What It Knows

Suppose:	$w_0 \models \boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$(*)\ \forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$
Suppose:	$w_0 \equiv_A w'$
Suppose:	$w' \equiv_A w$
by transitivity of $\equiv_A$:	$w_0 \equiv_A w$
by $(*)$:	$w \models \varphi$
discharging assumption:	$\forall w.\ w' \equiv_A w \Rightarrow w \models \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$w' \models \boxed{\mathsf{K}_A}\ \varphi$
discharging assumption:	$\forall w.\ w_0 \equiv_A w \Rightarrow w \models \boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$w_0 \models \boxed{\mathsf{K}_A}\boxed{\mathsf{K}_A}\ \varphi$
discharging assumption:	$\boxed{\mathsf{K}_A}\ \varphi \rightarrow \boxed{\mathsf{K}_A}\boxed{\mathsf{K}_A}\ \varphi$

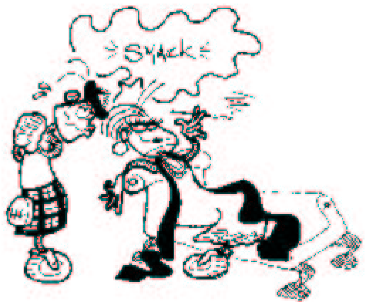
Property 5: An Agent Knows What It Doesn't Know.

Suppose:	$w_0 \models \neg\boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$\neg\forall w.\ w_0 \equiv_A w \Rightarrow w \models \varphi$
equivalently:	$\exists w.\ w_0 \equiv_A w \wedge w \models \neg\varphi$
i.e. for some: w_1 :	$(*)\ w_0 \equiv_A w_1 \wedge w_1 \models \neg\varphi$
Suppose:	$w_0 \equiv_A w'$
by symmetry $\equiv_A$:	$w' \equiv_A w_0$
by transitivity $\equiv_A$:	$(\dagger)\ w' \equiv_A w_1$
from $(*)\&(\dagger)$	$\exists w.\ w' \equiv_A w \wedge w \models \neg\varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$w' \models \neg\boxed{\mathsf{K}_A}\ \varphi$
discharging assumption:	$\forall w.\ w_0 \equiv_A w' \Rightarrow w' \models \neg\boxed{\mathsf{K}_A}\ \varphi$
by meaning $\boxed{\mathsf{K}_A}$:	$\boxed{\mathsf{K}_A}\neg\boxed{\mathsf{K}_A}\ \varphi$
discharging assumption:	$\neg\boxed{\mathsf{K}_A}\ \varphi \rightarrow \boxed{\mathsf{K}_A}\neg\boxed{\mathsf{K}_A}\ \varphi$

A Family of Model Logics

- Property **K** true in all modal logics.
- If $\equiv_A$ reflexive then **T** also true and logic called **KT**.
- If $\equiv_A$ reflexive and transitive then **4** also true and logic called **S4**.
- If $\equiv_A$ reflexive, symmetric and transitive then **5** also true and logic called **S5**.

Example from AI1 Lectures



- Mairi accuses Jock of cheating on her with Karen.
- Jock denies it.
- How can we account for the disagreement?

Differences in Their Beliefs

Mairi's Beliefs:

$$\boxed{\mathsf{K}_M}\ kissed(P_1, P_2) \Rightarrow affair(P_1, P_2)$$
$$\boxed{\mathsf{K}_M}\ kissed(jock, karen)$$

Jock's Beliefs:

$$\boxed{\mathsf{K}_J}\ kissed(P_1, P_2) \wedge love(P_1, P_2) \Rightarrow affair(P_1, P_2)$$
$$\boxed{\mathsf{K}_J}\ kissed(jock, karen)$$
$$\boxed{\mathsf{K}_J}\ \neg loves(jock, karen)$$

Lead to Difference in Their Conclusions

- Mairi infers that Jock is having an affair, but Jock doesn't,

  i.e. $\boxed{\mathsf{K}_M}\ affair(jock, karen)$ but not $\boxed{\mathsf{K}_J}\ affair(jock, karen)$.
- Note that property **T** cannot be true in this modal logic,

  since someone believes something that is false.

Conclusion

- Modal logics can be used to represent time, obligation and knowledge.
  We focus on knowledge.
- Given meaning via possible world semantics.
  Accessibility defined by $\equiv_A$.
- Properties **K**, **T**, **4** and **5**,
  depend on properties of $\equiv_A$: reflexive, symmetric, transitive.
- Problem of omniscience because of **K**.
- Family of logics depending which properties adopted.
  For instance, for belief reject **T**.
- Can use logic to account for differences in knowledge and belief.