Towards a Theory of Intention Revision

Workshop on the Dynamics of Intention and Preference

Eric Pacuit

Stanford University

February 27, 2008

Plan

- Existing literature
- Underlying ETL model
- Elements of a theory of intention revision
- Many agents

Some Literature

Stemming from Bratman's planning theory of intention a number of logics of rational agency have been developed:

- Cohen and Levesque; Rao and Georgeff (BDI); Meyer, van der Hoek (KARO); Bratman, Israel and Pollack (IRMA); and many others.

Some Literature

Stemming from Bratman's planning theory of intention a number of logics of rational agency have been developed:

- Cohen and Levesque; Rao and Georgeff (BDI); Meyer, van der Hoek (KARO); Bratman, Israel and Pollack (IRMA); and many others.

Some common features

- Underlying temporal model
- Belief, Desire, Intention, Plans, Actions are defined with corresponding operators in a language

J.-J. Meyer and F. Veltman. *Intelligent Agents and Common Sense Reasoning*. Handbook of Modal Logic, 2007.

Intention Revision

Many of the frameworks do discuss some form of intention revision.

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

- Beliefs are sets of Linear Temporal Logic formulas (eg., $\bigcirc\varphi$)

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

- Beliefs are sets of Linear Temporal Logic formulas (eg., $\bigcirc\varphi$)
- Desires are (possibly inconsistent) sets of Linear Temporal Logic formulas

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

- Beliefs are sets of Linear Temporal Logic formulas (eg., $\bigcirc\varphi$)
- Desires are (possibly inconsistent) sets of Linear Temporal Logic formulas
- Practical reasoning rules: $\alpha \leftarrow \alpha_1, \alpha_2, \ldots, \alpha_n$

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

- Beliefs are sets of Linear Temporal Logic formulas (eg., $\bigcirc\varphi$)
- Desires are (possibly inconsistent) sets of Linear Temporal Logic formulas
- Practical reasoning rules: $\alpha \leftarrow \alpha_1, \alpha_2, \ldots, \alpha_n$
- Intentions are derived from the agents current active plans (trees of practical reasoning rules)

Intention Revision

Many of the frameworks do discuss some form of intention revision.

W. van der Hoek, W. Jamroga and M. Wooldridge. *Towards a Theory of Intention Revision*. Synthese, 2007.

- Two types of beliefs: strong beliefs vs. weak beliefs (beliefs that take into account the agent's intentions)
- A dynamic update operator is defined ($[\Omega]\varphi$)

Underlying ETL Model (single agent)

- N is a set of **nodes**, or **states**
- A is a set of **primitive actions**.
- $\preceq$ is the successor relation on N (with the usual properties)
- l is a labeling function. Formally, l is a partial function from $N \times N$ to A where $l(n, n') \in A$ if $n \preceq n'$ and undefined otherwise.

Underlying ETL Model (single agent)

The agent's uncertainty is represented by a relation $\sim \subseteq N \times N$:

I.e., $n \sim n'$ if according to the agent's current information (i.e., *the events the agent has observed*), the agent cannot distinguish state n from n'

Underlying ETL Model (single agent)

The agent's uncertainty is represented by a relation $\sim \subseteq N \times N$:

I.e., $n \sim n'$ if according to the agent's current information (i.e., *the events the agent has observed*), the agent cannot distinguish state n from n'

Some assumptions:

- **Perfect Recall**: If $n \sim n'$, then $h_n = h'_n$. This means that the agents remembers all of its choices.

Underlying ETL Model (single agent)

The agent's uncertainty is represented by a relation $\sim \subseteq N \times N$:

I.e., $n \sim n'$ if according to the agent's current information (i.e., *the events the agent has observed*), the agent cannot distinguish state n from n'

Some assumptions:

- **Perfect Recall**: If $n \sim n'$, then $h_n = h'_n$. This means that the agents remembers all of its choices.
- **No Miracles**: If $n \sim n'$ and there are n_1 and n_2 with $l(n, n_1) = l(n', n_2)$, then $n_1 \sim n_2$.

Underlying ETL Model (single agent)

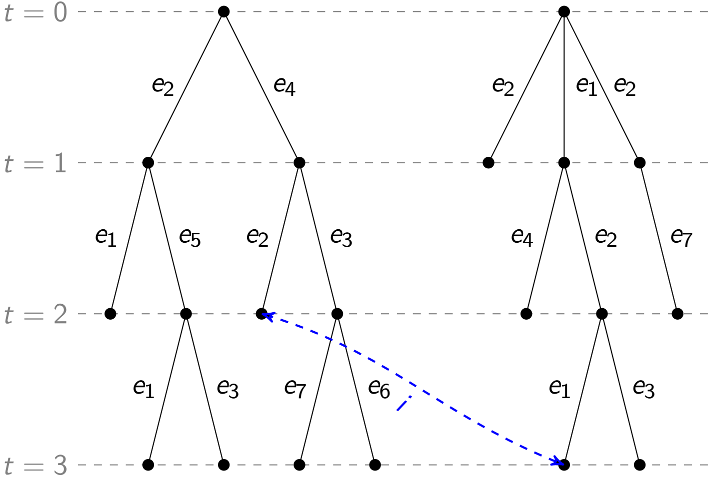
The agent's uncertainty is represented by a relation $\sim \subseteq N \times N$:

I.e., $n \sim n'$ if according to the agent's current information (i.e., *the events the agent has observed*), the agent cannot distinguish state n from n'

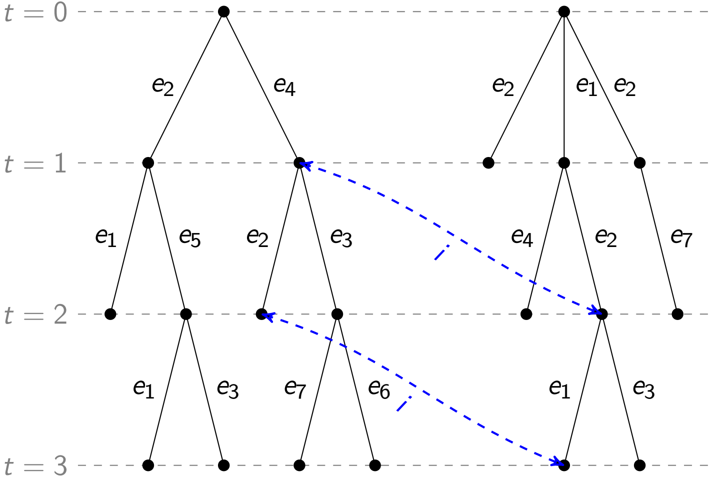
Some assumptions:

- **Perfect Recall**: If $n \sim n'$, then $h_n = h'_n$. This means that the agents remembers all of its choices.
- **No Miracles**: If $n \sim n'$ and there are n_1 and n_2 with $l(n, n_1) = l(n', n_2)$, then $n_1 \sim n_2$.
- **Uniform Actions**: If $n_1 \sim n_2$ and $l(n_1, n') = a$ then there is a n'' such that $l(n_2, n'') = a$. This means the agents knows which options are available.

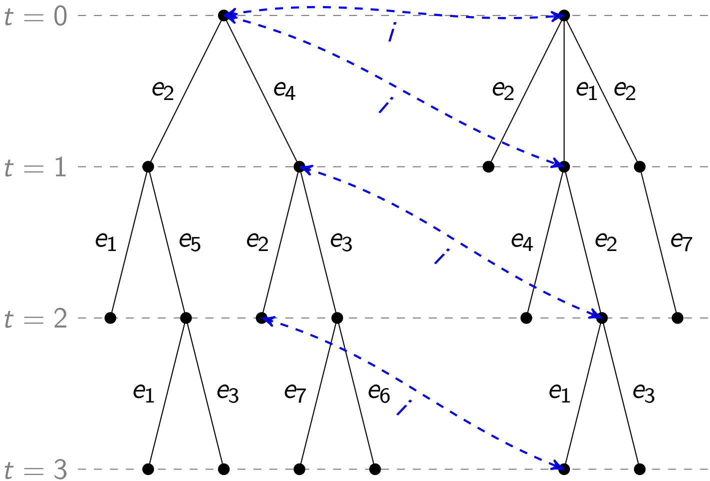
Perfect Recall



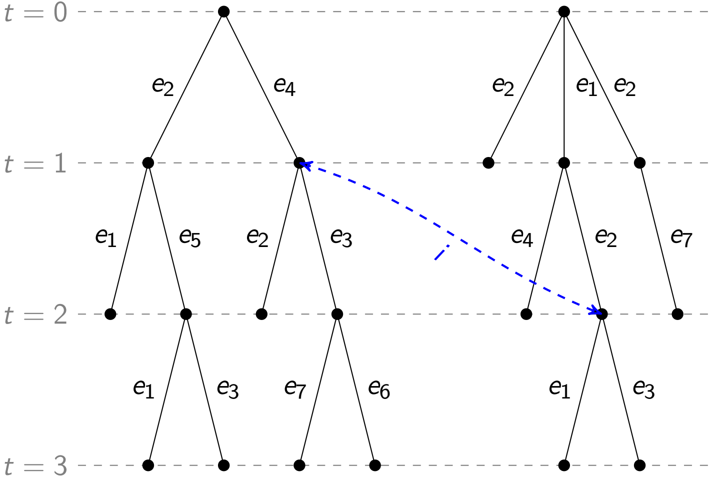
Perfect Recall



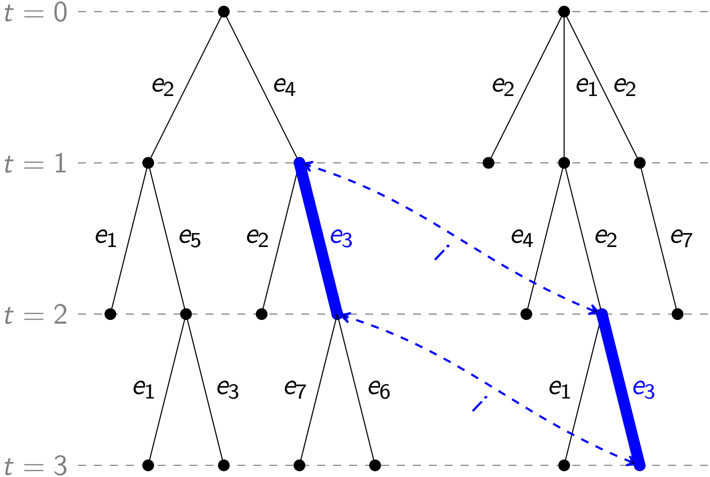
Perfect Recall



No Miracles



No Miracles



Uniform Actions

Uniform Actions: If $n_1 \sim n_2$ and $l(n_1, n') = a$ then there is a n'' such that $l(n_2, n'') = a$. This means the agents knows which options are available.

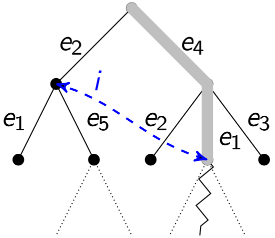
Histories

A **history** is a sequence of events or actions (for each node there is a history from a root to that node).

Two types of uncertainty

Given two finite histories h and h',

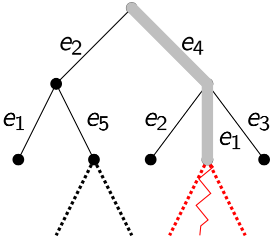
> $h \sim_i h'$ *means* given the events i has observed, *h and h' are indistinguishable*



Two types of uncertainty

Given two **maximal histories** H and H',

agent i may be uncertain which of the two will be the final outcome.



Elements of a theory of intention revision

Suppose $A = \{a_1, a_2, \ldots\}$ is a set of primitive actions and $P = \{p_1, p_2, \ldots\}$ a set of propositional variables.

Elements of a theory of intention revision

Suppose $A = \{a_1, a_2, \ldots\}$ is a set of primitive actions and $P = \{p_1, p_2, \ldots\}$ a set of propositional variables.

$\mathcal{L}_0$ be the propositional language generated by P.

Elements of a theory of intention revision

Suppose $A = \{a_1, a_2, \ldots\}$ is a set of primitive actions and $P = \{p_1, p_2, \ldots\}$ a set of propositional variables.

$\mathcal{L}_0$ be the propositional language generated by P.

Our basic expressions:

- $B(\varphi : a_1, \ldots a_n)$ is intended to mean "the agent believes φ *because* he intends to do a_1, and he intends to do a_2, ..., and he intends to do a_n."

Elements of a theory of intention revision

Suppose $A = \{a_1, a_2, \ldots\}$ is a set of primitive actions and $P = \{p_1, p_2, \ldots\}$ a set of propositional variables.

$\mathcal{L}_0$ be the propositional language generated by P.

Our basic expressions:

- $B(\varphi : a_1, \ldots a_n)$ is intended to mean "the agent believes φ *because* he intends to do a_1, and he intends to do a_2, ..., and he intends to do a_n."
- $I(a : \varphi_1, \ldots, \varphi_m)$ is intended to mean "the agent intends to do a *because* he does not believe $\neg\varphi_1$ and he does not believe $\neg\varphi_2, \ldots,$ and he does not believe $\neg\varphi_m$."

Elements of a theory of intention revision

Remark: We do not allowing the following formulas:

1. $B(I(a_1 : p) : a_2)$: "The agent believes that [he intends to do a_1 because he does not believe $\neg p$] because he intends to do a_2.
2. $I(a : Bp)$: "The agent intends to do a because he does not believe that $\neg Bp$.

Digression: Knowing/Believing for a reason

- S. Artemov has developed *justification logics* with a similar flavor: $t : \varphi$ is intended to mean "the agent knows φ for reason t".

Digression: Knowing/Believing for a reason

- S. Artemov has developed *justification logics* with a similar flavor: $t : \varphi$ is intended to mean "the agent knows φ for reason t".

- t is called a proof polynomial and is intended to represent a derivation of φ.
  - There is an algebraic structure on the set of proof polynomials: $t + s$, $!t$, $t \cdot s$

Digression: Knowing/Believing for a reason

- S. Artemov has developed *justification logics* with a similar flavor: $t : \varphi$ is intended to mean "the agent knows φ for reason t".

- t is called a proof polynomial and is intended to represent a derivation of φ.
  - There is an algebraic structure on the set of proof polynomials: $t + s$, $!t$, $t \cdot s$

- Mel Fitting has developed a Kripke style semantics.

  M. Fitting. *Logic of Proofs, Semantically*. APAL, 2006.

Elements of a theory of intention revision

Any Action: Let ? be a new action symbol (not in A). The intended interpretation of ? is "any action".

Elements of a theory of intention revision

Any Action: Let ? be a new action symbol (not in A). The intended interpretation of ? is "any action".

- $B(\varphi) =_{\text{def}} B(\varphi :?)$
- $I(a) =_{\text{def}} I(a : \top)$.

Elements of a theory of intention revision

A **belief set** $\mathcal{B}$ is any set of expressions of the form $\varphi : a_1, \ldots, a_n$ where $\varphi \in \mathcal{L}_0$ and $a_1, \ldots, a_n \in A$ satisfying the following properties:

1. If $\varphi_1 : a_1, \ldots, a_n \in \mathcal{B}$ and $\varphi_2 : b_1, \ldots, b_m \in \mathcal{B}$ then $\varphi_1 \wedge \varphi_2 : a_1, \ldots, a_n, b_1, \ldots, b_m \in \mathcal{B}$.
2. $\vdash \varphi_1 \rightarrow \varphi_2$ (in propositional logic) and $\varphi_1 : a_1, \ldots, a_n \in \mathcal{B}$ then $\varphi_2 : a_1, \ldots, a_n \in \mathcal{B}$
3. $\bot : a_1, \ldots, a_n \notin \mathcal{B}$

Elements of a theory of intention revision

An **intention set** $\mathcal{I}$ is any set of expressions of the form $a : \varphi_1, \ldots, \varphi_n$ where $a \in A$ and $\varphi_1, \ldots, \varphi_n \in \mathcal{L}_0$ satisfying the following properties:

1. If $a : \varphi_1, \ldots, \varphi_n \in \mathcal{I}$ and $a : \psi_1, \ldots \psi_m \in \mathcal{I}$ then $a : \varphi_1, \ldots, \varphi_n, \psi_1, \ldots, \psi_m \in \mathcal{I}$
2. If $a : \varphi_1, \ldots, \varphi_n \in \mathcal{I}$ then $\bigwedge_i \varphi_i$ is logically consistent.

Elements of a theory of intention revision

Intention-Belief Pairs: A pair $(\mathcal{I}, \mathcal{B})$ where $\mathcal{I}$ is an intention set and $\mathcal{B}$ is a belief set is called an intention-belief pair.

The main idea is that a pair $(\mathcal{I}, \mathcal{B})$ is intended to represent the agents current intentions and beliefs.

However, not every pair $(\mathcal{I}, \mathcal{B})$ will represent "coherent" intentions and beliefs.

Elements of a theory of intention revision

A pair $(\mathcal{I}, \mathcal{B})$ is **coherent** provided:

Elements of a theory of intention revision

A pair $(\mathcal{I}, \mathcal{B})$ is **coherent** provided:

1. The intentions are grounded in current beliefs: (if $a : \varphi \in \mathcal{I}$ there $\neg\varphi :? \notin \mathcal{B}$)

Elements of a theory of intention revision

A pair $(\mathcal{I}, \mathcal{B})$ is **coherent** provided:

1. The intentions are grounded in current beliefs: (if $a : \varphi \in \mathcal{I}$ there $\neg\varphi :? \notin \mathcal{B}$)

2. There are no cycles (eg., there is no $\varphi : a \in \mathcal{B}$ with $a : \varphi \in \mathcal{I}$)

Elements of a theory of intention revision

Revision operators:

Elements of a theory of intention revision

Revision operators:

1. **Change/update the reason for believing** φ: Suppose that the agent currently believes φ because he intends to do a. Updating by $\varphi : b$ involves changing/adding a reason for believing φ.

Elements of a theory of intention revision

Revision operators:

1. **Change/update the reason for believing** φ: Suppose that the agent currently believes φ because he intends to do a. Updating by $\varphi : b$ involves changing/adding a reason for believing φ.
2. **Update/revise beliefs**: The input is a formula φ and, following AGM, the agent changes its belief set $\mathcal{B}$ appropriately. Note that in order to maintain coherency, we may need to drop some intentions.

Elements of a theory of intention revision

Revision operators:

1. **Change/update the reason for believing** φ: Suppose that the agent currently believes φ because he intends to do a. Updating by $\varphi : b$ involves changing/adding a reason for believing φ.
2. **Update/revise beliefs**: The input is a formula φ and, following AGM, the agent changes its belief set $\mathcal{B}$ appropriately. Note that in order to maintain coherency, we may need to drop some intentions.
3. **Weak-add an intention**: Add an intention provided coherency is maintained otherwise do not add the intention.

Elements of a theory of intention revision

Revision operators:

1. **Change/update the reason for believing** φ: Suppose that the agent currently believes φ because he intends to do a. Updating by $\varphi : b$ involves changing/adding a reason for believing φ.
2. **Update/revise beliefs**: The input is a formula φ and, following AGM, the agent changes its belief set $\mathcal{B}$ appropriately. Note that in order to maintain coherency, we may need to drop some intentions.
3. **Weak-add an intention**: Add an intention provided coherency is maintained otherwise do not add the intention.
4. **Strong-add an intention**: Add the intention and change belief/intention set appropriately.

Many Agents

Many Agents

E. Pacuit, R. Parikh and E. Cogan. *The Logic of Knowledge Based Applications*. Knowledge, Rationality and Action (Synthese) 149: 311 - 341 (2006).

Many Agents

E. Pacuit, R. Parikh and E. Cogan. *The Logic of Knowledge Based Applications*. Knowledge, Rationality and Action (Synthese) 149: 311 - 341 (2006).

Issues: obligations, group obligations, knowledge, group knowledge, default obligations, etc.

An agent's obligations are often dependent on what the agent knows, and indeed one cannot reasonably be expected to respond to a problem if one is not aware of its existence.

Motivating Example

1. Uma is a physician whose neighbour is ill. Uma does not know and has not been informed. Uma has no obligation (as yet) to treat the neighbour.

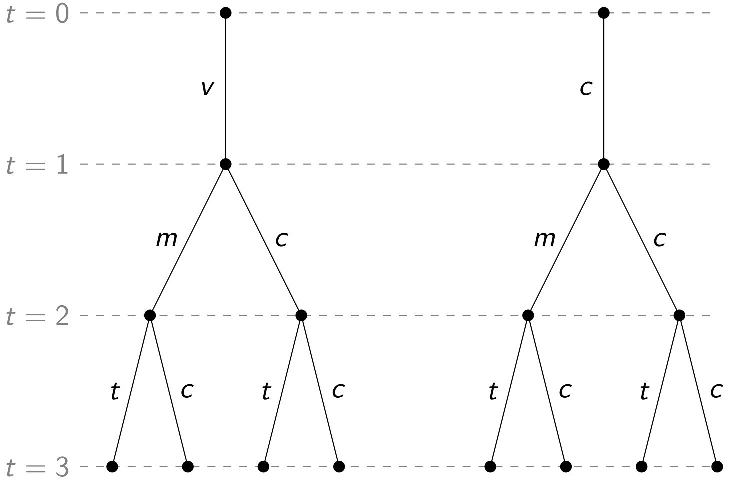
2. Uma is a physician whose neighbour Sam is ill. The neighbour's daughter Ann comes to Uma's house and tells her. Now Uma does have an obligation to treat Sam, or perhaps call in an ambulance or a specialist.

Motivating Example

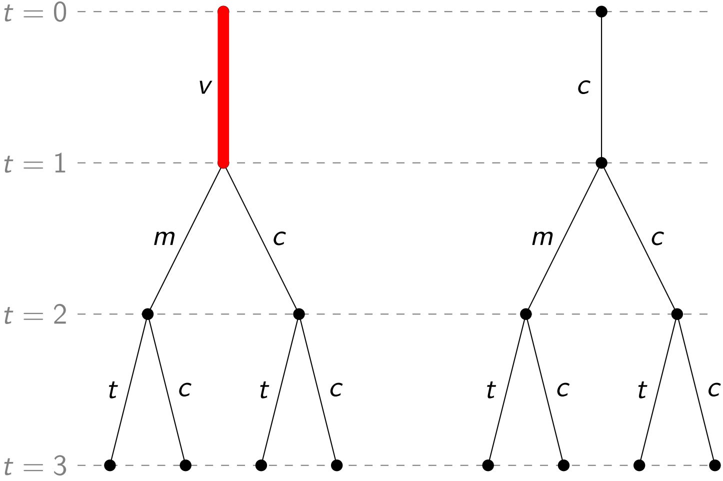
1. Uma is a physician whose neighbour is ill. Uma does not know and has not been informed. Uma has no obligation (as yet) to treat the neighbour.

2. Uma is a physician whose neighbour Sam is ill. The neighbour's daughter Ann comes to Uma's house and tells her. Now Uma does have an obligation to treat Sam, or perhaps call in an ambulance or a specialist.

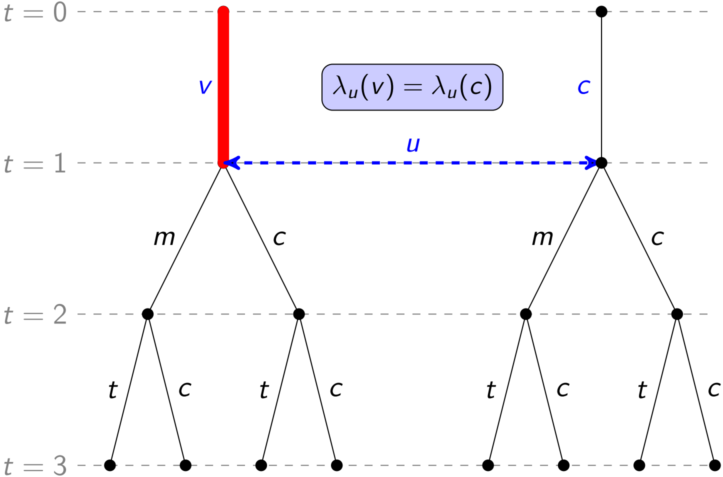
Formalizing Example 1 & 2



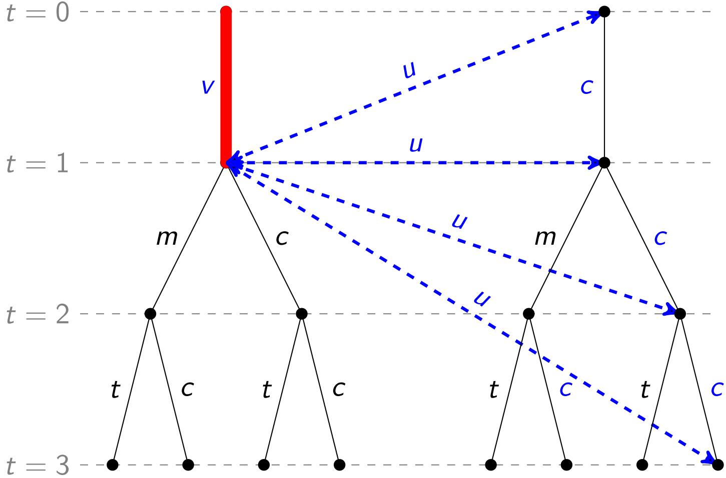
Formalizing Example 1 & 2



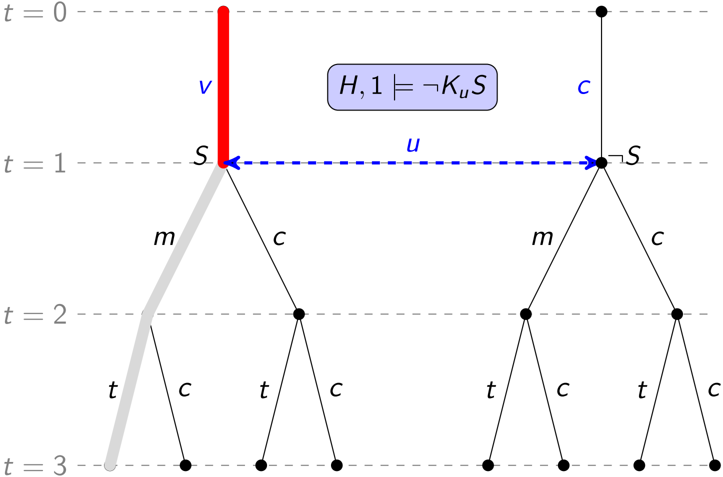
Formalizing Example 1 & 2



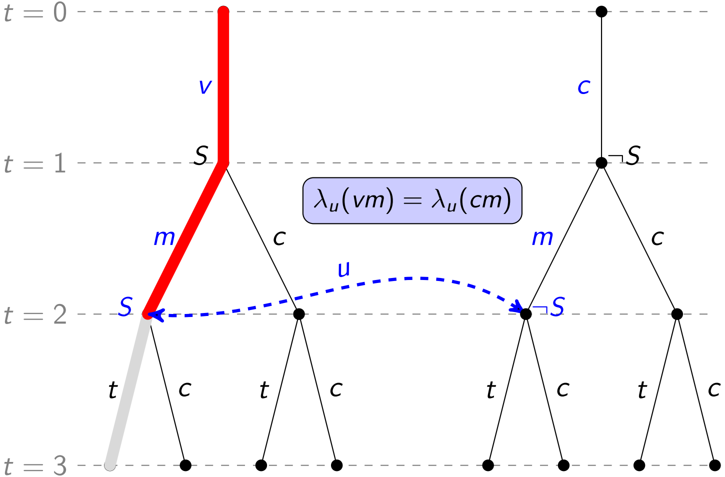
Formalizing Example 1 & 2



Formalizing Example 1 & 2

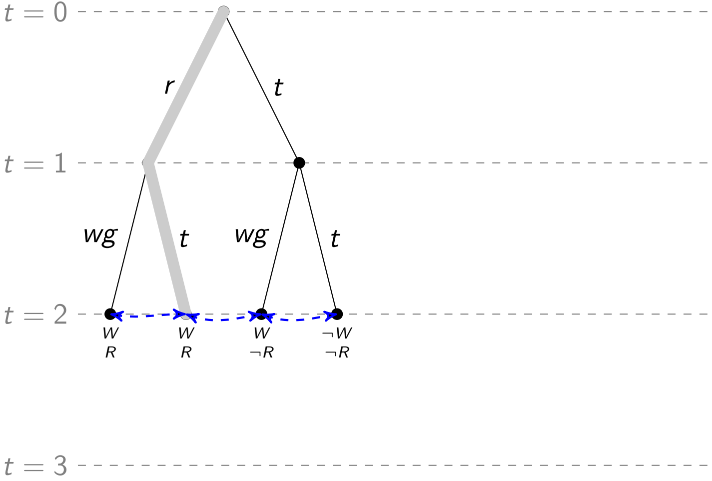


Formalizing Example 1 & 2

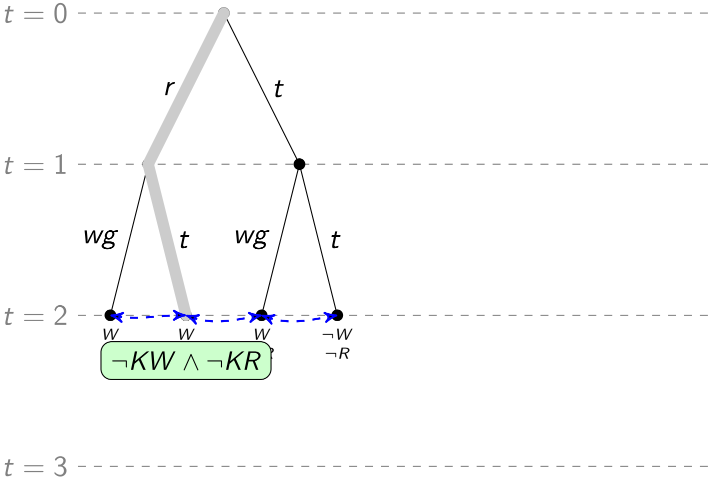


Learning from the Protocol

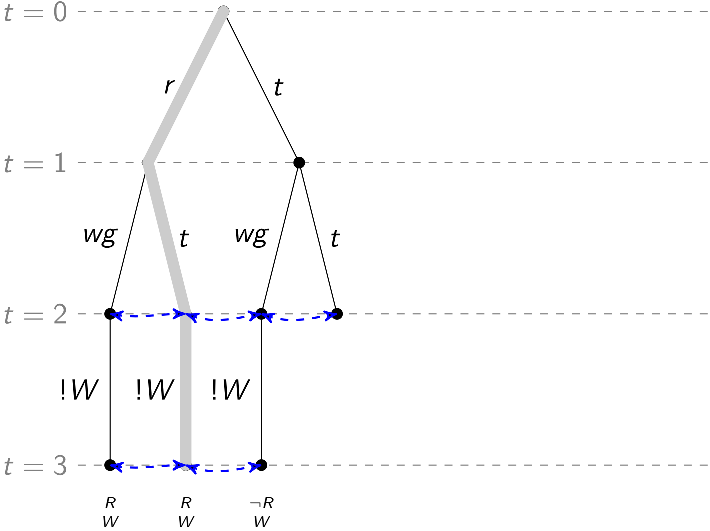
Learning from the Protocol



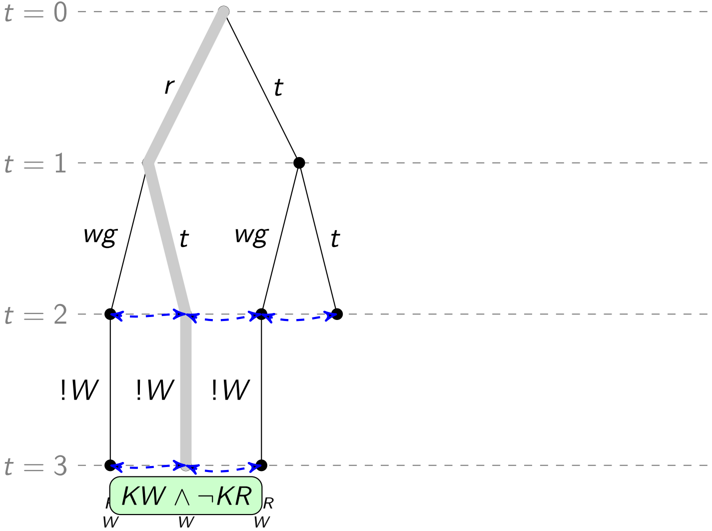
Learning from the Protocol



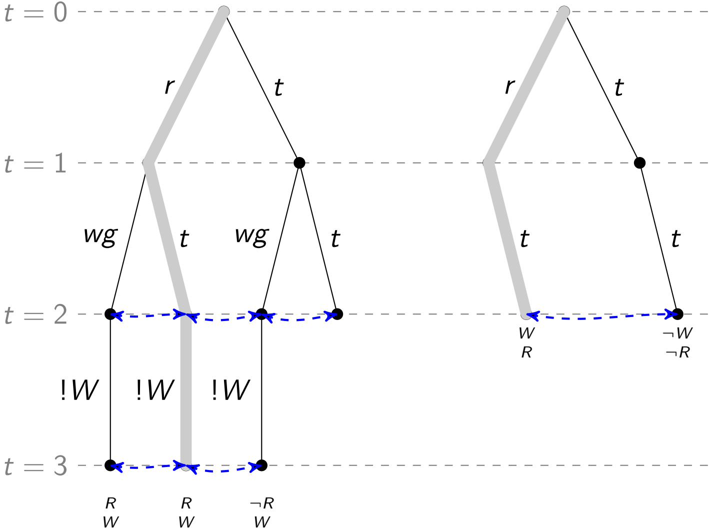
Learning from the Protocol



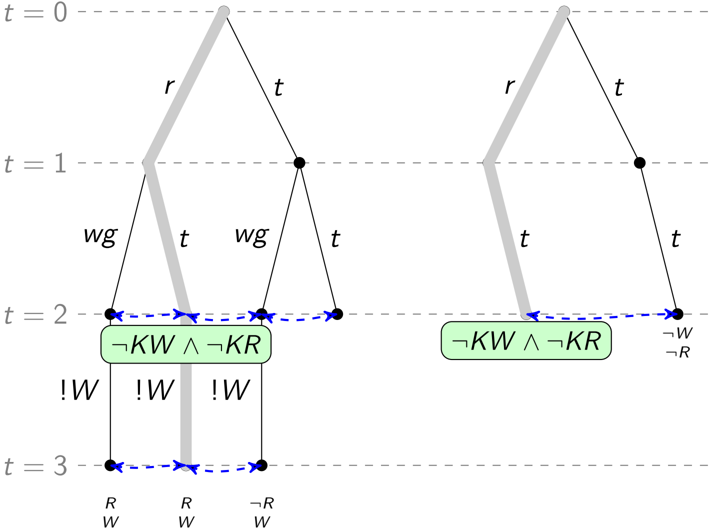
Learning from the Protocol



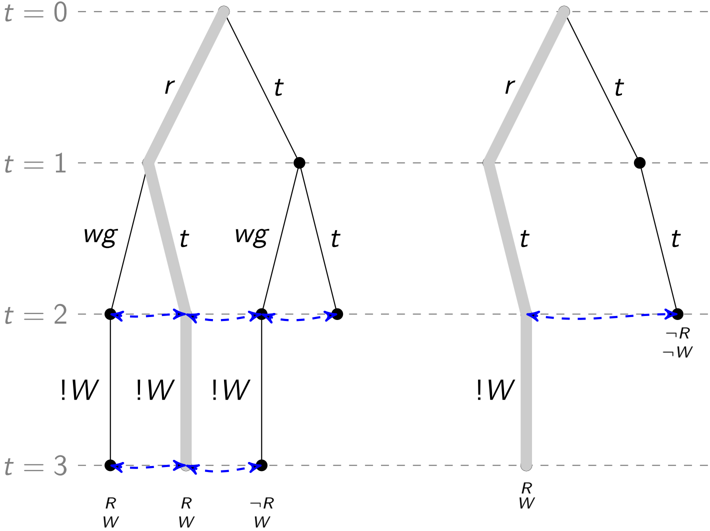
Learning from the Protocol



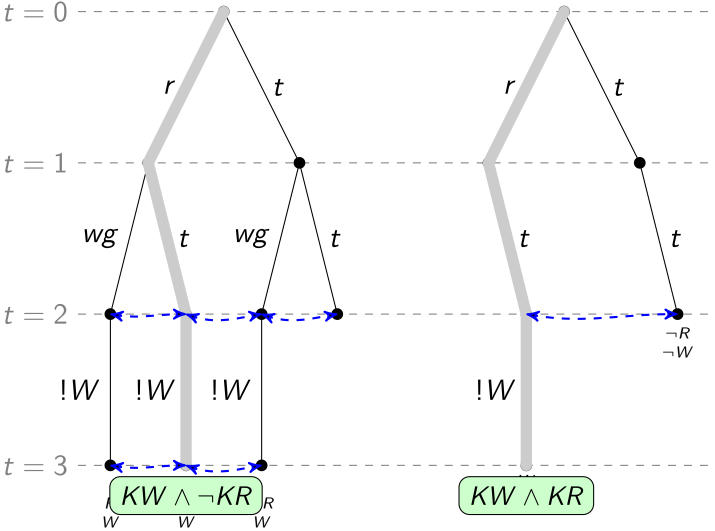
Learning from the Protocol



Learning from the Protocol



Learning from the Protocol



Actions

Assume a finite set, $Act \subseteq \Sigma$, of primitive actions.

Assume $Act = \cup_{i \in \mathcal{A}} C_i$ where $C_i \cap C_j \neq \emptyset$ for $i \neq j$.

Actions

Assume a finite set, $Act \subseteq \Sigma$, of primitive actions.

Assume $Act = \cup_{i \in \mathcal{A}} C_i$ where $C_i \cap C_j \neq \emptyset$ for $i \neq j$.

Given a finite global history H and $a \in Act$,

$$a(H) = \{H' \mid Ha \preceq H' \text{and } H' \text{a global history}\}$$

Actions

Assume a finite set, $Act \subseteq \Sigma$, of primitive actions.

Assume $Act = \cup_{i \in \mathcal{A}} C_i$ where $C_i \cap C_j \neq \emptyset$ for $i \neq j$.

Given a finite global history H and $a \in Act$,

$$a(H) = \{H' \mid Ha \preceq H' \text{and } H' \text{a global history}\}$$

$$H, t \models [a]\varphi \quad \text{iff} \quad H', t+1 \models \varphi \quad \text{for each} \quad H' \in a(H)$$

Actions

$$H, t \models [a]\varphi \;\text{ iff }\; H', t+1 \models \varphi \;\text{ for each }\; H' \in a(H)$$

Actions

$$H, t \models [a]\varphi \ \text{ iff } \ H', t+1 \models \varphi \ \text{ for each } \ H' \in a(H)$$

- When an action is performed, it is performed at the next moment of time.

Actions

$$H, t \models [a]\varphi \text{ iff } H', t+1 \models \varphi \text{ for each } H' \in a(H)$$

- When an action is performed, it is performed at the next moment of time.
- Only one agent can perform some action at any moment.

Actions

$$H, t \models [a]\varphi \text{ iff } H', t+1 \models \varphi \text{ for each } H' \in a(H)$$

- When an action is performed, it is performed at the next moment of time.
- Only one agent can perform some action at any moment.
- If no agents perform an action, then nature performs a 'clock tick'.

Actions

$$H, t \models [a]\varphi \text{ iff } H', t+1 \models \varphi \text{ for each } H' \in a(H)$$

- When an action is performed, it is performed at the next moment of time.
- Only one agent can perform some action at any moment.
- If no agents perform an action, then nature performs a 'clock tick'.
- Each agent knows *when* it can perform an action. $(\langle a_i \rangle \top \rightarrow K_i \langle a_i \rangle \top)$

Values: Informal Definition

All global histories will be presumed to have a value

Let $\mathcal{G}(H)$ be the set of extensions of (finite history) H which have the highest possible value. (Assumptions are needed to make $\mathcal{G}(H)$ well defined)

Values: Informal Definition

All global histories will be presumed to have a value

Let $\mathcal{G}(H)$ be the set of extensions of (finite history) H which have the highest possible value. (Assumptions are needed to make $\mathcal{G}(H)$ well defined)

We will say that a is good to be performed at H if $\mathcal{G}(H) \subseteq a(H)$, i.e., there are no H-good histories which do not involve the performing of a.

Knowledge Based Obligation

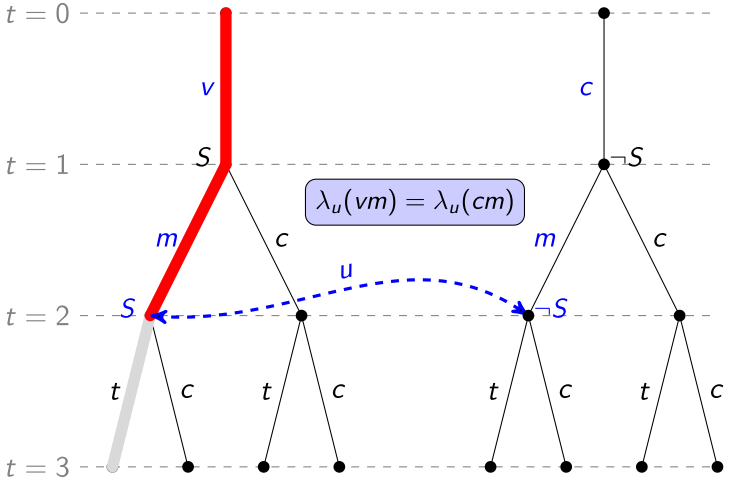
Agent i has a (knowledge based) obliged to perform action a at global history H and time t iff a is an action which i (only) can perform, and i *knows* that it is good to perform a.

For each $a \in \text{Act}$, let $G(a)$ be a formula:

$$H, t \models G(a) \text{ iff } \mathcal{G}(H_t) \subseteq a(H_t)$$

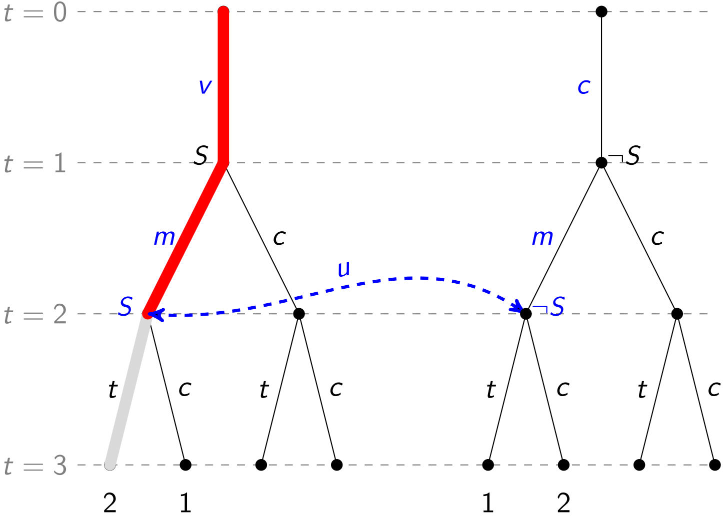
Then we say that i is obliged to perform action a (at H, t) if $K_i(G(a))$ is true (at H, t).

Example 2

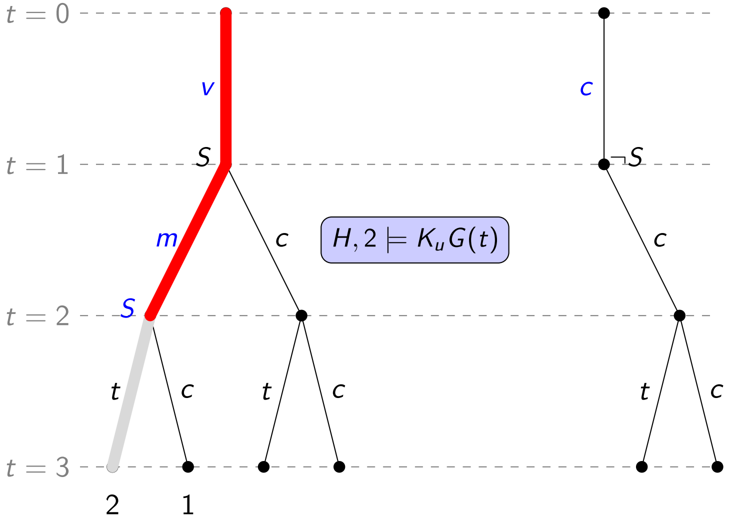


$t = 0$
$t = 1$
$t = 2$
$t = 3$
v
c
S
$\neg S$
$\lambda_u(vm) = \lambda_u(cm)$
m
c
m
c
u
S
$\neg S$
t
c
t
c
t
c
t
c

Example 2



Example 2



Recall that Ann has the (knowledge based) obligation to tell Jill about her father's illness ($K_a G(m)$).

Recall that Ann has the (knowledge based) obligation to tell Jill about her father's illness ($K_a G(m)$).

Clearly, Ann will not be under any obligation to tell Jill that her father is ill, if Ann justifiably believes that Jill would not treat her father even if she knew of his illness.

Recall that Ann has the (knowledge based) obligation to tell Jill about her father's illness ($K_a G(m)$).

Clearly, Ann will not be under any obligation to tell Jill that her father is ill, if Ann justifiably believes that Jill would not treat her father even if she knew of his illness.

Thus, to carry out a deduction we will need to assume

$$K_j(K_u \text{sick} \leftrightarrow \bigcirc \text{treat})$$

A similar assumption is needed to derive that Jill has an obligation to treat Sam.

A similar assumption is needed to derive that Jill has an obligation to treat Sam.

Obviously, if Jill has a good reason to believe that Ann always lies about her father being ill, then she is under no obligation to treat Sam.

A similar assumption is needed to derive that Jill has an obligation to treat Sam.

Obviously, if Jill has a good reason to believe that Ann always lies about her father being ill, then she is under no obligation to treat Sam.

In other words, we need to assume

$$K_j(\text{msg} \leftrightarrow \text{sick})$$

Common Knowledge of Ethicality

These formulas can all be derived for one common assumption which we call *Common Knowledge of Ethicality*.

Common Knowledge of Ethicality

These formulas can all be derived for one common assumption which we call *Common Knowledge of Ethicality*.

1. The agents must (commonly) know the protocol.
2. The agents are all of the same "type" (social utility maximizers)

Common Knowledge of Ethicality

These formulas can all be derived for one common assumption which we call *Common Knowledge of Ethicality*.

1. The agents must (commonly) know the protocol.
2. The agents are all of the same "type" (social utility maximizers)

Alternatively, we can argue that Ann has the knowledge based obligation to send the message because she knows that upon receiving the message, Uma will **change** her intentions accordingly (and so, will adopt the intention to treat Sam).

Conclusions

Conclusions

- As you may have noticed, I did not actually present a working theory of intention revision! (in progress with Yoav Shoham)

Conclusions

- As you may have noticed, I did not actually present a working theory of intention revision! (in progress with Yoav Shoham)

- Pointers to relevant literature left out here are very welcome.

Conclusions

- As you may have noticed, I did not actually present a working theory of intention revision! (in progress with Yoav Shoham)
- Pointers to relevant literature left out here are very welcome.
- Many technical questions remain about how to define the operators $B(\varphi : a)$ and $I(a : \varphi)$, which may fit nicely with Justification Logics.

Thank You!