

Defeasible Reasoning for Intelligent Agents

Aldo Antonelli

Dept. of Logic & Philosophy of Science

University of California, Irvine

aldo@uci.edu

New York, June 14-15, 2002



## Intelligent Agents

- Whether human or mechanical, IA interact in real time in game-like situations.
- The key to successful interaction is the ability to anticipate the other agents propensity to act.
- For this, agents must be able to *model* the other agents as well as themselves.

“You must consider what other people would do if you did something different from what you actually do” (Aumann)

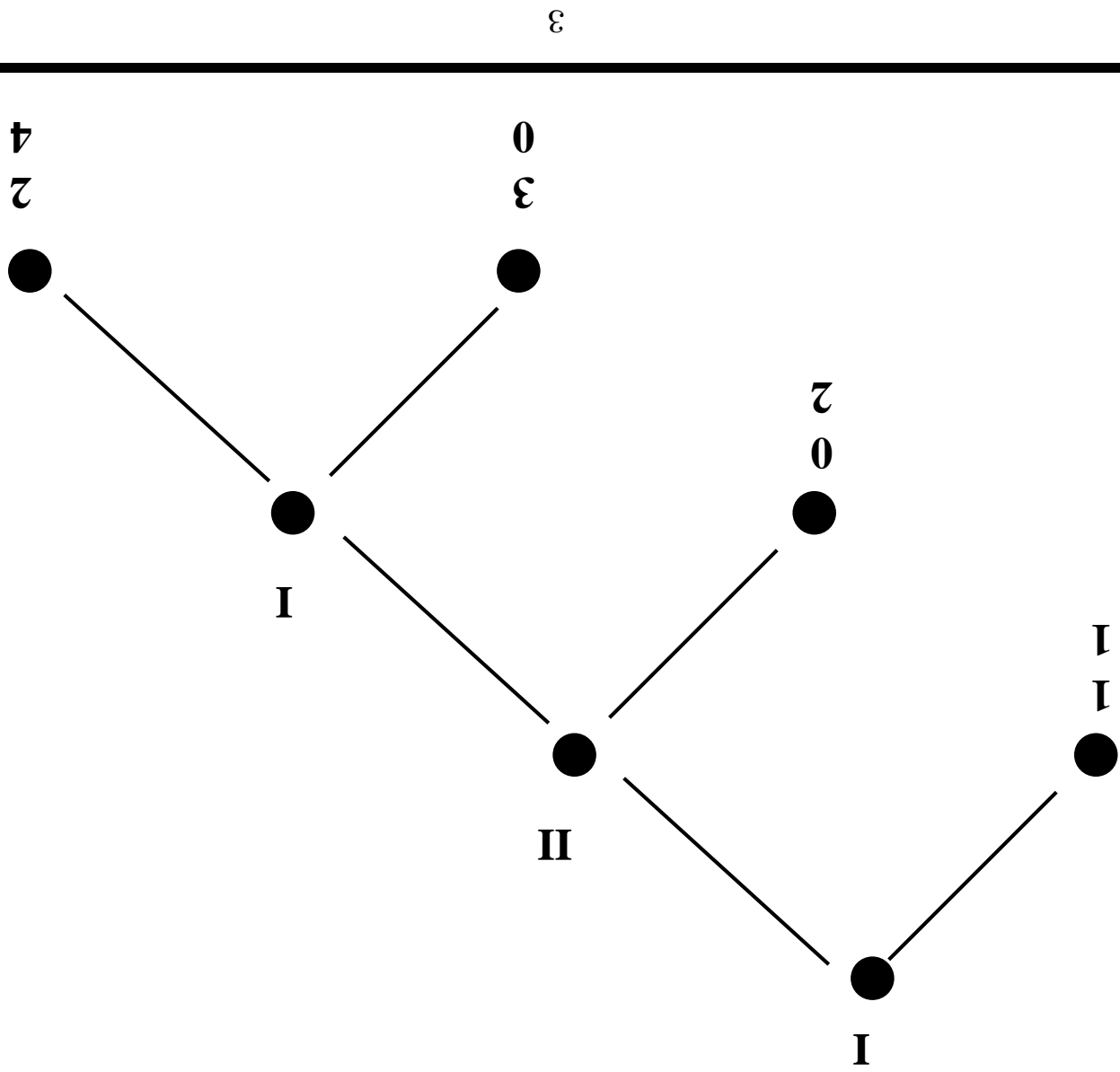
## Internal vs. external modelling

On many accounts, in order for players to infer the equilibrium play they must reason (counterfactually perhaps) about what would happen at each node, whether that node turns out eventually to lie on the equilibrium path or not:

“The specification of the equilibrium requires a description of what the agents expect to happen at each node, *were it to be reached*, even though in equilibrium play most if these nodes are never reached” (Bicchieri).

Players also build expectation about the other players' behavior, e.g., by fully expecting the other agents to be rational (utility maximizers).

A simple game



## Out-of-equilibrium play

Problems arise when agents pursue out-of-equilibrium play. Then a player's representation of the game and the other players' propensities to act must be revised to account for this new information. As is well-known, classical first-order logic, although in principle sufficient to infer the equilibrium path, is ill-suited to represent this kind of belief revision.

Several issues arise:

1. *Relevance*: We don't want the addition of a newly learned fact which is inconsistent with previously drawn inferences to *trivialize* a player's theory;
2. *Retraction*: the previously reached conclusions must actually be retracted in the light of the new evidence;
3. *Diagnostics*: The player's theory must be revised in the light of the new evidence.

## Out-of-equilibrium play, cont'd

The three requirements: Relevance, Retraction, and Diagnostics are conceptually distinct and call for different maneuvers.

1. Relevance immunizes the players from the threat of inconsistency. By itself, it can be achieved by a number of many *infra-classical* logics;
2. Retraction is a form of belief revision: it can be *extra-theoretical* & la AGM, or *intra-theoretical* as in a number of non-monotonic logics;
3. Diagnostics requires that the player's theory be modified in the appropriate fashion. Several options:

- Selten's "trembling hand"; or
- more radical revisions of the agent's theories (rationality, etc.)

While (1) and (2) are tasks for the logician, (3) requires a game theorist (but logic needs to supply the expressive resources).

## Meta-theoretic solutions

This family of solutions explicitly assign a *theory* to each agent, and deal with out-of-equilibrium play by characterizing how the theory is to be revised in the light of the new evidence:

### 1. Belief Revision: AGM;

### 2. “local theories” (Bicchieri & Antonelli 1995):

- Assign to each player a different theory of the game *at each node*;
- when an out-of-equilibrium play is observed, play continues according to the “local” theory;
- most naturally implements the trembling hand, but not only.

**Intra-theoretic solutions**

*Hypothetical* out-of-equilibrium play is represented by counterfactuals  $A > B$ , “if  $A$  were true, then so would  $B$ .”

Following the Lewis-Stalnaker approach, interpret  $A > B$  as true at a world  $w$  iff  $B$  is true at the world  $w'$  closest to  $w$  where  $A$  is true.

This approach is best suited for internal modelling: allows the players to consider what would be the case, were they to move in a certain way.

It's not obvious how to extend this to external modelling in such a way as to be robust in the presence of unexpected data.

Moreover, as is well known, the *similarity* relation is hard to capture in general.

*A better approach is given by logics for defeasible reasoning.*

## Consequence Relations

We take an abstract approach to non-monotonic logics, and consider the desirable properties of a relation of *feasible consequence*:

**Supraclassicality:** if  $\Gamma \models \phi$  then  $\Gamma \sim \phi$ .

**Reflexivity:** if  $\phi \in \Gamma$  then  $\Gamma \sim \phi$ ;

**Cut:** If  $\Gamma \sim \phi$  and  $\Gamma, \phi \sim \psi$  then  $\Gamma \sim \psi$ ;

**Cautious Monotony:** If  $\Gamma \sim \phi$  and  $\Gamma \sim \psi$ , then  $\Gamma, \phi \sim \psi$ .

In applications, it's important that  $\sim$  satisfy Cut, Reflexivity, and

Cautious Monotony (Gabbay).

For different reasons, the following are *not* desirable:

**Monotony:** If  $\Gamma \sim \phi$  and  $\Gamma \subseteq \Delta$  then  $\Delta \sim \phi$ .

**Rational Monotony:** If  $\Gamma \not\sim \neg\phi$  and  $\Gamma \sim \psi$ , then  $\Gamma, \phi \sim \psi$ .

## Stalnaker's Counter-example to RatMon

- Three composers: Verdi, Bizet, and Satie;
- Initial defeasible assumption:  $I(v), F(b), F(s)$ ;
- Further reliable information:  $C(v, b)$ ;
- No longer endorse  $I(v), F(b)$ , but still endorse  $F(s)$ ;
- Therefore, against RatMon:
  - $C(v, b) \sim F(s)$
  - $C(v, b) \not\sim \neg C(v, s)$
  - $C(v, b), C(v, s) \not\sim F(s)$

## The nature of conflict

Two kinds of conflict can arise in a defeasible setting:

1. conflicts between defeasible conclusions and hard facts;

2. conflicts between one defeasible conclusion and another one.

In both cases steps need to be taken to preserve or restore consistency.

The nature of non-monotonic reasoning assures that conflicts of type (1)

are handled by retracting the defeasible conclusion.

Several options are available for conflicts of type (2):

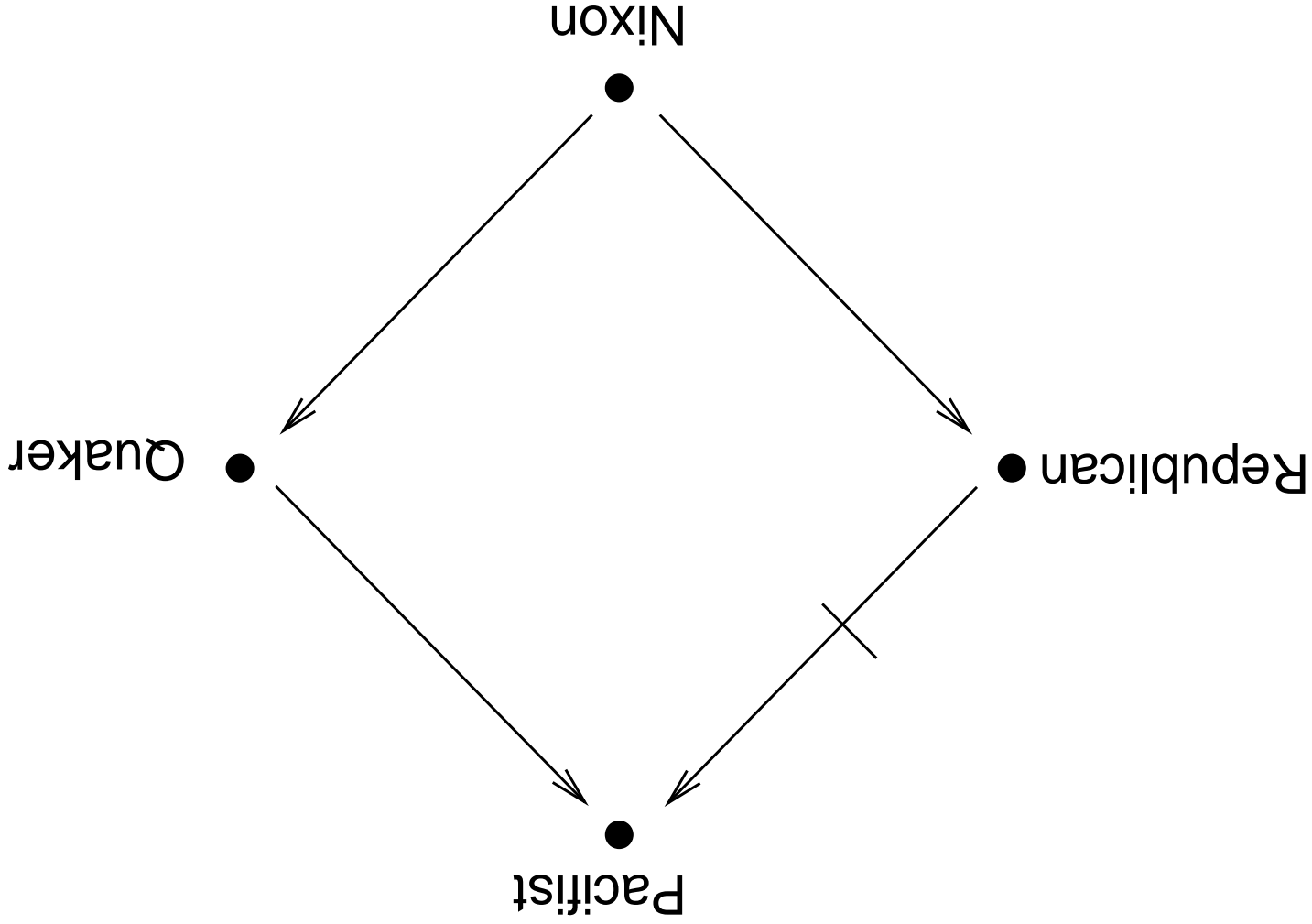
1. *bold* response (roughly): commit to a maximally consistent set of

defeasible inferences;

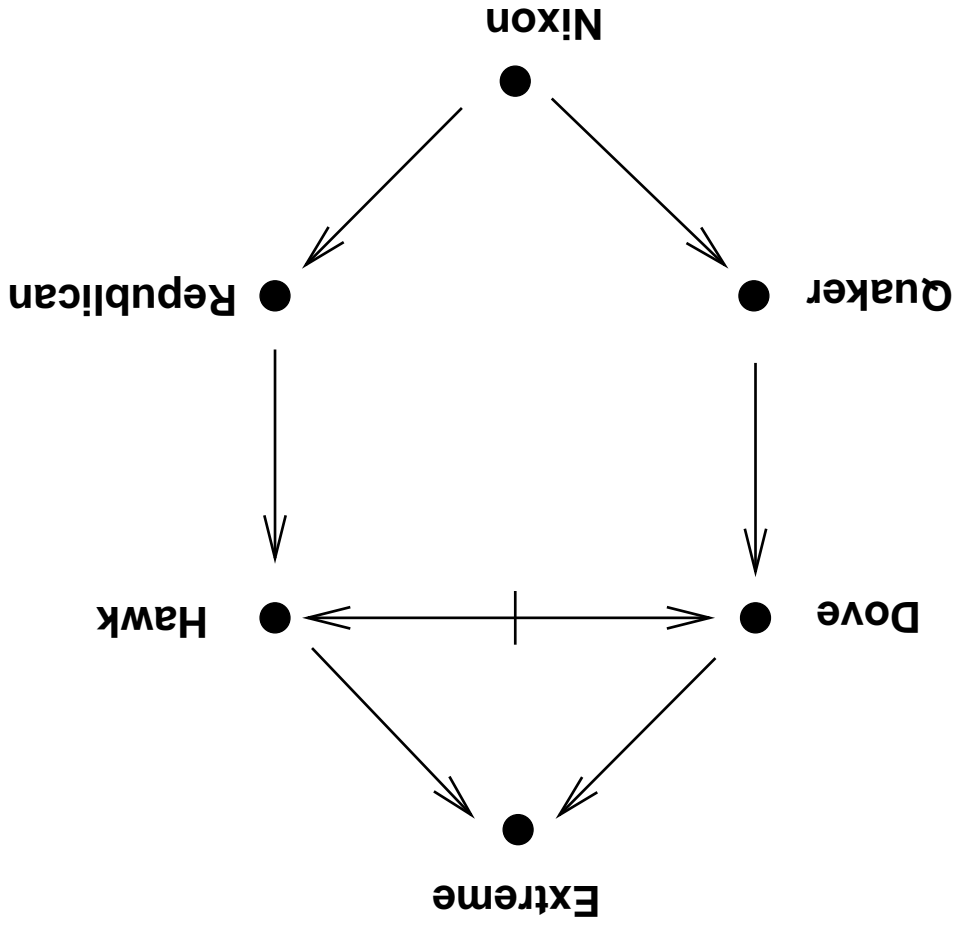
2. *cautious* response: in the face of conflicting inferences, suspend

judgment.

**The Nixon diamond**



**Horty's network**



Is Nixon "extreme"?

## Default Logic

Introduced by Reiter (1980) — extremely flexible formalism for the representation of defeasible inference. A default is a *rule* of the form:

$$\frac{\gamma}{\alpha : \beta}$$

saying that if  $\alpha$  is known and  $\beta$  is consistent with what is known, then  $\gamma$  can be inferred.  $\alpha$ ,  $\beta$  and  $\gamma$  are the prerequisite, justification, and conclusion.

A *default theory* is a pair  $(W, \Delta)$  where  $W$  is a set of sentences (a “world description”) and  $\Delta$  is a finite set of defaults. The default above is *triggered* by a pair  $(S_1, S_2)$  of sets of sentences iff  $S_1 \models \alpha$  but  $S_2 \not\models \neg\beta$ . A set  $E$  of sentences is an *extension*  $(W, \Delta)$  iff  $\mathfrak{F}(E) = E$ , where for any set  $S$  of sentences,  $\mathfrak{F}(S)$  is the smallest deductively closed set of sentences containing  $W$ , and such that if  $(S, S)$  triggers a default with consequent  $\gamma$ , then  $\gamma \in \mathfrak{F}(S)$ .

## Extensions

Default theories can have zero, one, or more extensions:

1. If  $W = \{ \alpha \}$  and  $\Delta$  contains

$$\frac{\beta}{\alpha : \beta}$$

then  $(W, \Delta)$  has *no* extensions.

2. If  $W = \{ \alpha \}$  and  $\Delta$  contains

$$\frac{\neg \gamma}{\alpha : \beta}$$

and

$$\frac{\neg \beta}{\alpha : \gamma}$$

Then  $(W, \Delta)$  has exactly *two* extensions.

Extensions are *orthogonal* and not iteratively constructed.

## Defeasible consequence for default logic

- Bold strategy: pick an extension  $E$  for  $(W, \Delta)$  and put  $(W, \Delta) \sim \phi$  iff  $\phi \in E$ ;

– Problem 1: On what basis is  $E$  picked?

– problem 2: What happens if  $(W, \Delta)$  has no extensions?

- Cautious strategy: put  $(W, \Delta) \sim \phi$  iff

$\phi \in \bigcup \{E : E \text{ is an ext for } (W, \Delta)\}$

– Problem: Cautious Monotony then fails for  $\sim$  (Makinson):

consider  $W = \emptyset$  and  $\Delta$  comprising

$$\frac{\theta}{\theta},$$

and

$$\frac{\neg\theta}{\theta \vee \eta : \neg\theta}.$$

Then  $(W, \Delta) \sim \theta$  as well as  $(W, \Delta) \sim \theta \vee \eta$ , but  $(\{\theta \vee \eta\}, \Delta) \not\sim \theta$ .

## General extensions

We look at prerequisite-free defaults of the form  $\frac{\alpha}{\beta}$ .

- A default like the above is *conflicted* in a set  $\Theta$  of defaults if  $\beta$  is inconsistent with the set of consequents of defaults in  $\Theta$ ;
- A default like the above is *pre-empted* in a set  $\Theta$  of defaults if  $\alpha$  is inconsistent with the set of consequents of defaults in  $\Theta$ .

A pair  $(\Gamma_+, \Gamma_-)$  of sets of defaults is a *general extension* for a default theory  $(W, \Delta)$  iff:

1.  $\Gamma_+$  is the set of all defaults neither conflicted in  $\Gamma_+$  nor pre-empted in  $(\Delta - \Gamma_-)$ ; and

2.  $\Gamma_-$  is the set of defaults neither conflicted nor pre-empted in  $\Gamma_+$ .

The definition can be extended to cover the case of prerequisites (Antonelli 1999)

## Features of general extensions

1. General extensions subsume extensions in Reiter's sense.
2. General extensions always exist.
3. General extensions can be obtained by a truly iterative process — although one that (except in some cases) is non-deterministic.
4. In general, the set of extensions of a default theory is not “flat” but has an interesting algebraic structure.
5. In fact, default theories of certain class (“semi-normal”) have a *unique* least general extension.
6. When a theory a least general extension one can give a natural implementation of the *cautious* approach that is not susceptible to the problem of floating conclusions.

## Examples

1. As before, let  $W = \{\alpha\}$  and let  $\Delta$  contain

$$\frac{\alpha : \beta}{\neg \gamma}, \quad \text{and} \quad \frac{\alpha : \gamma}{\neg \beta}.$$

Beside the two Reiter extensions, this theory has a general extension below both, in which neither default is triggered.

2. Consider the theory where  $W = \emptyset$  and  $\Delta$  contains

$$\frac{\beta}{\neg \beta}, \quad \text{and} \quad \frac{\alpha : \gamma}{\neg \gamma}.$$

The theory has no Reiter extension, but has a general extension (triggering the second default).

**Defeasible consequence with general extensions**

Define the relation  $\sim$  by putting  $(W, \Delta) \sim \phi$  iff for every *minimal* extension  $(\Gamma^+, \Gamma^-)$ ,  $\phi$  follows from the set of conclusions of defaults in  $\Gamma^+$ .

*The relation  $\sim$  so defined satisfies all three properties of*

*Reflexivity, Cut, and Cautious Monotony.*

Say that a default is *semi-normal* iff its justification implies its conclusion; and say that a default theory is semi-normal iff all of its defaults are. Every semi-normal default theory has a *least* general extension, which can be generated by an iterative deterministic process.

## Cautious Monotony

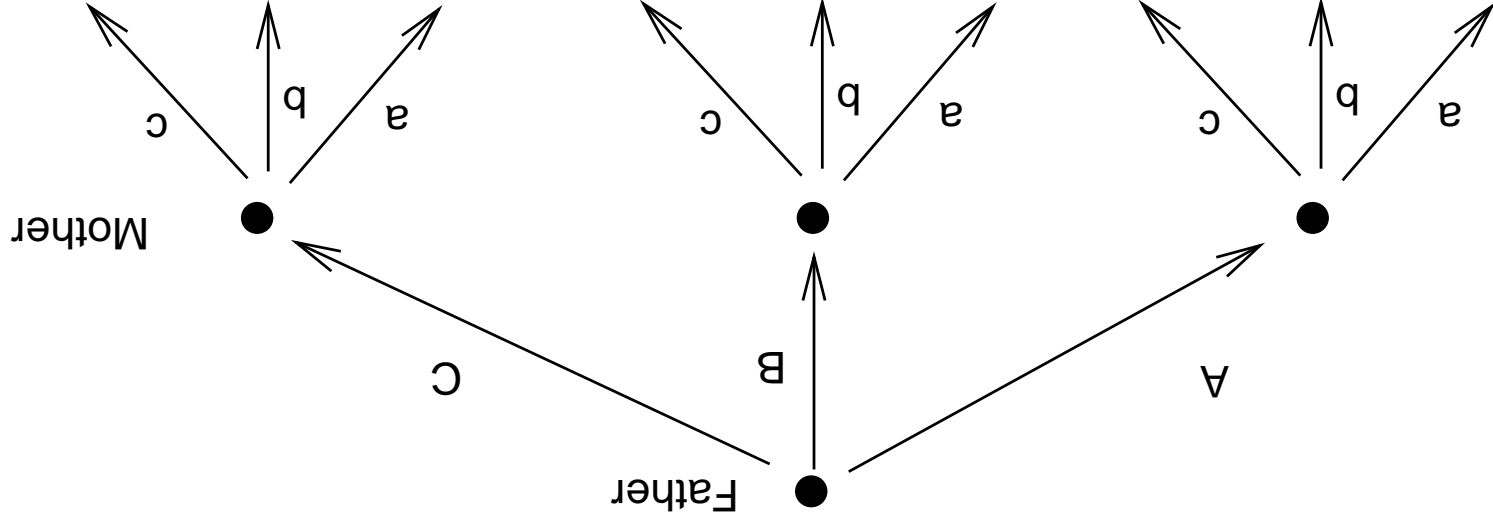
Here is the counter-example to Cautious Monotony for Reiter extensions:  
 consider  $W = \emptyset$  and  $\Delta$  comprising

$$\frac{\theta}{\theta} \quad \text{and} \quad \frac{\neg\theta}{\theta \vee \eta : \neg\theta}.$$

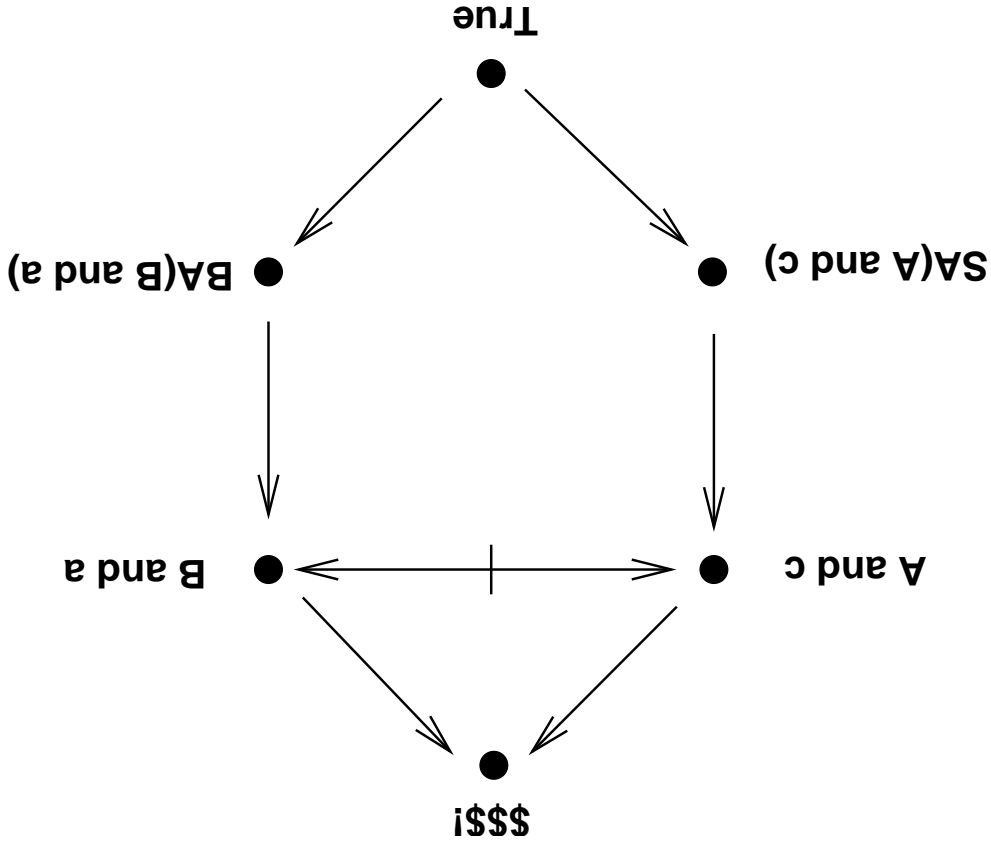
With Reiter extensions, adding  $\theta \vee \eta$  to the world description increases the number of extensions, and since Reiter extensions are flat, this knocks out  $\theta$  as a consequence.

With defeasible consequence based on general extensions, adding  $\theta \vee \eta$  also adds extensions to the theory, *but only non-minimal ones*, so that consequences are preserved.

**Horty's story: a game theoretic representation**



**Horty's story: an inferential representation**



The network can be formalized as a *semi-normal* default theory blocking the floating conclusion.

## Conclusions

- There is a lot of potential interaction between defeasible logic and game theory, especially in formalizing an agent's theory of the game and the players.
- For such an interaction to be successful, the logic must be well-behaved, and in particular it must satisfy Gabbay's 3 conditions: Reflexivity, Cut, and Cautious Monotony.
- A particularly well-behaved case with general extensions is the case of semi-normal default theories, which have a unique, iteratively generated, deterministic minimal extension.
- Although semi-normal default theories are not well-behaved under the Reiter notion of extension, it is well known that most if not all representational problems can be formalized in a semi-normal theory.