Causality in the context of multiple agents

Robert Demolombe[1]

[1]Institut de Recherche en Informatique de Toulouse

December 2013

Aim:

To find a formal definition of causality when several agents are acting together

Aim:

To find a formal definition of causality when several agents are acting together

Example:

It is forbidden to disseminate a password

Agent i has transmitted the beginning

Agent j has simultaneously transmitted the end

Formalization in the semantics of a Modal Logic

"Standard" causality definition

Agent i has caused that ϕ holds by doing action A iff

1. It is sufficient that i does A to obtain ϕ
2. It is necessary that i does A to obtain ϕ i.e.

counterfactual condition

if i had not done A then (ceteris paribus) ϕ might not have obtained

Academic example of joint actions

ϕ: there are 4 grams of poison, or more, in a given glass
Action A_n: to put n grams of poison in the glass

Case 1

Agent i and agent j have simultaneously performed A_2
Their joint action has caused that there are 4 grams of poison

Academic example of joint actions

ϕ: there are 4 grams of poison, or more, in a given glass

Action A_n: to put n grams of poison in the glass

Case 1

Agent i and agent j have simultaneously performed A_2

Their joint action has caused that there are 4 grams of poison

Case 2

Agent i has performed A_2

After: agent j has performed A_2

Agent j has caused that there are 4 grams of poison

Agent i did not cause that there are 4 grams of poison

Agent i has offered to j the opportunity to cause that there are 4 grams of poison

Academic example of joint actions

ϕ: there are 4 grams of poison, or more, in a given glass

Action A_n: to put n grams of poison in the glass

Case 1

Agent i and agent j have simultaneously performed A_2

Their joint action has caused that there are 4 grams of poison

Case 2

Agent i has performed A_2

After: agent j has performed A_2

Agent j has caused that there are 4 grams of poison

Agent i did not cause that there are 4 grams of poison

Agent i has offered to j the opportunity to cause that there are 4 grams of poison

Case 3

Agent i and agent j have simultaneously performed A_4

Does i has caused that there 4 grams of poison or more?

"Standard definition": in a counterfactual world j is acting

Then, the answer is "no" ... which is counterintuitive

Formal definition of the logic

Actions are defined by:

- the actor i
- the type of action A_6
- the effects ϕ

Agent i by doing an action of the type A_6 has brought it about that there are more than 4 grams of poison in the glass

A pair $i : A_6$ is called an "*act*"

Formal definition of the logic

Actions are defined by:

- the actor i
- the type of action A_6
- the effects ϕ

Agent i by doing an action of the type A_6 has brought it about that there are more than 4 grams of poison in the glass
A pair $i : A_6$ is called an "*act*"
Inspiration :
I. Pörn: $E_i\phi$ (no action type)
Agent i has brought it about that ϕ
K. Segerberg: $< i, A_6, p >$ (no counterfactual condition)
Agent i has performed the instance p of an action of type A_6

Language

Propositional Modal Language

$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid Done^{+}_{act}\phi \mid Done^{-}_{act}\phi \mid JE^{+}_{act}\phi \mid RJE^{+}_{act,act'} \mid SJE^{+}_{act^*}\phi$

act, act': sets of acts

Example: $act = \{i : A_2, i : B, j : A_2\}$

act^*: set of set of acts

Example: $act^* = \{\{i : A_2, i : B, j : A_2\}, \{k : A_4, k : C\}\}$.

$JE^{+}_{act}\phi$: the agents in act are going to bring it about that ϕ by doing exactly the set of acts act

Frame, Model

Frame

$F =< W, R^*_{act}, CR^*_{act-act'} >$

W: non empty set of worlds

R^*_{act} is a set of binary relations defined on $W \times W$

$R_{act}(w, w')$: performance of the set of acts *act* has started in w and ended in w'

Frame, Model

Frame

$F =< W, R^*_{act}, CR^*_{act-act'} >$

W: non empty set of worlds

R^*_{act} is a set of binary relations defined on $W \times W$

$R_{act}(w, w')$: performance of the set of acts *act* has started in w and ended in w'

$CR^*_{act-act'}$ is a set of ternary relations defined on $W \times W \times W$

$R_{act-act'}(w, w', w'')$: performance of the set of acts *act* has started in w and ended in w', and in w'' the acts in act' have not been performed (*ceteris paribus*)

w'' is a counterfactual world of w'

Frame, Model

Frame

$F =< W, R^*_{act}, CR^*_{act-act'} >$

W: non empty set of worlds

R^*_{act} is a set of binary relations defined on $W \times W$

$R_{act}(w, w')$: performance of the set of acts *act* has started in w and ended in w'

$CR^*_{act-act'}$ is a set of ternary relations defined on $W \times W \times W$

$R_{act-act'}(w, w', w'')$: performance of the set of acts *act* has started in w and ended in w', and in w'' the acts in act' have not been performed (*ceteris paribus*)

w'' is a counterfactual world of w'

Model

$M =< F, v >$ where F

v: function which assigns to each atomic proposition a subset of W

Joint Action Operator

Semantics of $JE^{+}_{act}\phi$

$act \subseteq all$

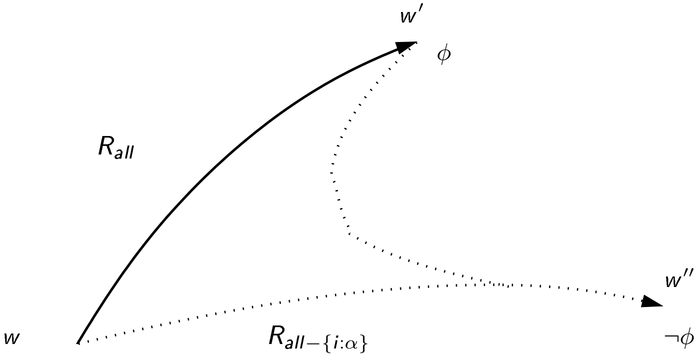
$M, w \models JE^{+}_{act}\phi$ iff

1) for all w' $(R_{all}(w, w') \quad \Rightarrow \quad M, w' \models \phi)$ and

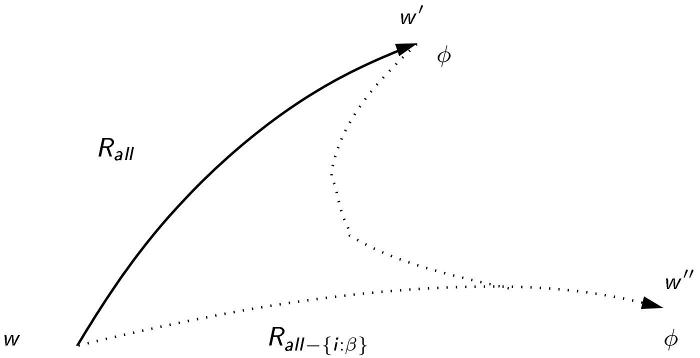
2) for all $i : \alpha$ in act, there exist w' and w'' such that $(R_{all-\{i:\alpha\}}(w, w', w'') \;\; and \;\; M, w'' \models \neg\phi)$ and

3) for all $j : \beta$ in all which are not in act for all w' and w'' $(R_{all-\{j:\beta\}}(w, w', w'') \quad \Rightarrow \quad M, w'' \models \phi)$.

1) guarantees that the set of acts *act* is sufficient to obtain ϕ and that the other acts in *all* do not prevent their performance



2) for all $i : \alpha$ in *act*, there exist w' and w'' such that
$(R_{all-\{i:\alpha\}}(w, w', w'')$ *and* $M, w'' \models \neg\phi)$
2) guarantees that every act $i : \alpha$ in *act* is necessary to obtain ϕ



3) for all $j : \beta$ in *all* which are not in *act* for all w' and w''
$(R_{all-\{j:\beta\}}(w, w', w'') \Rightarrow M, w'' \models \phi)$.
3) guarantees that the acts $j : \beta$ which are not in *act* are not necessary to obtain ϕ

Joint Action Operator

Theorems Non monotonicity property

If $act' \subset act$, we have: (NM1) $\models JE^{+}_{act}\phi \rightarrow \neg JE^{+}_{act'}\phi$.

If $act \subset act'$, we have: (NM2) $\models JE^{+}_{act}\phi \rightarrow \neg JE^{+}_{act'}\phi$.

$JE^{+}_{act}\phi$ characterizes exaclty the set of acts that have caused ϕ

Joint Action Operator

Theorem Closure

(CL) $\models (JE^{+}_{act_1}\phi \wedge JE^{+}_{act_2}\psi) \rightarrow JE^{+}_{act_1 \cup act_2}(\phi \wedge \psi)$

(CL) does not contradict non monotonicity because we have:

Joint Action Operator

Theorem Closure

(CL) $\models (JE^{+}_{act_1}\phi \wedge JE^{+}_{act_2}\psi) \rightarrow JE^{+}_{act_1 \cup act_2}(\phi \wedge \psi)$

(CL) does not contradict non monotonicity because we have:

If $\vdash \phi \leftrightarrow \psi$ and $act_1 \neq act_2$, then $\models (JE^{+}_{act_1}\phi \wedge JE^{+}_{act_2}\psi) \rightarrow \bot$

If $\not\vdash \phi \rightarrow \psi$, then $\not\models JE^{+}_{act_1 \cup act_2}(\phi \wedge \psi) \rightarrow JE^{+}_{act_1 \cup act_2}\phi$

Extended example

Sets of acts

$badmen = \{John : A_2, Jack : A_2\}$,

$badwomen = \{Mary : A_1, Miriam : A_3\}$

B: to put wine in the glass

C: to put whisky in the glass

$others = \{Robert : B, Andrew : C\}$

$bad = \{badmen, badwomen\}$

ϕ: there is at least 4 grams of poison in the glass

If the set of all the acts performed in w is: $all = badmen \cup others$

we have:

$M, w \models JE^{+}_{badmen}\phi$

Extended example

Sets of acts

$badmen = \{John : A_2, Jack : A_2\}$,

$badwomen = \{Mary : A_1, Miriam : A_3\}$

B: to put wine in the glass

C: to put whisky in the glass

$others = \{Robert : B, Andrew : C\}$

$bad = \{badmen, badwomen\}$

ϕ: there is at least 4 grams of poison in the glass

If the set of all the acts performed in w is: $all = badmen \cup others$ we have:

$M, w \models JE^{+}_{badmen}\phi$

If the set of all the acts performed in w is:

$all = badmen \cup badwomen \cup others$ we have:

$M, w \models \neg JE^{+}_{badmen}\phi$ and $M, w \models \neg JE^{+}_{badwomen}\phi$

counterintuitive

Restricted Joint Action Operator

$RJE^{+}_{act,act'}$: the agents in *act* are going to bring it about that ϕ by doing exactly *act* while the acts in *act'* are not performed

Like JE^{+}_{act} except that in 2) and 3) *all* is replaced by $all \setminus act'$

Restricted Joint Action Operator

$RJE^{+}_{act,act'}$: the agents in *act* are going to bring it about that ϕ by doing exactly *act* while the acts in *act'* are not performed

Like JE^{+}_{act} except that in 2) and 3) *all* is replaced by $all \setminus act'$

$M, w \models RJE^{+}_{act,act'}\phi$ iff
1) for all w' ($R_{all}(w, w') \quad \Rightarrow \quad M, w' \models \phi$) and
2) for all $i : \alpha$ in *act*, there exist w' and w'' such that
$(R_{(all \setminus act') - \{i:\alpha\}}(w, w', w'') \;\; and \;\; M, w'' \models \neg\phi)$ and
3) for all $j : \beta$ which is not in *act* for all w' and w''
$(R_{(all \setminus act') - \{j:\beta\}}(w, w', w'') \quad \Rightarrow \quad M, w'' \models \phi))$

Restricted Joint Action Operator

$RJE^{+}_{act,act'}$: the agents in *act* are going to bring it about that ϕ by doing exactly *act* while the acts in act' are not performed

Like JE^{+}_{act} except that in 2) and 3) *all* is replaced by $all \setminus act'$

$M, w \models RJE^{+}_{act,act'}\phi$ iff
1) for all w' ($R_{all}(w, w') \;\Rightarrow\; M, w' \models \phi$) and
2) for all $i : \alpha$ in *act*, there exist w' and w'' such that
$(R_{(all \setminus act') - \{i:\alpha\}}(w, w', w'')$ *and* $M, w'' \models \neg\phi)$ and
3) for all $j : \beta$ which is not in *act* for all w' and w''
$(R_{(all \setminus act') - \{j:\beta\}}(w, w', w'') \;\Rightarrow\; M, w'' \models \phi))$

Example: $act = badmen$, $act' = badwomen$,
$all = badmen \cup badwomen \cup others$
2) $all \setminus act' = badmen \cup others$, in w'' no bad woman is acting
3) $j : \beta$ may be any act in *badwoman* or *other*
We have:
$M, w \models RJE^{+}_{badmen,badwomen}\phi$
$M, w \models RJE^{+}_{badwomen,badmen}\phi$

Set of Joint Action Operator

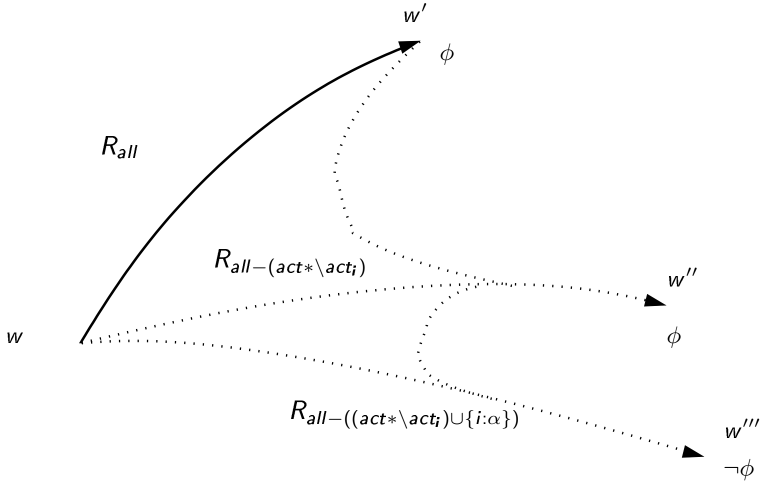
$SJE^{+}_{act^*}\phi$: every member *act* of act^* is going independently of other acts to bring it about that ϕ

Set of Joint Action Operator

$SJE^{+}_{act^*}\phi$: every member *act* of act^* is going independently of other acts to bring it about that ϕ

1) Performance of all the acts in *all* does not prevent to obtain ϕ
2) For every act_i in act^*, performance of act_i alone (*ceteris paribus*) is sufficient to obtain ϕ
3) In the context where act_i is the only element of act^* which is performed (*ceteris paribus*), performance of every act in act_i is necessary to obtain ϕ
4) There is no act, which is in *all* and which is not in act^* (*ceteris paribus*), which is necessary to obtain ϕ

$M, w \models SJE^{+}_{act^*}\phi$ iff
1) for all w' $(R_{all}(w, w') \Rightarrow M, w' \models \phi)$ and
2) for every act_i in act^*: for all w' and w''
$(R_{all-(act^*\setminus act_i)}(w, w', w'') \Rightarrow M, w'' \models \phi)$ and
3) for every $i : \alpha$ in act_i there exist w'' and w''' such that
$(R_{all-(act^*\setminus act_i)}(w, w', w'')$ and
$R_{all-((act^*\setminus act_i)\cup\{i:\alpha\})}(w, w'', w''')$ *and* $M, w''' \models \neg\phi)$ and
4) for all $j : \beta$ in *all* which are not in act^* for all w' and w''
$(R_{all-\{j:\beta\}}(w, w', w'') \Rightarrow M, w'' \models \phi)$.



In w'' the only acts in act^* which are performed are those in act_i
In w''' the same acts are performed as in w'' except $i : \alpha$

Set of Joint Action Operator

Theorem

If act_i is in act^*, then $\models SJE^+_{act^*}\phi \rightarrow RJE^+_{act_i,(act^*\backslash act_i)}\phi$

Example

$\models SJE^+_{bad}\phi \rightarrow RJE^+_{badmen,badwomen}\phi$

$\models SJE^+_{bad}\phi \rightarrow RJE^+_{badwomen,badmen}\phi$

Indirect Joint Action Operator

$IJE^{+}_{act}\phi$: the set of acts *act* is going to bring it about that further joint acts are going to bring it about that ϕ

Formal definition

$$IJE^{+}_{act}\phi \stackrel{\text{def}}{=} JE^{+}_{act}(JE^{+}_{act_1}(JE^{+}_{act_2} \dots (JE^{+}_{act_n}\phi)\dots))$$

Indirect Joint Action Operator

$IJE^{+}_{act}\phi$: the set of acts *act* is going to bring it about that further joint acts are going to bring it about that ϕ

Formal definition

$IJE^{+}_{act}\phi \stackrel{\text{def}}{=} JE^{+}_{act}(JE^{+}_{act_1}(JE^{+}_{act_2}\ldots(JE^{+}_{act_n}\phi)\ldots))$

Theorem

$M, w \models IJE^{+}_{act}\phi$ entails (*informally*):
1) performance of the sequence : act, act_1, ... ,act_n is sufficient to obtain ϕ
2) every $i:\alpha$ in act is necessary to cause performance of the sequence act_1, ... ,act_n
3) if $j:\beta$ is not in act, then $j:\beta$ performance is not necessary to have 1)

Conclusion

The operator $JE^{+}_{act}\phi$ characterizes exactly the set of acts that have caused ϕ

(no evaluation of the contribution of each act)

The operator $SJE^{+}_{act^*}\phi$ characterizes a set of set of acts such that every element in act^* causes ϕ (in a similar sense as in $JE^{+}_{act}\phi$)

The operator $IJE^{+}_{act}\phi$ characterizes indirect joint acts

Further works

Relationships with responsibility

In $act*$ there is no assumption about coordination between agents in a set

Example. Two representations of the same situation:

$badmen = \{John : A_2, Jack : A_2\}$,

$badwomen = \{Mary : A_1, Miriam : A_3\}$

versus

$fairhair = \{John : A_2, Mary : A_1\}$,

$brownhair = \{Jack : A_2, Miriam : A_3\}$