Causality in the context of multiple agents

Robert Demolombe[1]

Institut de Recherche en Informatique de Toulouse
France
robert.demolombe@orange.fr

Abstract. The formal definition of causality raises non trivial issues in the case of several agents acting together. Several action operators are defined in the semantics of a multi modal logic. The approach which is proposed is an extension to several agents of the "*bringing it about*" operators. A joint action operator is defined which holds the property of non monotonicity with respect to sets of agents. It is refined in a restricted joint operator for cases where several sets of agents cause independently a state of affairs and it is extended to sets of agents who are acting indirectly.
The formal definitions are evaluated with respect to several typical case studies and a detailed comparison with other approaches based on the STIT operators is presented.

1 Introduction

To assign responsibilities to agents we have to know who are the agents who have caused some damages. Then, even if responsibility cannot be identified to causality, there are deep relationships between norms and causality (see [21, 17, 13, 18, 5, 9, 8]) and the notion of causality plays a very important role when one wants to assign responsibilities.

In the case where several agents are acting together or when they are interacting it may be not easy to find who has caused such or such sate of affairs.

Let's consider, for instance, an academic example proposed by Lindahl in [14] where to kill a person it is sufficient that he absorbes 4 grams of poison and the agents i and j have simultaneously put 2 grams of poison in agent l's glass and l drinks what is in his glass and l dies. In that case neither i nor j have caused that there was 4 grams of poison in the glass but this state of affair has been obtained by their joint actions. Nevertheless, it is not the case that they have caused that l died because it might have been the case that l did not drink what is in his glass.

Let's assume now that, instead of acting simultaneously, i in a first step puts 2 grams of poison in the glass and that in the second step j add 2 grams more of poison. Does that situation is the same as the previous one? Certainly not, because i is not responsible of the fact that j has added 2 more grams of poison, and then j is the only responsible of the fact that there are 4 grams of poison in the glass.

Another, non trivial situation is when i and j put each one simultaneously 4 grams of poison in the glass. According to most of the definitions of causality an agent has caused some state of affairs if in a situation where he did not act, *ceteris paribus*, this state of affairs would not be obtained. In that example, if i did not have put poison in the glass would it have been the case that there are not 4 grams of poison in the glass? The answer is "no" because in the counterfactual situation where i is not acting, j is acting. The same kind of argument could be used to conclude that j has not caused that there are 4 grams of poison in the glass, and the final conclusion would be that neither i nor j are responsible of this state of affairs.

Of course, there is something wrong in this kind of argumentation. However, it is not trivial to make explicit where the mistake is. That is the reason why we need to reconsider the formal definition of causality in the case where several agents are acting "together". That is the aim of the work presented in this paper.

In section 2 is presented the logical framework which is intended to help reasoning about causality in the context of multi agents. This framework is applied in section 3 to the case study we have presented just before and it is compared in section 4 to other ones which have similar objectives.

2 Logical framework

The logical framework takes inspiration from von Wright [21], Pörn [17] and Hilpinen [11]. It can also be viewed as an extension to multiple agents of the framework proposed by Demolombe in [8]. The logic will be defined only in the semantics since the main goal is to try to clarify the meaning of the concepts (interesting surveys of the different approaches of the logic of action can be found in [11], [2] and [19]).

A basic idea is that the semantics of actions is defined, in the logical framework, by the effects that are caused by these actions and also by the name of the actions whose meaning is defined outside the logical framework.

From a theoretical point of view we accept that the meaning of an action could be completely defined by the set of all the effects that are caused by this action. For instance if we ask: *what is the meaning of the action which is called "to close the door"?*, the answer can be: *it is an action which has the effect that the door is closed.*

Nevertheless, there are many actions whose complete definition in terms of effects would be quite complex. For example, if we want to distinguish the actions: *to pay 100 euros cash*, *to pay 100 euros by credit card* and *to pay 100 euros by cheque*, it would be quite heavy to make explicit in the logical framework the effects that allow us to distinguish these actions. It is certainly easier to define, for instance, the action *to pay 100 euros by credit card* on one hand by a proposition like: *the amount of credit of the creditor has been increased by 100 euros* and, on the other hand, by the action name: *paying by credit card.*

However, it may be that in some contexts it is irrelevant to distinguish some actions and for that reason action names can be understood as the name of action types. For instance, in some context it may be irrelevant to distinguish the actions: *to pay 100 euros with 2 bank notes of 50 euros* and *to pay 100 euros with 10 bank notes of 10 euros*. Then, the fact that performance of the action *to pay 100 euros cash* can lead to several different worlds is not related to any assumption about non determinism but to the fact that action names denotes action types and not instances of actions.

Of course, we also need to identify who is the agent performing an action.

The formal consequences of this approach is that action operators are identified by the effects they cause and also by the name of an agent and of an action type (when there is no ambiguity we use the term "action" instead of "action type"). A pair: <agent, action type> is called an "act".

To define the language of the logic we have adopted the following notations.

$ATOM$: set of atomic propositions denoted by p, q, r...

$AGENT$: set of agents denoted by i, j, k, l,...

$ACTION$: set of actions denoted by α, β, γ, δ,...

ACT: set of acts denoted by pairs of the form: $i : \alpha$, where i is in $AGENT$ and α is in $ACTION$.

$SACT$: set of sets of acts denoted by act_1, act_2, act_3,...

$SSACT$: set of sets of sets of acts denoted by act_1^*, act_2^*, act_3^*,...

In the following it will be assumed that the set of agents $AGENT$ is finite.

An example of member of $SACT$ is: $\{i : \alpha, i : \beta, j : \gamma\}$.

An example of member of $SSACT$ is: $\{\{i : \alpha, i : \beta, j : \gamma\}, \{k : \alpha, k : \delta\}\}$.

The language L is the set of formulas defined by the following BNF:

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

where p ranges over $ATOM$, act and act' range over $SACT$ and act^* ranges over $SSACT$.

It is assumed that act, act' and act^* are not empty sets.

The intuitive reading of the modal operators is:

$Done^+_{act}\phi$: the agents in act are going to do the acts in act and after their performance ϕ will be true.

$Done^-_{act}\phi$: the agents in act have just done the acts in act and before their performance ϕ was true.

$JE^+_{act}\phi$: the agents in act are going to bring it about that ϕ by doing exactly the set of acts act (the "J" in JE stands for "joint" acts).

$RJE^+_{act,act'}$: the agents in act are going to bring it about that ϕ by doing exactly the set of acts act while the acts in act' are not performed (the "R" in RJE stands for "restricted" joint acts). It is assumed that we have: $act \cap act' = \emptyset$.

$SJE^+_{act^*}\phi$: every member act of act^* is going to bring it about that ϕ (the "SJ" in SJE stands for "set of joint" acts).

Definition 1 *A* **frame** *F is a tuple $F =< W, R^*_{act}, CR^*_{act-act'} >$, where W is a non empty set of worlds, act and act' are subsets of ACT, R^*_{act} is a set of binary relations defined on $W \times W$ and $CR^*_{act-act'}$ is a set of ternary relations defined on $W \times W \times W$.*

A **model** *M is a tuple $M =< F, v >$, where F is a frame and v is a function which assigns to each atomic proposition a subset of W.*

The intuitive meaning of these relations is:

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

$R_{act-act'}(w, w', w'')$ iff $R_{act}(w, w')$ and the only difference between w'' and w' is that in w'' none of the acts in act' have been performed and the acts in $act \setminus act'$ have been performed in w'' in the same way[1] as they have been performed in w'.

In informal terms we have the constraints: $R_{act-act'}(w, w', w'')$ entails $R_{act}(w, w')$ and $R_{act-act'}(w, w'')$ and, since the past is assumed to be unique, we have for every w, w', w'', act_1 and act_2: $R_{act_1}(w', w)$ and $R_{act_2}(w'', w)$ entail that $w' = w''$.

In the following, if $R_{act-act'}(w, w', w'')$ holds we shall say, for short, that w'' *is a counterfactual world of w' with respect to act'*.

Definition 2 *The fact that a formula ϕ is* **true** *in a world w of a model M is denoted by: $M, w \models \phi$. The fact that ϕ is a* **valid** *formula, that is, ϕ is true in every world of every model, is denoted by: $\models \phi$.*

The truth conditions for atomic propositions and logical connectives are defined as usual.

In the following *all* will be used to denote the set of all the acts that the agents start to perform in a given world[2].

Definition 3 *We have the following truth conditions for the action operators of the kind Done.*

$M, w \models Done^{+}_{act}\phi$ *iff*
there exists a world w' such that $R_{act}(w, w')$ and for all w' ($R_{act}(w, w') \Rightarrow M, w' \models \phi$).

$M, w \models Done^{-}_{act}\phi$ *iff*
there exists w' such that ($R_{act}(w', w)$ and $M, w' \models \phi$).

There is a large number of works about causality in the case of a unique acting agent. These works have tried to give a formal definition of the notion of counterfactual worlds (see [21, 11, 19]) however this definition has to be revisited in the context of multiple agents.

Definition 4 *We have the following truth conditions for the operators of* **joint action** *of the kind JE.*

$M, w \models JE^{+}_{act}\phi$ *iff*
1) for all w' ($R_{all}(w, w') \Rightarrow M, w' \models \phi$) and

[1] To be more precise: the instances of the action types performed in w'' are the same as in w'.

[2] In the following sentences of the kind "*acts that the agents start to perform in a given world*" will be abbreviated in "*acts that the agents perform in a given world*".

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

The intuitive meaning of condition 1) is that the set of acts in *all* is sufficient to guarantee that ϕ is obtained. This condition is not redundant with the condition 3) because it guarantees that the simultaneous performance of all the acts which are in *all* does not prevent to obtain ϕ.

The intuitive meaning of condition 2) is that every act $i : \alpha$ in *act* is necessary to obtain ϕ, that is, if $i : \alpha$ is the only act in *all* which is not performed, there exists a world where ϕ is not obtained[3].

The intuitive meaning of condition 3) is that for every act $j : \beta$ which is not in *act*, if all the acts in *all* but $j : \beta$ are performed, ϕ is always obtained, that is: $j : \beta$ is not necessary to obtain ϕ.

We have adopted the following notation in the case where *act* contains a single act: $E^+_{i:\alpha}\phi \stackrel{\text{def}}{=} JE^+_{\{i:\alpha\}}\phi$.

Theorem 1 *If $act' \subset act$, we have: (NM1) $\models JE^+_{act}\phi \to \neg JE^+_{act'}\phi$.*
If $act \subset act'$, we have: (NM2) $\models JE^+_{act}\phi \to \neg JE^+_{act'}\phi$.

Proof of (NM1). From $act' \subset act$ we can infer that there is an act $i : \alpha$ which is in *act* and which is not in act'. From the truth condition 2), for the modality JE^+_{act}, $M, w \models JE^+_{act}\phi$ entails that there exists two worlds w'_0 and w''_0 such that $R_{all-\{i:\alpha\}}(w, w'_0, w''_0)$ *and* $M, w''_0 \models \neg\phi$. Therefore the truth condition 3), for the modality $JE^+_{act'}$, is false in M, w due to the fact that $i : \alpha$ is not in act' and we have $M, w''_0 \models \neg\phi$. Therefore, for all M and w we have: $M, w \models JE^+_{act}\phi \to \neg JE^+_{act'}\phi$.

Proof of (NM2). From $act \subset act'$ we can infer that there is an act $i : \alpha$ which is in act' and which is not in *act*. From the truth condition 3), for the modality JE^+_{act} the proposition $M, w \models JE^+_{act}\phi$ entails that for all w' and w'' we have: $(R_{all-\{i:\alpha\}}(w, w', w'') \quad \Rightarrow \quad M, w'' \models \phi))$. Therefore the truth condition 2) for the modality $JE^+_{act'}$ is false in M, w due to the act $i : \alpha$. Therefore, for all M and w we have: $M, w \models JE^+_{act}\phi \to \neg JE^+_{act'}\phi$.

The intuitive reading of Theorem 1 is that if $JE^+_{act}\phi$ holds, ϕ has been caused by the performance of all the acts in *act* and by no other act. In more formal terms, if $act' \neq act$ we have : $\models JE^+_{act}\phi \to \neg JE^+_{act'}\phi$.

Theorem 2 *We have:*
(CL) $\models (JE^+_{act_1}\phi \wedge JE^+_{act_2}\psi) \to JE^+_{act_1 \cup act_2}(\phi \wedge \psi)$.
(CL') $\models (JE^+_{act}\phi \wedge JE^+_{act}\psi) \to JE^+_{act}(\phi \wedge \psi)$.
(NCL) If we have $\models \phi \to \psi$ and $act_2 \not\subseteq act_1$ or $\models \psi \to \phi$ and $act_1 \not\subseteq act_2$, we have: $\models (JE^+_{act_1}\phi \wedge JE^+_{act_2}\psi) \to \bot$.

[3] It is worth noting that, if the fact that some $i : \alpha$ is not performed entails that ϕ is not obtained in w'', it follows that the same property holds for any subset X of *act* that contains $i : \alpha$.

Proof of (CL). The truth conditions 1) for $JE^+_{act_1}\phi$ and $JE^+_{act_2}\psi$ entail that in w' we have both ϕ and ψ. If some $i:\alpha$ is in $act_1 \cup act_2$, it is either in act_1 or in act_2. If it is in act_1, from $JE^+_{act_1}\phi$ we can infer that there is some w'' where we have $\neg\phi$, which entails $\neg(\phi \wedge \psi)$. We can draw the same conclusion if $i:\alpha$ is in act_2. Therefore condition 2) is satisfied for $JE^+_{act_1 \cup act_2}(\phi \wedge \psi)$. The condition 3) is satisfied because if some $j:\beta$ is not in $act_1 \cup act_2$ it is not in act_1 and from $JE^+_{act_1}\phi$ we can infer that in w'' ϕ is true. We can derive in the same way that in w'' ψ is true. Therefore, $\phi \wedge \psi$ is true in w''.

Proof of (CL'). (CL') is an instance of (CL) for $act_1 = act_2$.

Proof of (NCL). Let's assume that we have $\models \phi \to \psi$ and $\neg(act_2 \subseteq act_1)$. From the assumption $\neg(act_2 \subseteq act_1)$ we infer that there exists some act $i:\alpha$ such that $i:\alpha \in act_2$ and $i:\alpha \notin act_1$. If in some M,w we have $M,w \models JE^+_{act_1}\phi \wedge JE^+_{act_2}\psi$, since we have $i:\alpha \notin act_1$, from the truth condition 3) for $M,w \models JE^+_{act_1}\phi$ we infer that for every w' and w'' such that $R_{all}(w,w')$ and $R_{all-\{i:\alpha\}}(w,w',w'')$ we have $M,w'' \models \phi$, and (a) $M,w'' \models \psi$, since we have $\models \phi \to \psi$. Since we have $i:\alpha \in act_2$, from the truth condition 2) for $M,w \models JE^+_{act_2}\psi$ we infer that there exist some w' and w'' such that $R_{all}(w,w')$ and $R_{all-\{i:\alpha\}}(w,w',w'')$ and (b) $M,w'' \models \neg\psi$. The facts (a) and (b) lead to an inconsistency. The proof is the same in the case of $\models \psi \to \phi$ and $\neg(act_1 \subseteq act_2)$.

It is worth noting that from (NCL) if $act_1 \neq act_2$ we have $\models (JE^+_{act_1}\phi \wedge JE^+_{act_2}\phi) \to \bot$. From (CL) we also have: $\models (JE^+_{act_1}\phi \wedge JE^+_{act_2}\phi) \to JE^+_{act_1 \cup act_2}\phi$. This property does not contradict the non monotonicity theorems (NM1) and (NM2). Indeed, if $act_1 \neq act_2$ the antecedent $JE^+_{act_1}\phi \wedge JE^+_{act_2}\phi$ of the implication is inconsistent, and if $act_1 = act_2$ we have: $act_1 \cup act_2 = act_1$.

Theorem 3 *We have:*
($\neg$N) $\not\models JE^+_{act}(\top)$.
($\neg$DM) If there are more than one acts in act and we have
$\not\models \phi \to \psi$ *and* $\not\models \psi \to \phi$, *we have:*
$\not\models JE^+_{act}(\phi \wedge \psi) \to (JE^+_{act}\phi \vee JE^+_{act}\psi)$.
($\neg$M) If we have $\not\models \phi \to \psi$ *and* $\not\models \psi \to \phi$, *we have:*
$\not\models JE^+_{act}(\phi \wedge \psi) \to JE^+_{act}\phi$ *and* $\not\models JE^+_{act}(\phi \wedge \psi) \to JE^+_{act}\psi$.

Proof of ($\neg$N). It is direct consequence of condition 2).

Proof of ($\neg$DM). Let's assume that we have $\not\models \phi \to \psi$ and $\not\models \psi \to \phi$ and for some M,w we have $M,w \models JE^+_{act}(\phi \wedge \psi)$. It may exist some act $i:\alpha$ such that there exist w' and w'' such that $R_{all}(w,w')$ and $R_{all-\{i:\alpha\}}(w,w',w'')$ and $M,w'' \models JE^+_{act}\neg\psi$ (which entails $M,w'' \models JE^+_{act}\neg(\phi \wedge \psi)$ and satisfies condition 2) for $JE^+_{act}(\phi \wedge \psi)$) and such that for every w' and w'', $R_{all}(w,w')$ and $R_{all-\{i:\alpha\}}(w,w',w'')$ entails that $M,w'' \models JE^+_{act}\phi$, since it is assumed that $\not\models \phi \to \psi$. This latter property shows that condition 2) for $M,w \models JE^+_{act}\phi$ is not satisfied and we have $M,w \not\models JE^+_{act}\phi$. We can show in the same way that there exists some other act $i':\alpha'$ which falsifies condition 2) for $M,w \models JE^+_{act}\psi$, since we have also assumed that $\not\models \psi \to \phi$ and we have $M,w \not\models JE^+_{act}\psi$.

Proof of ($\neg$M). It is a direct consequence of ($\neg$DM).

It is worth noting that (¬DM) is false in the particular case where there is only one act in *act*. Indeed, in that case we have $\models JE^+_{act}(\phi \wedge \psi) \rightarrow (JE^+_{act}\phi \vee JE^+_{act}\psi)$.

Definition 5 *An* **indirect joint action** *operator IJE is defined as follows: $IJE^+_{act}\phi$ denotes $JE^+_{act}\phi$ or there exist act' such that $IJE^+_{act}\phi$ denotes a formula of the form $IJE^+_{act}(IJE^+_{act'}\phi)$.*

The intuitive notion of indirect joint action is that the set of acts *act* is going to bring it about that further joint acts are going to bring it about that ϕ. A joint act represented by JE^+_{act} is seen as a special case of indirect joint act.

Theorem 4 *If we have: $M, w \models IJE^+_{act}\phi$ and $IJE^+_{act}\phi$ is of the form: $JE^+_{act}(JE^+_{act_1}(JE^+_{act_2}\dots(JE^+_{act_n}\phi)\dots))$, we have the following properties:*
1) for all w'_1, w'_2, ... ,w'_{n+1} $(R_{all}(w, w'_1) \wedge R_{all_1}(w'_1, w'_2) \wedge R_{all_2}(w'_2, w'_3) \dots \wedge R_{all_n}(w'_n, w'_{n+1}) \;\Rightarrow\; M, w'_{n+1} \models \phi)$.
2) for all $i : \alpha$ in act, there exist w'_1 and w''_1 such that $(R_{all}(w, w'_1)$ and $R_{all-\{i:\alpha\}}(w, w'_1, w''_1)$ and $M, w''_1 \models \neg(JE^+_{act_1}(JE^+_{act_2}\dots (JE^+_{act_n}\phi)\dots)))$.
3) for all $j : \beta$ which is not in act for all w'_1, w''_1, w''_2, ... ,w''_{n+1} $(R_{all}(w, w'_1)$ and $R_{all-\{j:\beta\}}(w, w'_1, w''_1) \wedge R_{all'_1}(w''_1, w''_2) \wedge R_{all'_2}(w''_2, w''_3) \dots \wedge R_{all'_n}(w''_n, w''_{n+1}) \;\Rightarrow\; M, w''_{n+1} \models \phi)$.

Proof. The proof is by induction on n.

For $n = 0$, $IJE^+_{act}\phi$ is $JE^+_{act}\phi$. Then, properties 1, 2 and 3 are exactly the truth conditions of $JE^+_{act}\phi$ (notice that for $n = 0$ the formula $JE^+_{act_1}(JE^+_{act_2}\dots(JE^+_{act_n}\phi)\dots)$ is ϕ).

Induction assumption: for every p such that $p \leq n$ the Theorem 4 holds.

For $p = n+1$ the form of $IJE^+_{act}\phi$ is: $JE^+_{act}(JE^+_{act_1}(JE^+_{act_2}\dots(JE^+_{act_{n+1}}\phi)\dots))$. Then, it can be rewritten as: $IJE^+_{act}(JE^+_{act_{n+1}}\phi)$.

Proof of property 1. From the induction hypothesis for all w'_1, w'_2, ... ,w'_{n+1} we have $M, w'_{n+1} \models JE^+_{act_{n+1}}\phi$. Then, from the truth condition 1) for $JE^+_{act_{n+1}}\phi$ we have:
for all w'_1, w'_2, ... ,w'_{n+2} we have $M, w'_{n+2} \models \phi$.

Proof of property 2. From the induction assumption we have: for all $i : \alpha$ in *act*, there exist w'_1 and w''_1 such that:
$(R_{all-\{i:\alpha\}}(w, w'_1, w''_1)$ *and* $M, w''_{n+1} \models \neg(JE^+_{act_1}(JE^+_{act_2}\dots(JE^+_{act_n}\phi')\dots)))$, where ϕ' is $JE^+_{act_{n+1}}\phi$. Then, the property 2) holds for $n+1$.

Proof of property 3. The proof is very close to the proof of property 1.

Definition 6 *We have the following truth conditions for the operators of* **restricted joint action** *of the kind RJE.*
$M, w \models RJE^+_{act,act'}\phi$ 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

($R_{all}(w,w')$ and $R_{(all \setminus act')-\{i:\alpha\}}(w,w',w'')$ and $M,w'' \models \neg\phi$) and
3) for all $j:\beta$ which is not in act for all w' and w''
($R_{all}(w,w')$ and $R_{(all \setminus act')-\{j:\beta\}}(w,w',w'') \Rightarrow M,w'' \models \phi$)).

The intuitive meaning of the conditions are the same as in the case of the operators JE^+_{act} except the fact that in the conditions 2) and 3) the acts in act' are not performed. The definition of the truth conditions of the operator $RJE^+_{act,act'}$ is intended to represent situations where other sets of acts than act have independently caused that ϕ is true.

For the operator $SJE^+_{act^*}$ it is assumed that the members act_i of act^* are all disjoint.

Definition 7 *We have the following truth conditions for the operators of* **set of joint actions** *of the kind SJE.*
$M,w \models SJE^+_{act^*}\phi$ *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''
($R_{all}(w,w')$ and $R_{all-(act^* \setminus act_i)}(w,w',w'') \Rightarrow M,w'' \models \phi$) and
3) for every $i:\alpha$ in act_i there exist w'' and w''' such that
($R_{all}(w,w')$ and $R_{all-(act^* \setminus act_i)}(w,w',w'')$ and $R_{all-((act^* \setminus act_i)\cup\{i:\alpha\})}(w,w'',w''')$ and $M,w''' \models \neg\phi$).

In the Definition 7 in w'' such that $R_{all-(act^* \setminus act_i)}(w,w',w'')$ the only acts in act^* which have been performed are the acts in act_i and in w''' such that $R_{all-((act^* \setminus act_i)\cup\{i:\alpha\})}(w,w'',w''')$ the only act in act_i which is not performed is $i:\alpha$.

The condition 1) is the same as in the case of the modality JE.

In the condition 2) the world w'' is a counterfactual world of w' with respect to the acts in act^* which are not in act_i, that is, roughly speaking, when act_i is the only member of act^* which is acting. The intuitive meaning of this condition is that each act_i in act^* is sufficient to obtain ϕ in w'' in the case where no other act in act^* is performed.

Notice that the condition 1) is not redundant with condition 2). Indeed, in the condition 2) the members act_i of act^* are acting independently, while the condition 1) guarantees that they can act together.

In the condition 3) the world w'' is a counterfactual world of w' and w''' is a counterfactual world of w'' with respect to the act $i:\alpha$. That is, in w''' none of the acts in $act^* \setminus act_i$ are performed and all the acts in act_i but $i:\alpha$ are performed. The intuitive meaning of this condition is that all the acts in act_i are necessary to obtain ϕ in a context where the members of act^* which are different of act_i are not acting.

From the conditions 2) and 3) it can be inferred that each act_i in act^* is a sufficient and necessary set of acts to bring it about that ϕ when the other members of act^* are not acting.

Theorem 5 *If act_i is in act^*, then $\models SJE^+_{act^*}\phi \to RJE^+_{act_i,(act^* \setminus act_i)}\phi$.*

Proof. This a trivial consequence of the truth conditions of the operator $SJE^{+}_{act^*}$.

This theorem validate the idea that each set of acts in act^* can cause ϕ independently of the other members of act^*.

Definition 8 *The following modal operators are defined from the operators $Done^{-}$, JE^{+} and SJE^{+}.*

$JE_{act}\phi \stackrel{def}{=} Done^{-}_{all}(JE^{+}_{act}\phi)$

If $IJE^{+}_{act}\phi$ is $JE^{+}_{act}(JE^{+}_{act_1}(JE^{+}_{act_2}\ldots(JE^{+}_{act_n}\phi)\ldots))$ we have:

$IJE_{act}\phi \stackrel{def}{=} Done^{-}_{all_n}(Done^{-}_{all_{n-1}}(\ldots(Done^{-}_{all_1}(JE^{+}_{act}\phi)\ldots)))$.

The intuitive meaning of these modal operators is:

$JE_{act}\phi$: the set of agents in *act* has brought it about that ϕ by doing exactly the set of acts *act*.

$IJE_{act}\phi$: the set of agents in *act* has indirectly brought it about that ϕ by doing exactly the set of acts *act*.

Theorem 6 *We have:*

(JT) $\models JE_{act}\phi \rightarrow \phi$

(IJT) $\models IJE_{act}\phi \rightarrow \phi$

Proof of (JT). From JE_{act} definition $JE_{act}\phi$ is $Done^{-}_{all}(JE^{+}_{act}\phi)$. We can easily show that and we have: $JE^{+}_{act}\phi \rightarrow Done^{+}_{all}\phi$. Then, we have $Done^{-}_{all}(Done^{+}_{all}\phi)$ which entails ϕ.

Proof of (IJT). The proof is by induction and it is based on the same idea as the proof of (JT).

3 Application to a case study

In this section the logical framework is applied to the example presented in the introduction. In order to show the expressive power of the logic it is assumed that there is another agent k who can add counter poison in the l's glass and that a given quantity of the counter poison inhibits the same quantity of poison. For instance, 5 grams of poisons and 3 grams of counter poison have the same effect as 2 grams of poison. The agents can do the following actions[4].

i has the possibility to put 6 grams (action A_6) or 5 grams (action A_5) of poison.

j has the possibility to put 6 grams (action B_6), or 5 grams (action B_5) of poison.

k has the possibility to put 7 grams (action C_7), or 9 grams (action C_9) or 2 grams (action C_2) of counter poison.

The formula p denotes the proposition: "*the quantity of poison and counter poison is equivalent to 4 or more grams of poison*".

[4] It is not assumed that these actions are the only actions that an agent can do and it may be that an agent does not do an action of type A nor an action of type B.

Case 1. Let's assume that in a world w the agents have selected the following action: $i : A_6$, $j : B_6$ and $k : C_7$. Notice that agent l is not acting.

We have $all = \{i : A_6, j : B_6, k : C_7\}$. The acts performance have started in the world w and for every w' such that $R_{all}(w, w')$ in w' agents i, j and k have respectively performed actions of the type A_6, B_6 and C_7. That is, in each w' the agents have performed variants of the **same** action types. For instance, the action type B_6 can be performed by putting 6 grams of poison in water, in whisky or in red wine. The condition 1) in the truth conditions of $JE^+_{\{i:A_6,j:B_6\}}p$ is true since in w' we have the equivalent of 6+6-7 = 5 grams of poison in the glass.

For the condition 2) we have to consider a counterfactual world w''_1 such that $R_{all-\{i:A_6\}}(w, w', w''_1)$. In w''_1 the acts that have been performed are the same as in w' except the act $i : A_6$ and the instance of the action types are also the same. If, for instance, in w' agent j has put 6 grams of poison in red wine, then in w''_1 he has also put 6 grams of poison in red wine. In w''_1 the only acts that have been performed are $j : B_6$ and $k : C_7$ and the quantity of poison in the glass is equivalent to 6-7 = -1 grams. Therefore, p is false in w''_1. We can easily show, in the same way, that in a world w''_2 such that $R_{all-\{j:B_6\}}(w, w', w''_2)$ p is also false. Then, we can conclude that the condition 2) in the truth condition is true.

Finally, for every world w''_3 such that $R_{all-\{k:C_7\}}(w, w', w''_3)$ the act $k : C_7$ is not performed and we have 6+6 = 12 grams of poison and p is true in w''_3. Then, condition 3) is also true.

Therefore, $M, w \models JE^+_{\{i:A_6,j:B_6\}}p$ is true. That fits the intuition since neither i nor j have caused alone that p is the case while their joint actions have caused p.

case 2. Let's assume now that in the world w agents have selected the following acts: $i : A_6$ and $k : C_7$, while agents j and l are not acting.

Then, in w' such that $R_{all}(w, w')$ and $all = \{i : A_6, k : C_7\}$ we have the equivalent of 6-7 = -1 gram of poison and we can easily check that $M, w \models JE^+_{\{i:A_6\}}p$ is false.

If in the world w' it is assumed that the only act which is performed is $j : B_6$, we have $all_1 = \{j : B_6\}$, and in the world w'_1 such that $R_{all_1}(w', w'_1)$ the quantity of poison in the glass is 6-1 = 5 grams (remember that in w' there is the equivalent of -1 gram of poison in the glass) and we can check that $M, w \models JE^+_{\{j:B_6\}}p$ is true.

These conclusions also fit the intuition that it is false that p is obtained by a joint act of i and j and that it is true that p has been caused by j's action even if j would not have the opportunity to cause that p if i did not perform the action A_6 before j starts to act.

Case 3. Let's assume now that in a world w the agents have selected the following action: $i : A_6$, $j : B_6$ and $k : C_2$ and agent l is not acting.

In that case we have: $all = \{i : A_6, j : B_6, k : C_2\}$. We can check that $M, w \models JE^+_{\{i:A_6\}}p$ is false though what agent i did is sufficient to make p true, since the equivalent of poison put by agents i and k, when j is not acting, is 6-2 = 4 grams. In formal terms, in the Definition 4 the truth condition 2) is not satisfied because for every worlds w' and w'' such that $R_{all}(w, w')$ and $R_{all-\{i:A_6\}}(w, w', w'')$, in w'' the equivalent quantity of poison is 6-2 = 4 grams because in w'' agent j is acting. The fact that $M, w \models JE^+_{\{i:A_6\}}p$ is false shows that we need a more appropriate definition of the causality than in the Definition 4 in cases where several agents can independently cause that p is true.

In the Definition 6 of the restricted joint operator RJE^+, the sufficient condition 2) is added to check whether what agent i does is sufficient to make p true when j is not acting and the condition 3) checks whether what agent i does is necessary to make p true when j is not acting.

Then, we can formally check that we have $M, w \models RJE^+_{\{i:A_6\},\{j:B_6\}}p$ and $M, w \models RJE^+_{\{j:B_6\},\{i:A_6\}}p$. Finally, we also have $M, w \models SJE^+_{\{i:A_6\},\{j:B_6\}}p$ which intuitively means that $i : A_6$ and $j : B_6$ can independently cause p.

4 Related works

Pauly has defined in [16] a Coalition Logic to represent groups of agents who are acting together (see also [1]). This logic can be seen as an extension of Harel's Dynamic Logic [10] to multiple agents in the sense that both of them ignore the counterfactual constraints. Other extensions are presented in [3]. The common feature of these logics is that they accept an axiom of the form $[G]\top$ which clearly shows that they do not represent causality since a group of agent cannot brings it about that a tautology is true.

In [12] (see also [13]) Horty has defined an action operator to represent the fact that a group of agents sees to it that ϕ is the case. This operator is usually abbreviated as a "stit" operator. We briefly introduce the semantics of its logical framework.

Frames are defined as sets of moments which are linearly ordered. This ordering defines histories, which have a tree structure. A moment m in an history h is denoted by m/h. The set of histories passing through the moment m is denoted by H_m.

In a model M a valuation function assigns a set of pairs m/h to each propositional atom of the language. The fact that a proposition represented by ϕ is true at m/h in the model M is denoted by: $M, m/h \models \phi$. A proposition which holds at the moment m of M and which is represented by ϕ is the set: $|\phi|^M_m = \{h \in H_m \ : \ M, m/h \models \phi\}$.

It is assumed that at a moment m of an history h an agent i has the possibility to do an action which is of a given type. Each action type is an element in a partition of H_m for i. This partition is formally represented by $Choice^m_i$ and the partition that represents the type of action that agent i does at m/h is represented by $Choice^m_i(h)$[5].

[5] In this approach it is assumed that $Choice^m_i(h)$ represents **all** the actions that agent i can do in m/h.

In this framework is defined the truth condition of a first action operator, usually called Chellas's stit [7, 6] and denoted by $[i \;\; cstit : \phi]$:

$M, m/h \models [i \;\; cstit : \phi]$ iff $Choice_i^m(h) \subseteq |\phi|_m^M$

The fact that agent i deliberatively sees to it that ϕ is represented by the formula $[i \;\; dstit : \phi]$ and its truth conditions are:

$M, m/h \models [i \;\; dstit : \phi]$ iff $Choice_i^m(h) \subseteq |\phi|_m^M$ and $|\phi|_m^M \neq H_m$.

The intuitive reading of the first condition is that doing an action of the type represented by the partition $Choice_i^m(h)$ is sufficient in m/h to guarantee the truth of ϕ and the reading of the second condition is that the proposition represented by ϕ may be false for an another history at the moment m. The latter condition is intended to represent the fact that what agent i did was necessary to obtain ϕ.

The cstit operator is extended to a group of agents G. Roughly speaking the set of possible choices of the agents in G is represented by the intersection of the choices of each agent in G and they are represented by $Choice_G^m$. Then, the truth conditions for this operator are defined by:

$M, m/h \models [G \;\; cstit : \phi]$ iff $Choice_G^m(h) \subseteq |\phi|_m^M$

However, there is no similar extension in [12] of the dstit operator to a group of agents.

In [5, 4] (see also [18]) Carmo has defined this extension for joint actions performed by a group of agents. The truth conditions for this operator are defined as follows:

$M, m/h \models [G \;\; dstit : \phi]$ iff

1) $M, m/h' \models \phi$ for every history h' such that $h' \in \cap_{i \in G} Choice_i^m(h)$, and

2) for every i in G there exists some h' such that h' pass through m (i.e. $h' \in H_m$) and $h' \in \cap_{j \in G - \{i\}} Choice_j^m(h)$ and $M, m/h' \not\models \phi$. [6]

The differences with the joint action operator JE presented in this paper can be shown with the case study we have analyzed in the previous section.

Let's first consider the case 1 where in m/h the agents have selected the following action types: $i : A_6$, $j : B_6$ and $k : C_7$.

Then, in m/h the quantity of poison in the glass is equivalent to 6+6-7 = 5 grams of poison and $M, m/h \models p$ is true. Intuitively, it is clear that p has been caused by the joint actions of i and j. However, we may have $M, m/h \models [\{i, j\} \;\; dstit : p]$ false.

Indeed, there may be an history h' such that $h' \in \cap_{x \in \{i,j\}} Choice_x^m(h)$ and in m/h' agent k has selected the action type C_9. In that situation the equivalent quantity of poison is 6+6-9 = 3 grams and p is false in m/h'. Hence, condition 1) is false and $M, m/h \models [\{i, j\} \;\; dstit : p]$ is false.

The basic reason why we have this counter intuitive consequence is that condition 1) is too strong. To guarantee that in m/h the actions selected by i and j in m/h are sufficient to obtain p we should only consider actions performed by the agents which are **all** of the same type as the actions they have performed in m/h.

---

[6] In [4] $M, m/h \models \phi$, $Choice_i^m(h)$ and m/h are respectively denoted by: $M \models_{h,m} \phi$, $C_{m,i}(h)$ and h, m.

Let's assume now that for every history h' such that $h' \in \cap_{x \in \{i,j\}} Choice_x^m(h)$ in m/h' agent k selects the action type C_7. Then, condition 1) is satisfied. Let's assume in addition that for every h' such that $h' \in \cap_{x \in G-\{i\}} Choice_x^m(h)$ or $h' \in \cap_{y \in G-\{j\}} Choice_y^m(h)$, in m/h' agent k has selected the action type C_5. In that situation i and j have respectively selected either A_6 and B_5 or A_5 and B_6 (remember that i (resp. j) can only chose action type A_6 or A_5 (resp. B_6 or B_5)). Then, condition 2) is false because for all these histories the equivalent quantity of poison is 6+5-5 = 6 grams and p is true in m/h'. Therefore, in that case also $M, m/h \models [\{i,j\} \ \ dstit : p]$ is false.

Again in this situation the formal consequence is counter intuitive. The basic reason is that the counterfactual histories in 2) should satisfy the *ceteris paribus* constraint, which is not imposed by condition 2). For instance, the fact that i is not doing A_6 should not entail that i is doing A_5.

In [15] Lorini and Schwarzentruber have used the $[G \ \ cstit : \phi]$ operator [7] defined by Horty to formalize different kinds of counterfactual emotions. For that purpose, they need to represent the fact that the group of agents G could have prevented a state of affairs represented by ϕ. This fact is denoted by $CHP_G\phi$ and it is formally defined as $CHP_G\phi \stackrel{\text{def}}{=} \phi \wedge \neg[AGT \setminus G \ \ cstit : \phi]$, where AGT denotes the set of all the agents. According to Horty's truth condition the fact that we have: $M, m/h \models \phi \wedge \neg[AGT \setminus G \ \ cstit : \phi]$ is equivalent to: $M, m/h \models \phi$ and there exists h' in $Choice_{AGT \setminus G}^m(h)$ such that $M, m/h' \models \neg\phi$. However, this latter property does not guarantee that in m/h' the group G sees to it that $\neg\phi$ in the sense of the dstit operator.

In [11] Hilpinen has defined a "*necessitating agency*" operator D (see AD9 in [11]) which has some common features with the operator JE though this operator is defined only for a unique agent. According to its truth conditions $D\phi$ is true in a world w'[8] iff there exists a world w and an action α such that i) $w' \in g(\alpha, w)$ and ϕ is true in w' and ii) there exists a world w'' such that $< w, w'' >$ is "*maximally similar to the course of action exemplified by* $< w, w' >$" and ϕ is false in w''. In the condition i) $g(\alpha, w)$ denotes the set of worlds where we are after performance of an action of the type α and possibly other actions. There is a deep similarity between a tuple of worlds $< w, w', w'' >$ that satisfies these conditions and a tuple that satisfies $R_{all-\{i:\alpha\}}(w, w', w'')$ for some $i : \alpha$ in the framework we have presented here. This is a basic difference with the *stit* operator. Indeed, if it is assumed, for example, that in m/h agent i does both actions A_4 and A_2 (respectively: to put 4 grams and 2 grams of poison) instead of doing A_6, we can define a partition of H_m for i in such a way that in some partition i only does A_4, in another one he only does A_2 and in another one he does nothing. Nevertheless, there is no means in the *stit* formalism to express which one satisfies the *ceteris paribus* constraint with respect to m/h. That is the reason why we have proposed, like Hilpinen in [11], these kinds of ternary relationships.

---

[7] The notations have been changed in order to make easier the comparison with previous works.

[8] The notations have been changed to make easier the comparison with other works.

In [20] Sergot has proposed formal definitions for joint actions operator in contexts where agents are acting collectively. These definitions take inspiration in the definition of Pörn's bringing it about operator. However, a significant difference is that these operators are intended to characterize *how* the agents do the actions and not about the final state of affairs which is obtained.

5 Conclusion

It has been shown from case studies that it is not easy to find a general definition of causality that fits the intuition when it is applied to different kinds of situations.

We have proposed a definition of the operator of joint actions JE which is appropriate to represent situations where the set of acts performed by a group of agents is sufficient to obtain a situation where a proposition ϕ holds and where all the acts in this set are necessary to obtain this situation.

Another operator of indirect joint actions IJE has been defined for situations where a set of agents has indirectly caused a state of affairs by bringing it about that other agents have caused that situation.

The analysis of the case study 3 shows that the definition of the joint action operator JE has to be modified in the case where several sets of agents have caused independently that ϕ holds in the sense that each set of agents if it would have been acting alone would have caused ϕ. To represent causality in these kind of situations we have defined the restricted joint operator RJE.

In addition to the formal definitions of these operators we have formally proved some logical properties of the operators and it has been shown that the application of these definitions to the case study leads to non counter intuitive conclusions. Finally, we have presented a detailed analysis of other similar proposals in particular those which refer to the stit operator.

The presented case study is an academic example. Nevertheless, we could easily find similar cases in the field of computer science. For instance, in a context where it is forbidden to communicate a password it may happen that two agents have jointly informed another agent about a password in the case where one of them has informed about the beginning of the password while the other one has informed about the rest of the password. Another example may be that it is forbidden to remove all the copies of a given file. If there are only two copies which are in two different sites and two agents start simultaneously a command to remove each one a copy, they have removed all the copies by their joint actions and this state of affairs has been obtained indirectly since each command has caused the performance of one or several software agents.

The work which has been presented could be extended into several directions. One of them is to find a complete axiomatization of what has been defined in the semantics. Another one is to extend the logical framework to the representation of what the agents believe or know. For instance, for the analysis of responsibilities it may be relevant, in the case 2, where j add poison after i, to distinguish cases where j knows what i did and cases where he does not know that.

Acknowledgements. We would like to thank José Carmo, Andreas Herzig and the reviewers of EUMAS conference for all their fruitful comments. If there are mistakes they are the only responsibility of the author.

References

1. T. Agotnes, W. van der Hoek, and M. Wooldridge. On the logic of coalitional games. In *Proceedings of the Conference on Autonomous Agents and Multi Agent Systems*. Association for Computing Machinery, 2006.
2. L. Aqvist. Old foundations for the logic of agency and action. *Studia Logica*, 72, 2002.
3. J. Broeersen, A. Herzig, and N. Troquard. Normal Coalition Logic and its conformant extension. In D. Samet, editor, *Theoretical Aspects of Rationality and Knowledge*. Presses Universitaires de Louvain, 2007.
4. J. Carmo. Collective agency, direct action and dynamic operators. *Logic Journal of the IGPL*, 18(1):66–98, 2010.
5. J. Carmo and O. Pacheco. Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles. *Fundamenta Informaticae*, 48:129–163, 2001.
6. B. Chellas. On bringing it about. *Journal of Philosophical Logic*, 24, 1995.
7. B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
8. R. Demolombe. Relationships between obligations and actions in the context of institutional agents, human agents or software agents. *Journal of Artificial Intelligence and Law*, 19(2), 2011.
9. R. Demolombe and A.J. Jones. Actions and normative positions. A modal-logical approach. In D. Jacquette, editor, *Companion to Philosophical Logic*. Blackwell, 2002.
10. D. Harel. Dynamic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2. Reidel, 1984.
11. R. Hilpinen. On Action and Agency. In E. Ejerhed and S. Lindstrom, editors, *Logic, Action and Cognition: Essays in Philosophical Logic*. Kluwer, 1997.
12. J. Horty. *Agency and deontic logic*. Oxford University Press, 2001.
13. J.F. Horty and N. Belnap. The deliberative STIT: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.
14. L. Lindahl. *Position and Change- A Study in Law and Logic*. Synthese Library 112, D. Reidel, 1977.
15. E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175:814–847, 2011.
16. M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
17. I. Pörn. Action Theory and Social Science. Some Formal Models. *Synthese Library*, 120, 1977.
18. F. Santos and J. Carmo. Indirect Action, Influence and Responsibility. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems, Workshops in Computing Series*. Springer, 1996.
19. K. Segerberg. Outline of a logic of action. In F. Wolter, H. Wansing, W. de Rijke, and M. Zakharyaschev, editors, *Advances in Modal Logic, Volume 3*. World Scientific Publishing Co., 2002.

20. M. Sergot. The logic of unwitting collective agency. Technical report 2088/6, Imperial College, London, 2008.
21. G. H. von Wright. *Norm and Action*. Routledge and Kegan, 1963.