Responsibility Judgments: Steps towards a formalization

Henri Prade
IRIT, CNRS - University of Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9, France
prade@irit.fr

Abstract

The concept of responsibility refers to different, although related, meanings. Judgments about responsibility, demerit or merit may involve several facets, such as causality, intentionality, awareness of consequences, as well as utility and deontic notions. The proposed approach relies on a recent modeling of causality ascription by agents put in face of a sequence of reported events, on the basis of their own beliefs regarding the normal course of things. The study intends to cover situations, where something bad, or good, happened, and the action of an agent caused it, or could have prevented it. Moreover, the expected consequences of the action may have taken place or not, the action may be costly or beneficial to the agent, and the action may have different deontic status. On the basis of these different concerns, we are interested in this paper in describing how responsibility, merit or blame can be attributed by agents about the actions of other agents.

Keywords: Responsibility, Causality, Merit, Blame, Nonmonotonic Consequence Relation.

1 Introduction

The formalization of the concept of responsibility has become an issue in artificial intelligence with the increasing interest for the modeling of multiple agent systems; see in particular [3], [4], [9], [12]. Indeed, it may be important to model how agents perceive the actions of other agents, and may attribute responsibility, merit or blame to these agents for their actions. Moreover, the concept of responsibility plays a key role in norm-governed organizations and legal reasoning. The violation of norms can be itself related to causality [7].

Responsibility is usually tightly connected with the idea of causality. Indeed it is hard to think of being responsible of something in some sense, without having any relation with what has caused this thing to happen. Different models of causality have been proposed in the last decade in artificial intelligence [11], [6], [2]. The emphases of these three approaches are quite different: The first one is more concerned with the introduction of a logical language for describing temporal trees of events in terms of five basic relations between events, while the second one proposes a modeling of causality in systems described by structural equations, taking advantage of the idea of intervention [10]. The last proposal uses a non-monotonic logic representation of what is the normal course of things for an agent, leading to view the potential cause(s) of a reported change as an abnormal (conjunction of) event(s) in a given context, in agreement with cognitive psychology findings. In the following, we privilege this latter view, which is in any case compatible with the two others.

The paper is organized as follows. In the next section a general discussion points out the different meanings of the concept of responsibility, and the focus of the paper, which is more about causal responsibility than moral responsibility. Section 3 identifies the different

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 145–152
Torremolinos (Málaga), June 22–27, 2008

facets that are involved in responsibility, merit, or blame judgments. Section 4 provides the necessary background on causality ascription, and extends the existing model for taking into account new features of the problem, while Section 5 presents the proposed formalization of responsibility, and offers a preliminary and informal discussion on what matters in merit or blame attribution, before concluding.

2 Responsibility Judgments: A General Discussion

Cholvy, Cuppens and Saurel [4] have distinguished three different meanings for the idea of responsibility. They summarize the first meaning in the following way.

Definition 1. Something bad happened, and you caused it or could have prevented it.

This view may be termed « causal involvement », although the above authors did not give any particular name to it. Unsurprisingly, this definition directly relates the idea of responsibility to the notion of cause. It is worth noticing that the authors are only interested in the case where "something bad happened", since they focused on damages (in the sense of a given norm) caused, being mainly concerned by failures of security requirements. However, even if society is more prompt to blame people for bad things that occurred due to their responsibility, than to recognize merits, a formal approach should be general enough to encompass the case where good things took place as well.

Besides, an interesting feature of Definition 1 is that it includes situations where an agent could have prevented bad things to happen. This implicitly means that both the action(s) performed by the agent and those that were not performed (although they were feasible) have to be considered. The lack of prevention of bad things happening is referred as "indirect responsibility" in [4]. Note also that an agent who prevented a good (resp. bad) thing to happen may be blamed (resp. complimented), but this is not covered by Definition 1.

Definition1 may be exemplified by

Example 1.a “The child throw a stone with force into the window, the pane is broken”.

Example 1.b “Peter left his house lately, and missed the train” (which suggests that if he had left early, he would have got the train).

The second understanding of responsibility considered in [4], termed « answerability », is stated and illustrated as follows:

Definition 2. Obligation or moral duty to report or explain your actions or someone else’s action to a given authority.

Example 2. “Parents are considered as being legally responsible for the damages caused by their minor children”.

The last sense of the term "responsibility" discussed in [4], called « accountability » , is in some sense a strengthening of Definition 2, where the sanctions that apply when something bad occurs are properly defined in the policy regulations. This corresponds to

Definition 3. Position, which enables you to make decisions in a given organization but implies that you must be prepared to justify your actions.

Example 3. “The President is responsible in front of the Assembly” (and if the President seriously failed, he may be dismissed).

It is also useful to have in mind the following classical distinctions between i) civil responsibility (that refers to the obligation of repairing the damaged caused to somebody) and ii) penal responsibility (for those who can be prosecuted for their crimes or offences) from a juridical point of view, and iii) moral responsibility from an ethical point of view (the agent recognizes himself as the author of his acts, and assumes their merits or demerits).

In the following, we concentrate on the view expressed by definition 1 (extended to the happening of good things), mainly in a causal responsibility perspective.

3 The Different Facets of Responsibility

As implicitly shown by the previous discussion, there are several features that may have an impact on the attribution of responsibility to an agent who had a role in some good or bad reported event. Thus, the following issues appear to be relevant:

a) What happened? What did not happen, but could have happened? Could what happened be expected or not? Indeed, if something really unexpected took place after an action, the author of the action can hardly be considered as responsible for the results, be there good or bad. For instance, an agent who intended to fire another person, but his action fails because his gun was jammed cannot be usually prosecuted for that (but only for the threat committed), nor credited for having spared his potential victim!

b) Clearly, using Definition 1 requires being able to give a precise meaning to the verbs "to cause", and "to prevent".

c) An agent (or a group of agents) is identified as being the author of an action, or as not having performed some action. In the following, we do not discuss the important problem of the responsibility shared between several agents and then how to allocate parts of the responsibility to each agent; see [3] for a preliminary discussion and proposal.

d) Was the performed action intended? Or its non performance intended? Was the will of the agent free? Or, on the contrary, was the agent under the command, or the pressure of another (group of) agents? In the following, it is assumed for simplicity that the concerned agent had his free will when he performed the considered action (or did nothing).

e) What happened, or what could have happened, is bad, or good according to some norm, or accepted scale.

f) The action performed (or not performed!) may be obligatory, recommended, or on the contrary not recommended at all, or even forbidden. It may be also completely free. In this latter case, although there is no violation of any regulation, the consequences of the action may still be good or bad for other agents, who will regard the author of the action as responsible for their benefits or their troubles.

g) The action performed (or not performed!) may be more or less costly, or on the contrary somewhat rewarding by itself for the agent who performed it. Indeed, for instance, an agent who made a good thing for others may be considered as having all the less merit, as this thing is also more beneficial for him. In case what is forbidden is only a matter of penalty, performing a forbidden action may be viewed as being only a matter of cost for his author, the action being judged on the goodness, or the badness of its consequences for the other agents.

The above seven points summarize the different main issues that have to be taken into account when judging of the merit, demerit, and responsibility of an agent.

4 Causality Ascription

In this section, we first address the points a and b introduced in the previous section. Moreover we assume a unique acting agent, being completely free in his will (points c and d).

4.1 What happened

It is assumed that one is in a (reported) context C (C represents the partial available knowledge about the context). Moreover, it is supposed that a sequence such as,

$$\neg B_t \quad A_t \quad B_{t'}$$

where t' denotes a time instant strictly after t (B_t means that B is reported true at time t). This reads: B was false, A took place, then B became true.

A_t denotes an action that took place at time t. Although A might be an event that it is not under the control of an agent, as in the example “the storm broke out (A_t), then the flood took place ($B_{t'}$)”, we assume that A is an action performed by an agent, as in the report “Peter drank (A_t), he got a fine ($B_{t'}$)”. Note also that the report may be incomplete. For instance, in the above example a more complete report may be “Peter drank (alcoholic beverages) (A^1_t), he drove his car ($A^2_{t'}$), he became inebriated ($B^1_{t''}$), he was controlled by a policeman ($B^2_{t'''}$), he got a fine ($B^3_{t''''}$)”, with t’’’ > t’’ > t’ > t.

Besides, the agent *a*, who performs A, as well as the person that receives the sequential information $\neg B_t$, A_t, $B_{t'}$, and who is supposed to judge the potential responsibility of agent *a*, have some knowledge on what is the normal course of the world in context C, and maybe also in context C ∧ A, regarding B. For simplicity, it is assumed that agent *a* and the judge have exactly the same knowledge.

Namely, one may either believe that $C \mid\approx B$ (B is expected to be true in context C), or that $C \mid\approx \neg B$

(B is expected to be false), or that $C \not|\!\approx B$ and $C \not|\!\approx \neg B$ (the truth or the falsity of B is contingent), where $|\!\approx$ is a so-called nonmonotonic consequence relation [8] describing what is normal, and $\not|\!\approx$ stands for its negation. Similarly, in context $C \wedge A$, the agent may have the same form of belief. It is assumed, that $C \wedge A$ is consistent (otherwise, one would have to take into account that the fact that A becomes true should modify C into a known way C').

In case, one knows $C \wedge A |\!\approx B$, B is expected to be true after A took place, and the sequence $\neg B_t$, A_t, $B_{t'}$ is not surprising. On the contrary, if the sequence $\neg B_t$, A_t, $\neg B_{t'}$ is reported, it would mean that action A had not its normal, expected effect.

A similar analysis may be conducted with respect to what did not happen (rather than w. r. t. what happened). Indeed, let us assume the sequence $\neg B_t$, $\neg H_t$, $\neg B_{t'}$, with t' > t, where $\neg H_t$ means that the hypothetical action H has not been performed at time t. Thus, B was false, H did not take place, B remains false. This is expected if one knows $C |\!\approx \neg B$ and $C \wedge \neg H |\!\approx \neg B$. Moreover, in case $C \wedge H |\!\approx B$, one is allowed to think that if H had taken place, it is likely that B would have become true, leading to a reported sequence $\neg B_t$, H_t, $B_{t'}$.

4.2 Causation and related notions

Bonnefon, Da Silva Neves, Dubois and Prade [2] have recently justified, both theoretically and experimentally, the two following definitions of facilitation and causality ascriptions made on the basis of pieces of default knowledge when a sequence where a change took place is reported.

Definitions 4 (Facilitation and Causation). Let us assume that an agent learns of the sequence $\neg B_t$, A_t, $B_{t'}$. Let us call C (the context) the conjunction of all other facts known by, or reported to the agent at time t' > t. Given a nonmonotonic consequence relation $|\!\approx$, if the agent believes that $C |\!\approx \neg B$, and that $C \wedge A \not|\!\approx \neg B$ (resp. $C \wedge A |\!\approx B$), the agent will perceive A as having *facilitated* the occurrence of (resp. as *being the cause of*) B in context C, which will be denoted $C: A \Rightarrow fa\ B$ (resp. $C : A \Rightarrow ca\ B$).

Indeed, [2] presents experiments that indicate the cognitive validity of the two above notions. Moreover, these definitions have expected properties [2]. In particular, it is shown that

- If $C: A \Rightarrow ca\ B$ or if $C: A \Rightarrow fa\ B$ then $C |\!\approx \neg A$.
- A restricted transitivity property holds: If $C: A \Rightarrow ca\ B$, if $C: B \Rightarrow ca\ D$ and if $B \wedge C |\!\approx A$ then $C: A \Rightarrow ca\ D$. The two properties hold for $\Rightarrow ca$ provided that $|\!\approx$ is a preferential entailment in the sense of Kraus, Lehmann and Magidor [8]. The first property holds for facilitation ($\Rightarrow fa$) if $|\!\approx$ is a rational closure entailment.

Note that here transitivity requires $B \wedge C |\!\approx A$, i.e. A is not too specific with respect to B (it means that the normal way to have B (in context C), is to have A). For instance, for A = drinking, B = inebriated, D: staggering, we have 'drinking' ⇒ca 'inebriated' and 'inebriated' ⇒ca staggering' entail 'drinking' ⇒ca 'staggering', since 'inebriated' $|\!\approx$ 'drinking'.

Example. Again consider the example "Peter drank ($A^1{}_t$), he drove his car ($A^2{}_{t'}$), he became inebriated ($B^1{}_{t''}$), he was controlled by a policeman ($B^2{}_{t'''}$), he got a fine ($B^3{}_{t'''}$)". From the commonsense knowledge

$$C |\!\approx \neg B^1,\ C \wedge A^1 |\!\approx B^1$$
$$C \wedge A^2 |\!\approx \neg B^3,\ C \wedge A^2 \wedge B^2 \not|\!\approx \neg B^3$$
$$C \wedge B^1 \wedge A^2 \wedge B^2 |\!\approx B^3$$

one can conclude that

- "the fact Peter drank caused that he became inebriated" ($C |\!\approx \neg B^1$, $C \wedge A^1 |\!\approx B^1$),
- "the fact Peter became inebriated and was controlled caused that he got a fine"
($C \wedge A^2 |\!\approx \neg B^3$, $C \wedge A^2 \wedge\_\ B^1 \wedge B^2 |\!\approx B^3$),
- "the fact Peter was controlled facilitates that he got a fine" ($C \wedge A^2 |\!\approx \neg B^3$, $C \wedge A^2 \wedge B^2 \not|\!\approx \neg B^3$),

but one cannot conclude, which is quite reasonable, that
"the fact Peter drove caused that he got a fine"
(because $C \wedge A^2 |\!\approx \neg B^3$).

Still one could argue that if Peter had not taken his car, he would not have got any fine. This illustrates the fact that the idea of counterfactual is not sufficient to reveal causes. In fact, one could say that the fact that Peter drove his car is here something like a necessary condition for having him getting a fine.

An attempt is now made in order to formally define the idea of being a necessary condition for something to take place.

Definition 5 (Necessary condition). N is said to be a necessary condition for having B in the scenario $\neg B_t$, A_t, $B_{t'}$ (with t' > t), if

i) $C |\!\approx N$, ii) $C |\!\approx \neg B$, iii) $C \wedge A |\!\approx B$,

and

iv) $C \wedge \neg N |\!\approx \neg B$; $C \wedge \neg N \wedge A |\!\approx \neg B$.

Note that here the necessary condition N is normal in the context, while a potential cause should be abnormal as said before. Note also that N does not cause, nor facilitates B in context C. Indeed, from (i) and (ii), one can conclude from a characteristic property[1] of preferential entailment [8] that $C \wedge N \mid\!\approx \neg B$, as well as $C \wedge N \wedge A \mid\!\approx B$.

This can be illustrated by the following example (inspired from [5]) where C = 'paper' ; N = 'oxygen'; A = 'matches'; B = 'fire', where using matches will be perceived as the cause of fire, while having oxygen is a necessary condition.

Let us go back to our drunk driver example. The following common sense knowledge seems reasonable:

$C \mid\!\approx \neg B^3$ (in general one does not get a fine)

$C \wedge \neg A^2 \mid\!\approx \neg B^3$ (in general if one does not drive one does not get a fine)

$C \wedge \neg A^2 \wedge B^1 \wedge B^2 \mid\!\approx \neg B^3$ (in general if one does not drive one does not get a fine even if one is inebriated and controlled)

We observe that it corresponds to conditions (ii) and (iv) in Definition 5. Conditions (i) and (iii), which would write $C \mid\!\approx A^2$ and $C \wedge B^1 \wedge B^2 \mid\!\approx B^3$ respectively, do not hold here. Still a weaker condition holds, namely

$C \wedge A^2 \wedge B^1 \wedge B^2 \mid\!\approx B^3$ (in general if one is controlled while driving and being inebriated, one gets a fine)

It would correspond to weaken Definition 5 by dropping requirement (i) and replacing (iii) by a more specific condition involving N, namely $C \wedge N \wedge A \mid\!\approx B$.

So, strictly speaking, using definitions 4 ad 5, the fact that Peter drove, is neither a cause for getting a fine, nor a necessary condition, although it does play a role in the process (and it satisfies some of the key conditions of Definition 5).

Remark. The above definition of causality is appropriate in a "static" world that does evolve by itself (i.e. things tend to persist), i.e. in the situation considered in Definition 4, $\neg B_t$ persists to be true if there is no action that takes place (because in general $C \mid\!\approx \neg B$). In a "dynamic" world, $\neg B_t$ would tend to change "spontaneously", i.e. $\exists\, t^*, \forall t^\circ > t^*, C \mid\!\approx B_{t^\circ}$. A simple example is provided by the case of a serious disease that leads to death if nothing is done. In such a case, Definition 4 should be adapted in the following way for causality (facilitation can be handled similarly).

Definition 6. Let us assume that an agent learns of the sequence $\neg B_t, A_t, \neg B_{t'}$. Let us call C (the context) the conjunction of all other facts known by the agent at time t' > t. Given a non-monotonic consequence relation $\mid\!\approx$, if the agent believes that for some $t^*, \forall t^\circ > t^*, C \wedge \neg A \mid\!\approx B_{t^\circ}$, and that $C \wedge A \mid\!\approx \neg B$, the agent will perceive A as *having caused* ¬B in context C.

4.3 Prevention

In the previous informal discussion of the concept of responsibility in sections 2 and 3, it was mentioned that the fact that an agent could have prevented that something (especially a bad thing) to happen, by doing (or not doing) some action, may lead to some indirect responsibility of the agent. Let us try to clarify what "prevents" could mean here.

First, remember that stating that "A caused B" in the reported sequence $\neg B_t, A_t, B_{t'}$ with t' > t, amounts to believe $C \mid\!\approx \neg B$, and $C \wedge A \mid\!\approx B$. A first understanding of "prevent" is given by the definition:

Definition 7. (Prevention to persist)
A prevents B if A causes ¬B.

Indeed, if A causes B in a sequence $\neg B_t, A_t, B_{t'}$ (t' > t), A prevents ¬B to persist. This view is close to the idea of an action H that annuls the effects of action A, a soon as

$$C \mid\!\approx \neg B,\ \ C \wedge A \mid\!\approx B,\ \ C \wedge A \wedge H \mid\!\approx \neg B$$

Indeed, H prevents B to persist in such a case.

Another slightly different understanding of 'prevent' is "A prevents B to take place". It corresponds to the definition

Definition 8. (Prevention to take place)
A prevented B to take place in the reported sequence $\neg B_t, A_t, \neg B_{t'}$ (t' > t) if

i) $C \mid\!\not\approx \neg B$ (i.e. ¬B does not persist by itself)
ii) $C \wedge A \mid\!\approx \neg B$.

[1] The property used here is called "cautious monotony": from $C \mid\!\approx A$ and $C \mid\!\approx B$, one can deduce $C \wedge A \mid\!\approx B$.

In such a case, having $\neg B$ initially was not particularly expected, i.e. was not normal, and once A took place, having $\neg B$ was normal. Note that the condition (i) covers two situations: either $C \mid\!\approx B$ (and $\neg B_t$ is exceptional), or $C \not\mid\!\approx B$ (and $\neg B_t$ is contingent). Doing A prevents to have B becoming true by the normal course of things in the first case, and by accident in the second case (up to the potential failure of A w. r. t. its expected consequence).

Starting with a situation where B is false, and given an action, there are four scenarios that can be considered, according as H is performed or not, and B becomes true or not. Let us examine them, and see what prevents what.

1) H was performed, B became true, i.e. the sequence $\neg B_t, H_t, B_{t'}$ $(t' > t)$ is reported. In such a case if the normal course of things is known as being described by $C \mid\!\approx \neg B$ and $C \wedge H \mid\!\approx B$, H is perceived as the *cause* of B. It entails that $C \wedge \neg H \mid\!\approx \neg B$, i.e. "if H is not performed, B does not take place (normally)". In other words, doing H prevents $\neg B$ to persist.

2) H was not performed, B did not become true, i.e. the sequence $\neg B_t, \neg H_t, \neg B_{t'}$ $(t' > t)$ is reported.

If the normal course of things is described by $C \mid\!\approx \neg B$, then $\neg B$ has just been persisting. Moreover, one may assume that $\neg H$ is innocuous with respect to $\neg B$, i.e. $C \wedge \neg H \mid\!\approx \neg B$ (maybe because $C \mid\!\approx \neg H$; in any case it should at least hold that $C \not\mid\!\approx H$, i.e. there is no special reason to have H performed). Besides, note that, in contrast with the previous case, $C \mid\!\approx \neg B$ and $C \wedge \neg H \mid\!\approx \neg B$ does entail at all that $C \wedge H \mid\!\approx B$ should hold. However, if it is known that $C \wedge H \mid\!\approx B$ does hold, then one can argue that "if H have been performed, B would have taken place", and then H would have been regarded as the cause of B (since $C \mid\!\approx \neg B$ and $C \wedge H \mid\!\approx B$). Then, an agent could have *prevented* $\neg B$ to *persist* by doing H.

Otherwise, $\neg B$ does not tend to persist by itself, i.e. $C \not\mid\!\approx \neg B$ and the report of $\neg B_{t'}$ is not especially expected. Then, if moreover one knows $C \wedge \neg H \mid\!\approx \neg B$, one could say that not doing H *prevents* B to *take place*.

3) H was not performed, B became true, i.e. the sequence $\neg B_t, \neg H_t, B_{t'}$ $(t' > t)$ is reported. Such a sequence is consistent with not having B in context C normally ($C \not\mid\!\approx \neg B$). In such a case, if it is known that $C \wedge H \mid\!\approx \neg B$, i.e. doing H in context C normally lead to have $\neg B$, one can argue that "if H have been performed, B would not have taken place", as in the example "if Peter had inserted his foot in the door ajar, the door would not have shut off". In such a situation, not doing H *has prevented* $\neg B$ to *persist*, but it does not mean that $\neg H$ caused B (since $C \wedge \neg H \mid\!\approx B$ is not known).

Note that the reported sequence $\neg B_t, \neg H_t, B_{t'}$ might be due to some unreported action A_t. Then if $C \wedge A \wedge H \mid\!\approx \neg B$, doing H would have annulled the effect of A, and prevented $\neg B$ to persist.

4) H was performed, B did not become true, i.e. the sequence $\neg B_t, H_t, \neg B_{t'}$ $(t' > t)$ is reported. If it is the case that $C \not\mid\!\approx \neg B$, and moreover $C \wedge H \mid\!\approx \neg B$, then one can argue that "if H have not been performed, B might have taken place", i.e. doing H *prevented* B *to take place* (in the sense of Definition 8). For instance, "If Peter had not inserted his foot in the door ajar, the door would have shut off".

Again, if there is some unreported action A_t in the sequence $\neg B_t, H_t, \neg B_{t'}$ and if $C \wedge A \wedge H \mid\!\approx \neg B$ (while A_t caused B, if $C \mid\!\approx \neg B$ and $C \wedge A \mid\!\approx B$), not doing H would have prevented $\neg B$ to take place, by not annulling the effect of A.

5 Ascribing Responsibility, Merit and Blame

In the following, an agent is responsible inasmuch the agent caused, or could have prevented, by performing or not an action, that something happened. The notions of merit and blame are primarily connected with the goodness or badness of what happened. They also involve the deontic status of the considered action, and its positive or negative benefit for the agent.

5.1 Direct and Indirect Responsibility

In the view developed here, the idea of responsibility is disconnected from the idea that something *bad* happens, but is only related to the fact that something happens, be it good or bad. This leads to the following definition.

Definition 9. An agent *a* is perceived as *directly responsible* for the happening of B in the reported sequence $\neg B_t, A_t, B_{t'}$ $(t' > t)$, if

- *a* performs A free from the coercion of any other agent;
- A caused B (in the sense of Definition 4).

Note that this definition makes sure that the expected effect of action A has been reported. Besides, one may think of a weaker form of responsibility in case 'A caused B' is changed into 'A facilitated B'. Note also that here an agent may be only responsible for something done that is not in the normal course of things in the current context (as a consequence of Definition 4).

The above definition applies to a "static" world, and should be adapted for a "dynamic" world (see final remark in section 4.2), as follows

Definition 10. An agent *a* is perceived as *directly responsible* for the happening of $\neg B$ in the reported sequence $\neg B_t$, A_t, $\neg B_{t'}$ ($t' > t$), if
- *a* performs A free from the coercion of any other agent;
- A caused $\neg B$ (in the sense of Definition 6).

It can be illustrated by the following example: The door was open, but going to slam, and the agent maintains it open by inserting his foot.

The idea of indirect responsibility corresponds to the situation where something happened, and the agent could have prevented it.

Definition 11. An agent *a*, free from the coercion of any other agent, is perceived as *indirectly responsible* that B took place in context C, if

- in case of a reported sequence $\neg B_t$, $\neg H_t$, $B_{t'}$ ($t' > t$), *a* could have prevented B to take place by performing H, provided that $C \wedge H \mid\!\approx \neg B$;

- in case of a reported sequence $\neg B_t$, H_t, $B_{t'}$ ($t' > t$), *a* could have prevented B to take place by not performing H if $C \wedge \neg H \mid\!\approx \neg B$, or by performing an act A that annuls the effect of H, i.e. such that $C \wedge A \wedge H \mid\!\approx \neg B$.

A similar definition can be adapted to the case where $\neg B$ persisted, namely

Definition 12. An agent *a*, free from the coercion of any other agent, is perceived as *indirectly responsible* that $\neg B$ persisted in context C, if

- in case of a reported sequence $\neg B_t$, $\neg H_t$, $\neg B_{t'}$ ($t' > t$), *a* could have prevented $\neg B$ to persist by performing H, provided that $C \wedge H \mid\!\approx B$;

- in case of a reported sequence $\neg B_t$, H_t, $\neg B_{t'}$ ($t' > t$), *a* could have prevented B to persist by not performing H if $C \wedge \neg H \mid\!\approx B$, or by performing an act A that annuls the effect of H, i.e. such that $C \wedge A \wedge H \mid\!\approx B$.

5.2 Meritoriousness and Blameworthiness

Getting a result by an action that delivered its expected consequences, of which the agent who performed the action was aware of, is all the more meritorious for the agent as the result is better, and all the more blameworthy as it is worse.

Indeed, an agent cannot be blamed (or congratulated) for an action that he performed, but which fails. For instance, if John threw a stone into a windowpane, but did not break it, he is responsible of nothing, and moreover one could check that throwing the stone is not the cause that the pane remained intact.

Note also that there are things that happen, due to the responsibility of agents, which are neither especially good, nor especially bad, but just neutral. In such situations, the agents are responsible, but do not deserve any blame or compliments.

Moreover, the evaluation of results should not be only based on what is obtained, but also on what was avoided. Indeed $B_{t'}$ may be judged as being just "not bad", while $\neg B_t$ was really undesirable, for instance. Similar cases can be encountered with "good" things. Thus, an agent who is responsible that something bad did not happen, because he prevented it, is as meritorious as an agent who is responsible that something good happened. Similarly, an agent responsible for something good that did not happen, because he prevented it, is as blameworthy as an agent responsible for something bad that happened.

Besides, blame and merit have to be modulated according to the deontic status of the considered action. Doing something obligatory inhibits merit or blame, if one considers that the agent did nothing but his duty. On the contrary, doing something forbidden that led to a bad result should reinforce the blame.

Lastly, blame and merit seem also to have to be modulated by taking into account if what the agent did or did not was profitable, free, or costly for him. Clearly, if for instance, something bad happened, because the agent did not prevent it, but if what he could have done was very costly for him, this provide him with some excuse. On the contrary, if something good happened, but the agent knew that his action would be very beneficial also for him, the merit of the agent is more debatable.

6 Concluding Remarks

The paper has presented a preliminary discussion of how a recently proposed model of causality (based on nonmonotonic consequence relations) that focuses on the abnormal features that in a context cause changes, can be used for assessing agents' responsibility. This has led to discuss new aspects of the proposed modeling of causality, and to a new view in modeling responsibility since it is based on a model of causality that departs from the previously used ones for this purpose.

Clearly many questions remain open, in particular a more formal approach for determining when agents are meritorious or are blameworthy. This would also include a careful comparison with the other approaches to responsibility, even if they are generally based on quite different intuitions. For instance, [4] assumes that responsibility is associated with the happening of bad consequences, themselves due to the violations of regulations that state what is obligatory or forbidden. However, regulations are just guidelines on the way to act, and are something that is not as primitive as causality in the attribution of responsibility, while they become certainly more crucial in the assessment of blames. Lastly, these issues are by nature, and because of their different facets, highly arguable, and could be also addressed using an argumentation-based approach [1].

Acknowledgements

This work has been supported by a grant from the French "Agence Nationale pour la Recherche (ANR)", project number NT05-3-44479 (project 'MICRAC' http://www.irit.fr/MICRAC).

References

1. L. Amgoud, H. Prade. Arguing About Potential Causal Relations. *Journées "Intelligence Artificielle Fondamentale"* (IAF'07), Grenoble, Jul. 2-3, 2007. http://gdri3iaf.info.univ-angers.fr/spip.php?rubrique7

2. J.-F. Bonnefon, R. M. Da Silva Neves, D. Dubois and H. Prade. Background Default Knowledge and Causality Ascriptions. *Proc. of the 17th European Conference on Artificial Intelligence* (ECAI'06), G. Brewka, S. Coradeschi, A. Perini, P. Traverso (Eds.), IOS Press, Zurich, pages 11–15, Riva del Garda, Italy, Aug. 29 – Sept.1, 2006. Revised and extended version to appear in the *Int. J. of Approximate Reasoning.*

3. H. Chockler, and J. Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22, 93-115, 2004.

4. L. Cholvy, F. Cuppens, and C. Saurel. Towards a Logical Formalization of Responsibility. *Proceedings of the 6th International Conference on Artificial Intelligence and Law*, Melbourne, Australia, June 1997, 233-242.

5. N. Goodman, *Fact, Fiction and Forecast.* The Bobbs-Merrill Comp., 3rd edition, 1973

6. J. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. Part II: Explanations. *British Journal for the Philosophy of Science*, 56, 843-887 & 889-911, 2005.

7. D. Kayser and F. Nouioua. About Norms and Causes. *International Journal of Artificial Intelligence Tools*, 14, 7-24, 2005.

8. S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44, 167-207, 1990.

9. W. Mao, and J. Gratch. Social Causality and Responsibility: Modeling and Evaluation. *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents* (IVA 2005), T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), Lecture Notes in Computer Science 3661, Springer, pages 191-204, Kos, Greece, September 12-14, 2005.

10. J. Pearl. *Causality.* Cambridge University Press, New York, 2000.

11. G. Shafer. *The Art of Causal Conjecture.* MIT Press, Cambridge. 1996.

12. G. Shafer. Causality and Responsibility. *Cardozo Law Rev.,* 22, 101-123, 2001.