Background default knowledge and causality ascriptions

Jean-François Bonnefon and **Rui Da Silva Neves**[1] and **Didier Dubois** and **Henri Prade**[2]

Abstract. A model is defined that predicts an agent's ascriptions of causality (and related notions of facilitation and justification) between two events in a chain, based on background knowledge about the normal course of the world. Background knowledge is represented by nonmonotonic consequence relations. This enables the model to handle situations of poor information, where background knowledge is not accurate enough to be represented in, e.g., structural equations. Tentative properties of causality ascriptions are explored, i.e., preference for abnormal factors, transitivity, coherence with logical entailment, and stability with respect to disjunction and conjunction. Empirical data are reported to support the psychological plausibility of our basic definitions.

1 INTRODUCTION

Models of causal ascriptions crucially depend on the choice of an underlying representation for the causality-ascribing agent's knowledge. Unlike standard diagnosis problems (wherein an unobserved cause must be inferred from observed events and known causal links), causality ascription is a problem of describing as 'causal' the link between two observed events in a sequence. The first step in modeling causal ascription is to define causality in the language chosen for the underlying representation of knowledge. In this article, we define and discuss a model of causal ascription that represents knowledge by means of nonmonotonic consequence relations.[3] Indeed, agents often must cope with poor knowledge about the world, under the form of default rules. Clearly, this type of background knowledge is less accurate than, e.g., structural equations. It is nevertheless appropriate to predict causal ascriptions in situations of restricted knowledge. Section 2 presents the logical language we will use to represent background knowledge. Section 3 defines our main notions of causality and facilitation ascriptions. Empirical data are reported to support the distinction between these two notions. Section 4 establishes some formal properties of the model. Section 5 distinguishes the notion of epistemic justification from that of causality. Section 6 relates our model to other works on causality in AI.

2 MODELING BACKGROUND KNOWLEDGE

The agent is supposed to have observed or learned of a sequence of events, e.g.: $\neg B_t$, A_t, B_{t+1}. This expresses that B was false at time t, when A took place, and that B became true afterwards ($t+1$ denotes a time point after t). There is no uncertainty about these events.

Besides, the agent maintains a knowledge-base made of conditional statements of the form 'in context C, if A takes place then B is generally true afterwards', or 'in context C, B is generally true'. These will be denoted by $A_t \wedge C_t \mid\!\sim B_{t+1}$, and by $C_t \mid\!\sim B_t$, respectively. (Time indices will be omitted when there is no risk of confusion.) The conditional beliefs of an agent with respect to B when an action A takes place or not in context C can take three forms: (i) If A takes place B is generally true afterwards: $A_t \wedge C_t \mid\!\sim B_{t+1}$; (ii) If A takes place B is generally false afterwards: $A_t \wedge C_t \mid\!\sim \neg B_{t+1}$; (iii) If A takes place, one cannot say whether B is generally true or false afterwards: $A_t \wedge C_t \not\mid\!\sim B_{t+1}$ and $A_t \wedge C_t \not\mid\!\sim \neg B_{t+1}$.

We assume that the nonmonotonic consequence relation $\mid\!\sim$ satisfies the requirements of 'System P' [18]; namely, $\mid\!\sim$ is reflexive and the following postulates and characteristic properties hold ($\models$ denotes classical logical entailment):

Left Equivalence	$E \mid\!\sim G$ and $E \equiv F$	imply	$F \mid\!\sim G$
Right Weakening	$E \mid\!\sim F$ and $F \models G$	imply	$E \mid\!\sim G$
AND	$E \mid\!\sim F$ and $E \mid\!\sim G$	imply	$E \mid\!\sim F \wedge G$
OR	$E \mid\!\sim G$ and $F \mid\!\sim G$	imply	$E \vee F \mid\!\sim G$
Cautious Monotony	$E \mid\!\sim F$ and $E \mid\!\sim G$	imply	$E \wedge F \mid\!\sim G$
Cut	$E \mid\!\sim F$ and $E \wedge F \mid\!\sim G$	imply	$E \mid\!\sim G$

In addition, we assume $\not\mid\!\sim$ to obey the property of Rational Monotony, a strong version of Cautious Monotony[19]:

Rational Monotony	$E \not\mid\!\sim \neg F$ and $E \mid\!\sim G$	imply	$E \wedge F \mid\!\sim G$

Empirical studies repeatedly demonstrated [1, 2, 5, 10, 24] that System P and Rational Monotony provide a psychologically plausible representation of background knowledge and default inference. Arguments for using nonmonotonic logics in modeling causal reasoning were also discussed in the cognitive science literature [26].

3 ASCRIBING CAUSALITY OR FACILITATION

In the following definitions, A, B, C, and F are either reported actions or statements describing states of affairs, even though notations do not discriminate between them, since the distinction does not yet play a crucial role in the model. When nothing takes place, the persistence of the truth status of statements is assumed in the normal course of things, i.e., $B_t \mid\!\sim B_{t+1}$ and $\neg B_t \mid\!\sim \neg B_{t+1}$.

Assume that in a given context C, the occurrence of event B is known to be exceptional (i.e., $C \mid\!\sim \neg B$). Assume now that F and A are such that $F \wedge C \not\mid\!\sim \neg B$ on the one hand, and $A \wedge F \wedge C \mid\!\sim B$ on the other hand; we will say that in context C, A together with F are perceived as the *cause* of B (denoted $C : A \wedge F \Rightarrow_{\text{ca}} B$), while F alone is merely perceived to have *facilitated* the occurrence of B (denoted $C : F \Rightarrow_{\text{fa}} B$).

Definition 1 (Facilitation ascription). *An agent that, in context C, learns of the sequence $\neg B_t$, F_t, B_{t+1} will judge that $C : F \Rightarrow_{\text{fa}} B$ if it believes that $C \mid\!\sim \neg B$, and that both $F \wedge C \not\mid\!\sim \neg B$ and $F \wedge C \not\mid\!\sim B$.*

[1] LTC-CNRS and DSVP, respectively, 5 allées A. Machado 31058 Toulouse Cedex 9, France. E-mail: {bonnefon,neves}@univ-tlse2.fr.
[2] IRIT-CNRS, 118 Route de Narbonne, 31062 Toulouse Cedex, France. E-mail: {dubois,prade}@irit.fr
[3] This model was advocated in a recent workshop paper [9].

Definition 2 (Causality ascription). *An agent that, in context C, learns of the sequence* $\neg B_t$ *,* A_t*,* B_{t+1} *will judge that* $C : A \Rightarrow_{\text{ca}} B$ *if it believes that* $C \vdash \neg B$*, and* $A \wedge C \vdash B$*.*

Example 1 (Driving while intoxicated). *When driving, one has generally no accident, Drive* $\vdash \neg$*Accident. This is no longer true when driving while drunk, which is not as safe, Drive* $\wedge$ *Drunk* $\nvdash$ $\neg$*Accident); moreover, fast driving while drunk will normally lead to an accident, Drive* $\wedge$ *Fast* $\wedge$ *Drunk* $\vdash$ *Accident. Suppose now that an accident took place after the driver drove fast while being drunk. Fast* $\wedge$ *Drunk will be perceived as the* cause *of the accident, while Drunk will only be judged as having* facilitated *the accident.*

Note that Def. 1 is weaker than saying F 'prevents' $\neg B$ from persisting: $\nvdash$ does not allow the jump from 'not having $\neg B$' to 'B'. In Def. 2, the fact that B is exceptional in context C precludes the possibility for C to be the cause of B – but not the possibility that $B \models C$, i.e., that C is a necessary condition of B. Thus, context can be a necessary condition of B without being perceived as its cause.

An interesting situation arises when an agent only knows that $C \vdash \neg B$ and $F \wedge C \nvdash \neg B$, and learns of the sequence of events $\neg B_t$ (in context C), F_t, B_{t+1}. Although this situation should lead the agent to judge that $C : F_t \Rightarrow_{\text{fa}} B_{t+1}$, it may be tempting to judge that $C : F_t \Rightarrow_{\text{ca}} B_{t+1}$, as long as no other potential cause reportedly took place. Another interesting situation arises when, in context C, an agent learns of the sequence $\neg B_t$, A_t, and B_{t+1}, while it believes that $\neg B_t \wedge C \vdash \neg B_{t+1}$, and that $A_t \wedge C \vdash \neg B_{t+1}$. Then the agent cannot consider that $C : A_t \Rightarrow_{\text{ca}} B_{t+1}$, and it may suspect some fact went unreported: finding about it would amount to a diagnosis problem.

There is no previous empirical support to the distinction we introduce between ascriptions of cause and facilitation. To check whether this distinction has intuitive appeal to lay reasoners, we conducted two experiments in which we presented participants with different sequences of events. We assessed their relevant background knowledge, from which we predicted the relations of cause and facilitation they should ascribe between the events in the sequence. We then compared these predictions to their actual ascriptions.

3.1 Experiment 1

Methods Participants were 46 undergraduate students. None was trained in formal logic or in philosophy. Participants read the stories of three characters, and answered six questions after reading each story. The three characters were described as constantly feeling very tired (an uncommon feeling for them) after two recent changes in their lives: working at night and having a stressful boss (for the first character), working at night and becoming a dad (for the second character), and having a stressful boss and becoming a dad (for the third character). The first three questions assessed participants' background knowledge with respect to (i) the relation between the first event and feeling constantly tired; (ii) the second event and feeling constantly tired; and (iii) the conjunction of the two events and feeling constantly tired. For example:

> What do you think is the most common, the most normal: Working at night and feeling constantly tired, or working at night and not feeling constantly tired? or are those equally common and normal?

Participants who chose the first, second, and third answer were assumed to endorse *WorkNight* $\vdash$ *Tired*; *WorkNight* $\vdash$ $\neg$*Tired*; and (*WorkNight* $\nvdash$ *Tired*) $\wedge$ (*WorkNight* $\nvdash$ $\neg$*Tired*), respectively. The fourth, fifth, and sixth questions assessed participants' ascriptions of causality or facilitation between (i) the first event and feeling constantly tired; (ii) the second event and feeling constantly tired; and (iii) the conjunction of the two events and feeling constantly tired. E.g., one of these questions read:

> Fill in the blank with the word 'caused' or 'facilitated', as seems the most appropriate. If neither seems appropriate, fill in the blank with 'xxx': *Working at night . . . the fact that Julien feels constantly tired.*

Results Out of the 116 ascriptions that the model predicted to be of facilitation, 68% indeed were, 11% were of causality, and 21% were neither. Out of the 224 ascriptions that the model predicted to be of causality, 46% indeed were, 52% were of facilitation, and 2% were neither. The global trend in the results is thus that background knowledge that theoretically matches a facilitation ascription indeed largely leads people to make such an ascription, while background knowledge that theoretically matches a causality ascription leads people to divide equally between causality and facilitation ascriptions. This trend is statistically reliable for almost all ascriptions required by the task. Relevant statistics (χ^2 scores) are higher than 7.7 for 7 out of the 9 ascriptions ($p < .05$, one-tailed, in all cases), and higher than 3.2 for the remaining two ascriptions ($p < .10$, one-tailed, in both cases). From these results, it appears that the notion of facilitation does have intuitive appeal to lay reasoners, and that it is broadly used as defined in our model. In particular, it clearly has a role to play in situations where an ascription of causality sounds too strong a conclusion, but no ascription at all sounds too weak.

3.2 Experiment 2

Experiment 2 was designed to consolidate the results of Experiment 1 and to answer the following questions: Does the fact that background knowledge match Def. 1 or Def. 2 affect the strength of the link participants perceive between two reported events, and does this perceived strength in turn determine whether they make an ascription of causality or facilitation?

Methods Participants were 41 undergraduates. Elements of their background knowledge were assessed as in Exp. 1, in order to select triples of propositions $< Context, Factor, Effect >$ that matched either Def. 1 or Def. 2. E.g., a participant might believe that one has generally no accident when driving, but that one will generally have an accident when driving after some serious drinking; for this participant, $< Drive, SeriousDrinking, Accident >$ is a match with Def. 2. Participants then rated on a 9-point scale how strongly *Factor* and *Effect* were related. Finally, as a measure of ascription, they chose an appropriate term to describe the relation between *Factor* and *Effect*, from a list including 'causes' and 'facilitates'.

Results Out of the 16 ascriptions that the model predicted to be of facilitation, 14 were so, and 2 were of causality. Out of the 25 ascriptions that the model predicted to be of causality, 11 were so, and 14 were of facilitation. Beliefs thus had the expected influence on ascriptions, $\chi^2 = 4.5$, $p < .05$. The trend observed in Experiment 1 is replicated in Experiment 2. We also conducted a *mediation analysis* of our data, which consists in a series of 3 regression analyzes. The direct effect of background knowledge on ascription was significant, $\beta = .33$, $p < .05$. The effect of background knowledge on perceived strength was also significant, $\beta = .41$, $p < .01$. In the third regression, background knowledge and perceived strength were entered simultaneously. Perceived strength was a reliable predictor of ascription, $\beta = .29$, $p < .05$, which was no longer the case for background knowledge, $\beta = .23$, $p > .05$. Data thus meet the requirement

of a mediational effect: Whether the background knowledge of participants matches Def. 1 or Def. 2 determines their final ascription of $C : \mathit{Factor} \Rightarrow_{\mathrm{fa}} \mathit{Effect}$ or $C : \mathit{Factor} \Rightarrow_{\mathrm{ca}} \mathit{Effect}$ through its effect on the perceived strength of the link between *Factor* and *Effect*.

4 PROPERTIES OF CAUSAL ASCRIPTIONS

4.1 Impossibility of mutual causality

Proposition 1. *If $C : A \Rightarrow_{\mathrm{ca}} B$, then it cannot hold that $C : B \Rightarrow_{\mathrm{ca}} A$.*

Proof. If $C : A \Rightarrow_{\mathrm{ca}} B$, it holds that $C \vdash\!\sim \neg A$, $C \wedge A \vdash\!\sim B$, and the sequence $\neg B_t, A_t, B_{t+1}$ has been observed. This is not inconsistent with $C \vdash\!\sim \neg A$, $C \wedge B \vdash\!\sim A$ (the background knowledge part of $C : B \Rightarrow_{\mathrm{ca}} A$), but it is inconsistent with the sequence $\neg A_t, B_t, A_{t+1}$ that would allow the ascription $C : B \Rightarrow_{\mathrm{ca}} A$. □

4.2 Preference for abnormal causes

Psychologists established that abnormal conditions are more likely to be selected by human agents as the cause of an event [17] and more so if this event is itself abnormal [12] (see also [16] in the area of legal philosophy). Our model reflects this preference: Only what is abnormal in a given context can be perceived as facilitating or causing a change in the normal course of things in this context.

Proposition 2. *If $C : A \Rightarrow_{\mathrm{ca}} B$ or $C : A \Rightarrow_{\mathrm{fa}} B$, then $C \vdash\!\sim \neg A$.*

Proof. $C \vdash\!\sim \neg A$ is false when either $C \vdash\!\sim A$ or $C \not\vdash\!\sim \neg A$. If $C \vdash\!\sim A$, it cannot be true that both $C \vdash\!\sim \neg B$ and either $A \wedge C \not\vdash\!\sim \neg B$ (the definition of $C : A \Rightarrow_{\mathrm{fa}} B$) or $A \wedge C \vdash\!\sim B$ (the definition of $C : A \Rightarrow_{\mathrm{ca}} B$). This is due to the Cautious Monotony property of $\vdash\!\sim$, which forces $C \wedge A \vdash\!\sim \neg B$ from $C \vdash\!\sim A$ and $C \vdash\!\sim \neg B$. Likewise, the Rational Monotony of $\vdash\!\sim$ forces $C \wedge A \vdash\!\sim \neg B$ from $C \not\vdash\!\sim \neg A$ and $C \vdash\!\sim \neg B$; thus, it cannot be the case that $C : A \Rightarrow_{\mathrm{fa}} B$ or $C : A \Rightarrow_{\mathrm{ca}} B$ when $C \not\vdash\!\sim \neg A$. □

Example 2 (The unreasonable driver). *Let us imagine an agent who believes it is normal to be drunk in the context of driving ($\mathit{Drive} \vdash\!\sim \mathit{Drunk}$). This agent may think that it is exceptional to have an accident when driving ($\mathit{Drive} \vdash\!\sim \neg\mathit{Accident}$). In that case, the agent cannot but believe that accidents are exceptional as well when driving while drunk: $\mathit{Drive} \wedge \mathit{Drunk} \vdash\!\sim \neg\mathit{Accident}$. As a consequence, when learning that someone got drunk, drove his car, and had an accident, this agent will neither consider that $C : \mathit{Drunk} \Rightarrow_{\mathrm{fa}} \mathit{Accident}$ nor that $C : \mathit{Drunk} \Rightarrow_{\mathrm{ca}} \mathit{Accident}$.*

4.3 Transitivity

Def. 2 does not grant general transitivity to $\Rightarrow_{\mathrm{ca}}$. If $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : B \Rightarrow_{\mathrm{ca}} D$, it does not always follow that $C : A \Rightarrow_{\mathrm{ca}} D$. Formally: $C \vdash\!\sim \neg B$ and $A \wedge C \vdash\!\sim B$ and $C \vdash\!\sim \neg D$ and $B \wedge C \vdash\!\sim D$ do not entail $C \vdash\!\sim \neg D$ and $A \wedge C \vdash\!\sim D$, because $\vdash\!\sim$ itself is not transitive. Although $\Rightarrow_{\mathrm{ca}}$ is not generally transitive, it becomes so in one particular case.

Proposition 3. *If $C : A \Rightarrow_{\mathrm{ca}} B$, $C : B \Rightarrow_{\mathrm{ca}} D$, and $B \wedge C \vdash\!\sim A$, then $C : A \Rightarrow_{\mathrm{ca}} D$.*

Proof. From the definition of $C : B \Rightarrow_{\mathrm{ca}} D$, it holds that $B \wedge C \vdash\!\sim D$. From $B \wedge C \vdash\!\sim A$ and $B \wedge C \vdash\!\sim D$, applying Cautious Monotony yields $A \wedge B \wedge C \vdash\!\sim D$, which together with $A \wedge C \vdash\!\sim B$ (from the definition of $C : A \Rightarrow_{\mathrm{ca}} B$) yields by Cut $A \wedge C \vdash\!\sim D$; since it holds from the definition of $C : B \Rightarrow_{\mathrm{ca}} D$ that $C \vdash\!\sim \neg D$, the two parts of the definition of $C : A \Rightarrow_{\mathrm{ca}} D$ are satisfied. □

Example 3 (Mud on the plates). *Driving back from the countryside, you get a fine because your plates are muddy, $\mathit{Drive} : \mathit{Mud} \Rightarrow_{\mathrm{ca}} \mathit{Fine}$. Let us assume that you perceive your driving to the countryside as the cause for the plates to be muddy, $\mathit{Drive} : \mathit{Countryside} \Rightarrow_{\mathrm{ca}} \mathit{Mud}$. For transitivity to apply, i.e., to judge that $\mathit{Drive} : \mathit{Countryside} \Rightarrow_{\mathrm{ca}} \mathit{Fine}$, it must hold that $\mathit{Mud} \wedge \mathit{Drive} \vdash\!\sim \mathit{Countryside}$: If mud on your plates usually means that you went to the countryside, then the trip can be considered the cause of the fine. If the presence of mud on your plates does not allow to infer that you went to the countryside (perhaps you also regularly drive through muddy streets where you live), then transitivity is not applicable; you will only consider that the mud caused the fine, not that the trip did.*

4.4 Entailment and causality ascriptions

Classical entailment $\models$ does not preserve $\Rightarrow_{\mathrm{ca}}$. If $C : A \Rightarrow_{\mathrm{ca}} B$ and $B \models B'$, one cannot say that $C : A \Rightarrow_{\mathrm{ca}} B'$. Indeed, while $A \wedge C \vdash\!\sim B'$ follows by right weakening [18] from $A \wedge C \vdash\!\sim B$, it is not generally true that $C \vdash\!\sim \neg B'$, given that $C \vdash\!\sim \neg B$. Besides, according to Definition 2, if $A' \models A$, the fact that $C : A \Rightarrow_{\mathrm{ca}} B$ does not entail that $C : A' \Rightarrow_{\mathrm{ca}} B$, since $C \vdash\!\sim \neg B$ and $A \wedge C \vdash\!\sim B$ do not entail $A' \wedge C \vdash\!\sim B$ when $A' \models A$. This fact is due to the extreme cautiousness of System P. It is contrasted in the following example with Rational Monotony.

Example 4 (Stone throwing). *An agent believes that a window shattered because a stone was thrown at it ($\mathit{Window} : \mathit{Stone} \Rightarrow_{\mathrm{ca}} \mathit{Shatter}$), based on its beliefs that $\mathit{Window} \vdash\!\sim \neg\mathit{Shatter}$ and $\mathit{Stone} \wedge \mathit{Window} \vdash\!\sim \mathit{Shatter}$. Using the Cautious Monotony of System P, it is not possible to predict that the agent would make a similar ascription if a* small *stone had been thrown (SmallStone), or if a* white *stone had been thrown (WhiteStone), or even if a* big *stone had been thrown (BigStone), although it holds that $\mathit{SmallStone} \models \mathit{Stone}$, $\mathit{WhiteStone} \models \mathit{Stone}$, and $\mathit{BigStone} \models \mathit{Stone}$. Adding Rational Monotony [19] to System P allows the ascriptions $\mathit{Window} : \mathit{BigStone} \Rightarrow_{\mathrm{ca}} \mathit{Shatter}$ and $\mathit{Window} : \mathit{WhiteStone} \Rightarrow_{\mathrm{ca}} \mathit{Shatter}$, but also $\mathit{Window} : \mathit{SmallStone} \Rightarrow_{\mathrm{ca}} \mathit{Shatter}$. To block this last ascription, it would be necessary that the agent has specific knowledge about the harmlessness of small stones, such as $\mathit{Window} \wedge \mathit{Smallstone} \not\vdash\!\sim \mathit{Shatter}$ or even $\mathit{Window} \wedge \mathit{Smallstone} \vdash\!\sim \neg\mathit{Shatter}$.*

4.5 Stability w.r.t. disjunction and conjunction

$\Rightarrow_{\mathrm{ca}}$ is stable with respect to disjunction, both on the right and on the left, and stable w.r.t. conjunction on the right.

Proposition 4. *The following properties hold:*

1. *If $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$, then $C : A \Rightarrow_{\mathrm{ca}} B \vee B'$.*
2. *If $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A' \Rightarrow_{\mathrm{ca}} B$, then $C : A \vee A' \Rightarrow_{\mathrm{ca}} B$.*
3. *If $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$, then $C : A \Rightarrow_{\mathrm{ca}} B \wedge B'$.*

Proof. Applying AND to the first part of the definitions of $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$, i.e., $C \vdash\!\sim \neg B$ and $C \vdash\!\sim \neg B'$, yields $C \vdash\!\sim \neg B \wedge \neg B'$, and thus $C \vdash\!\sim \neg(B \vee B')$. Now, applying AND to the second part of the definitions of $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$, i.e., $A \wedge C \vdash\!\sim B$ and $A \wedge C \vdash\!\sim B'$, yields $A \wedge C \vdash\!\sim B \wedge B'$, which together with Right Weakening yields $A \wedge C \vdash\!\sim B \vee B'$. The definition of $C : A \Rightarrow_{\mathrm{ca}} B \vee B'$ is thus satisfied. The proof of Fact 2 is obtained by applying OR to the second part of the definitions of $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$. Finally, applying AND to the first part of the definitions of $C : A \Rightarrow_{\mathrm{ca}} B$ and $C : A \Rightarrow_{\mathrm{ca}} B'$, i.e., $C \vdash\!\sim \neg B$ and $C \vdash\!\sim \neg B'$, yields $C \vdash\!\sim \neg B \wedge \neg B'$, which together with Right Weakening, yields

$C \mid\!\sim \neg B \vee \neg B'$, and thus $C \mid\!\sim \neg(B \wedge B')$. Now, applying AND to the second part of the definitions of $C : A \Rightarrow_{\text{ca}} B$ and $C : A \Rightarrow_{\text{ca}} B'$, i.e., $A \wedge C \mid\!\sim B$ and $A \wedge C \mid\!\sim B'$, yields $A \wedge C \mid\!\sim B \wedge B'$. The definition of $C : A \Rightarrow_{\text{ca}} B \wedge B'$ is thus satisfied. □

$\Rightarrow_{\text{ca}}$ is not stable w.r.t. conjunction on the left. If $C : A \Rightarrow_{\text{ca}} B$ and $C : A' \Rightarrow_{\text{ca}} B$, then it is not always the case that $C : A \wedge A' \Rightarrow_{\text{ca}} B$ (see example 5). This lack of stability is once again due to the cautiousness of System P; for $C : A \wedge A' \Rightarrow_{\text{ca}} B$ to hold, it is necessary that $C \wedge A \mid\!\sim A'$ or, alternatively, that $C \wedge A' \mid\!\sim A$. Then Cautious Monotony will yield $A \wedge A' \wedge C \mid\!\sim B$. Rational Monotony can soften this constraint and make it enough that $C \wedge A \not\mid\!\sim \neg A'$ or $C \wedge A' \not\mid\!\sim \neg A$.

Example 5 (Busy professors). *Suppose that professors in your department seldom show up early at the office (Prof $\mid\!\sim \neg$Early). However, they generally do so when they have tons of student papers to mark (Prof $\wedge$ Mark $\mid\!\sim$ Early), and also when they have a grant proposal to write (Prof $\wedge$ Grant $\mid\!\sim$ Early). When learning that a professor had tons of papers to grade and that she came in early, you would judge that Prof : Mark $\Rightarrow_{\text{ca}}$ Early. Likewise, when learning that a professor had a grant proposal to write and came in early, you would judge that Prof : Grant $\Rightarrow_{\text{ca}}$ Early. But what if you learn that a professor had tons of papers to grade* and *a grant proposal to write and that she came in early? That would depend on whether it is an exceptional situation to have to deal with both tasks on the same day. If it is not exceptional (Mark $\not\mid\!\sim \neg$Grant), then you will judge that Prof : Mark $\wedge$ Grant $\Rightarrow_{\text{ca}}$ Early. If, on the contrary, Mark $\wedge$ Grant is an exceptional event, it does not hold anymore that Mark $\wedge$ Grant $\mid\!\sim$ Early, and it is thus impossible to feel sure about Prof : Mark $\wedge$ Grant $\Rightarrow_{\text{ca}}$ Early. For example, it might be the case that faced with such an exceptional workload, a professor will prefer working at home all day rather than coming to the office. In that case, her coming in early would be due to another factor, e.g., a meeting that could not be cancelled.*

5 ASCRIPTIONS OF JUSTIFICATION

Perceived causality as expressed in Def. 2 should be distinguished from the situation that we term 'justification.' We write that $C : A \Rightarrow_{\text{ju}} B$ when an agent judges that the occurrence of A in context C gave reason to expect the occurrence of B.

Definition 3 (Justification). *An agent that learns in context C of the sequence $\neg B_t$, A_t, B_{t+1} will judge that $C : A \Rightarrow_{\text{ju}} B$ if it believes that $C \not\mid\!\sim \neg B$, $C \not\mid\!\sim B$ and $A \wedge C \mid\!\sim B$.*

Faced with facts C, $\neg B_t$, A_t, B_{t+1}, an agent believing that $C \not\mid\!\sim \neg B$, $C \not\mid\!\sim B$ and $A \wedge C \mid\!\sim B$ may doubt that the change from $\neg B_t$ to B_{t+1} is really due to A_t, although the latter is indeed the very reason for the lack of surprise at having B_{t+1} reported. Indeed, situation $\neg B_t$ at time t appears to the agent to be contingent, since it is neither a normal nor an abnormal course of things in context C. This clearly departs from the situation where $C \mid\!\sim \neg B$ and $A \wedge C \mid\!\sim B$, wherein the agent will judge that $C : A \Rightarrow_{\text{ca}} B$. In a nutshell, the case whereby $C \not\mid\!\sim \neg B$, $C \not\mid\!\sim B$ and $A \wedge C \mid\!\sim B$ cannot be interpreted as the recognition of a causal phenomenon by an agent: All that can be said is that reporting A caused the agent to start believing B, and that she should not be surprised of having B_{t+1} reported.

What we call justification is akin to the notion of explanation following Spohn [27]: Namely, 'A is a reason for B' when raising the epistemic rank for A raises the epistemic rank for B. Gärdenfors [11] captured this view to some extent, assuming that A is a reason for B if B is not retained in the contraction of A. Williams et al. [29] could account for the Spohnian view in a more refined way using kappa-rankings and transmutations, distinguishing between weak and strong explanations. As our framework can easily be given a possibilistic semantics [4], it could properly account for this line of thought, although our distinction between perceived causation and epistemic justification is not the topic of the above works.

6 RELATED WORKS

Causality plays a central role in at least two problems studied in AI, diagnosis and the simulation of dynamical systems. Diagnosis problems are a matter of abduction: One takes advantage of the knowledge of some causal links to infer the most plausible causes of an observed event [23]. In this setting, causality relations are often modelled by conditional probabilities $P(\text{effect}|\text{cause})$.[4] Dynamical systems are modelled in AI with respect, e.g., to qualitative physics [6], and in logics of action. The relation of nonmonotonic inference to causality has already been emphasized by authors dealing with reasoning about actions and the frame problem [13, 20, 28]. Material implication being inappropriate to represent a causal link, these approaches define a 'causal rule' as 'there is a cause for effect B to be true if it is true that A has just been executed', where 'there is a cause for' is modelled by a modal operator.

The problem discussed in this paper is not, however, one of classical diagnosis. Neither does it deal with the qualitative simulation of dynamical systems, nor with the problem of describing changes caused by the execution of actions, nor with what does not change when actions are performed. We are concerned here with a different question, namely the explanation of a sequence of reported events, in terms of pairs of events that can be considered as related by a causality relation. In that sense, our work is reminiscent of the 'causal logic' of Shafer [25], which provides a logical setting that aims at describing the possible relations of concomitance between events when an action takes place. However, Shafer's logic does not leave room for abnormality. This notion is central in our approach, as it directly relates to the relations of qualitative independence explored in [7] – causality and independence being somewhat antagonistic notions.

Following [22], Halpern and Pearl [14, 15] have proposed a model that distinguishes real causes ('cause in fact') from potential causes, by using an a priori distinction between 'endogenous' variables (the possible values of which are governed by structural equations, for example physical laws), and 'exogenous' variables (determined by external factors). Exogenous variables cannot be deemed causal. Halpern and Pearl's definition of causality formalizes the notion of an active causal process. More precisely, the fact A that a subset of endogenous variables has taken some definite values is the real cause of an event B if (i) A and B are true in the real world, (ii) this subset is minimal, (iii) another value assignment to this subset would make B false, the values of the other endogenous variables that do not directly participate to the occurrence of B being fixed in some manner, and (iv) A alone is enough for B to occur in this context. This approach, thanks to the richness of background knowledge when it is represented in structural equations, makes it possible to treat especially difficult examples. Our model is not to be construed as an alternative or a competitor to models based on structural equations. Indeed, we see our approach as either a 'plan B' or a complement to structural equation modeling. One might not have access to the accurate

[4] Nevertheless, Bayesian networks [21] (that represent a joint probability distribution by means of a directed graph) do not necessarily reflect causal links between their nodes, for different graphical representations can be obtained depending on the ordering in which variables are considered [8].

information needed to build a structural equation model; in this case, our less demanding model might still be operable. Alternatively, a decision support system may be able to build a structural equation model of the situation, although its users only have access to qualitative knowledge. In that case, the system will be able to compare its own causality ascriptions to the conclusions of the qualitative model, and take appropriate explanatory steps, would those ascriptions be too different. Indeed, our model does not aim at identifying the true, objective cause of an event, but rather at predicting what causal ascription an agent would make based on the limited information it has at its disposal.

Models based on structural equations are often supplemented with the useful notion of *intervention*. In many situations, finding the cause of an event will be much easier if the agent can directly intervene in the manner of an experimenter. In future work, we intend to explore the possibility of supplementing our own model with a similar notion by means of a $\mathbf{do}(\bullet)$ operator. An ascription of causality (resp., facilitation) would be made iff the requirements of Definition 2 (resp., 1) are met both for A, B, C *and* for $\mathbf{do}(A)$, B, C, where $\mathbf{do}(A)$ means that the occurrence of A is forced by an intervention [22]. As for now, we only give a brief example of how such an operator can be used in our approach.

Example 6 (Yellow teeth). *An agent learns that someone took up smoking, that this person's teeth yellowed, and that this person developed lung cancer. The agent believes that generally speaking, it is abnormal to be a smoker, to have yellow teeth, and to develop lung cancer (resp., ($C \vdash \neg Smoke$, $C \vdash \neg Yellow$, $C \vdash \neg Lung$). The agent believes that it is normal for smokers to have yellow teeth ($C \wedge Smoke \vdash Yellow$) and to develop lung cancer ($C \wedge Smoke \vdash Lung$), and that it is not abnormal for someone who has yellow teeth to develop lung cancer ($C \wedge Yellow \nvdash \neg Lung$). From these beliefs and observations, Definitions 1 and 2 would allow for various ascriptions, including the following one: Smoking caused the yellow teeth which* in turn *facilitated lung cancer. With the additional constraint based on the* $\mathbf{do}(\bullet)$ *operator, only one set of ascriptions remains possible:* Both *the yellow teeth and the lung cancer were caused by smoking. Yellow teeth cannot be said anymore to facilitate lung cancer because, inasmuch as lung cancer is generally abnormal, it holds that* $C \wedge \mathbf{do}(Yellow) \vdash \neg Lung$*: There is no reason to think that one will develop lung cancer after painting one's teeth yellow.*

7 CONCLUDING REMARKS

We have presented a simple qualitative model of the causal ascriptions an agent will make from its background default knowledge, when confronted with a series of events. In addition to supplementing this model with a $\mathbf{do}(\bullet)$ operator, we intend to extend our present work in three main directions. First, we should be able to equip our framework with possibilistic qualitative counterparts to Bayesian networks [3], since System P augmented with Rational Monotony can be represented in possibilistic logic [4]. Second, we will derive postulates for causality from the independence postulates presented in [7]. Finally, in parallel to further theoretical elaboration, we will maintain a systematic experimental program that will test the psychological plausibility of our definitions, properties, and postulates.

ACKNOWLEDGMENTS

This work was supported by a grant from the Agence Nationale pour la Recherche, project number NT05-3-44479.

REFERENCES

[1] S. Benferhat, J. F. Bonnefon, and R. M. Da Silva Neves, 'An experimental analysis of possibilistic default reasoning', in *KR2004*, pp. 130–140. AAAI Press, (2004).

[2] S. Benferhat, J. F. Bonnefon, and R. M. Da Silva Neves, 'An overview of possibilistic handling of default reasoning: An experimental study', *Synthese*, **146**, 53–70, (2005).

[3] S. Benferhat, D. Dubois, L. Garcia, and H. Prade, 'On the transformation between possibilistic logic bases and possibilistic causal networks', *International Journal of Approximate Reasoning*, **29**, 135–173, (2002).

[4] S. Benferhat, D. Dubois, and H. Prade, 'Nonmonotonic reasoning, conditional objects and possibility theory', *Artificial Intelligence*, **92**, 259–276, (1997).

[5] R. M. Da Silva Neves, J. F. Bonnefon, and E. Raufaste, 'An empirical test for patterns of nonmonotonic inference', *Annals of Mathematics and Artificial Intelligence*, **34**, 107–130, (2002).

[6] J. de Kleer and J. S. Brown, 'Theories of causal ordering', *Artificial Intelligence*, **29**, 33–61, (1986).

[7] D. Dubois, L. Fariñas Del Cerro, A. Herzig, and H. Prade, *A roadmap of qualitative independence*, volume 15 of *Applied Logic series*, 325–350, Kluwer, Dordrecht, The Netherlands, 1999.

[8] D. Dubois and H. Prade, 'Probability theory in artificial intelligence. Book review of J. Pearl's 'Probabilistic Reasoning in Intelligent Systems", *Journal of Mathematical Psychology*, **34**, 472–482, (1999).

[9] D. Dubois and H. Prade, 'Modeling the role of (ab)normality in the ascription of causality judgements by agents', in *NRAC'05*, pp. 22–27, (2005).

[10] M. Ford, 'System LS: A three tiered nonmonotonic reasoning system', *Computational Intelligence*, **20**, 89–108, (2004).

[11] P. Gärdenfors, 'The dynamics of belief systems: Foundations vs. coherence theories', *Revue Internationale de Philosophie*, **44**, 24–46, (1990).

[12] I. Gavansky and G. L. Wells, 'Counterfactual processing of normal and exceptional events', *Journal of Experimental Social Psychology*, **25**, 314–325, (1989).

[13] E. Giunchiglia, J. Lee, N. McCain, V. Lifschitz, and H. Turner, 'Nonmonotonic causal theories', *Artificial Intelligence*, **153**, 49–104, (2004).

[14] J. Halpern and J. Pearl, 'Causes and explanations: A structural-model approach — part 1: Causes', *British Journal for the Philosophy of Science*, (to appear).

[15] J. Halpern and J. Pearl, 'Causes and explanations: A structural-model approach — part 2: Explanations', *British Journal for the Philosophy of Science*, (to appear).

[16] H. L. A. Hart and T. Honoré, *Causation in the law*, Oxford University Press, Oxford, 1985.

[17] D. J. Hilton and B. R. Slugoski, 'Knowledge-based causal attribution: The abnormal conditions focus model', *Psychological Review*, **93**, 75–88, (1986).

[18] S. Kraus, D. Lehmann, and M. Magidor, 'Nonmonotonic reasoning, preferential models and cumulative logics', *Artificial Intelligence*, **44**, 167–207, (1990).

[19] D. Lehmann and M. Magidor, 'What does a conditional knowledge base entails?', *Artificial Intelligence*, **55**, 1–60, (1992).

[20] N. McCain and H. Turner, 'A causal theory of ramifications and qualifications', in *IJCAI'95*, San Francisco, CA, (1995). Morgan Kaufmann.

[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[22] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, 2000.

[23] Y. Peng and J. A. Reggia, *Abductive Inference Models for Diagnostic Problem-Solving*, Springer Verlag, Berlin, 1990.

[24] N. Pfeifer and G. D. Kleiter, 'Coherence and nonmonotonicity in human reasoning', *Synthese*, **146**, 93–109, (2005).

[25] G. Shafer, 'Causal logic', in *ECAI'98*, pp. 711–719, Chichester, England, (1998). Wiley.

[26] Y. Shoham, 'Nonmonotonic reasoning and causation', *Cognitive Science*, **14**, 213–252, (1990).

[27] W. Spohn, 'Deterministic and probabilistic reasons and causes', *Erkenntnis*, **19**, 371–393, (1983).

[28] H. Turner, 'A logic of universal causation', *Artificial Intelligence*, **113**, 87–123, (1999).

[29] M.-A. Williams, M. Pagnucco, N. Foo, and B. Sims, 'Determining explanations using transmutations', in *IJCAI'95*, pp. 822–830. Morgan Kaufmann, (1995).