A Comparative Study of Six Formal Models of Causal Ascription

Salem Benferhat[1], Jean-François Bonnefon[3], Philippe Chassy[3],
Rui Da Silva Neves[3], Didier Dubois[3], Florence Dupin de Saint-Cyr[3],
Daniel Kayser[2], Farid Nouioua[2], Sara Nouioua-Boutouhami[2], Henri Prade[3],
and Salma Smaoui[1]

[1] Université d'Artois
[2] Université Paris 13
[3] Université de Toulouse

Abstract. Ascribing causality amounts to determining what elements in a sequence of reported facts can be related in a causal way, on the basis of some knowledge about the course of the world. The paper offers a comparison of a large span of formal models (based on structural equations, non-monotonic consequence relations, trajectory preference relations, identification of violated norms, graphical representations, or connectionism), using a running example taken from a corpus of car accident reports. Interestingly enough, the compared approaches focus on different aspects of the problem by either identifying all the potential causes, or selecting a smaller subset by taking advantages of contextually abnormal facts, or by modeling interventions to get rid of simple correlations. The paper concludes by a general discussion based on a battery of criteria (several of them being proper to AI approaches to causality).

1 Introduction

Causality is a protean and complex notion. Accordingly, multiple models of causation were developed in Artificial Intelligence (AI). Indeed, the idea of causality pervades several important AI problems, e.g., in the diagnosis of the potential causes from observed effects; in the induction of causal laws from series of observations; in logics of action; in the qualitative simulation of dynamical systems (when propagating constraints in influence graphs).

In this article, we focus on the perception of causal relations and causal ascription. Unsurprisingly, models proposed for causal ascription generally agree in some way with the idea of relating causality to *counterfactuality:* the counterfactual 'Had A not taken place, B would not have occurred' sounds as a necessary condition for declaring that A causes B. This idea underlies many approaches, from that initiated in modal logic years ago [1], to the approach more recently advocated by Pearl [2] in a probabilistic setting. However, as we will see, providing a full account of the way causality is perceived may also benefit from the identification of facts found 'abnormal' by agents in given contexts, among a series of reported events.

S. Greco and T. Lukasiewicz (Eds.): SUM 2008, LNAI 5291, pp. 47–62, 2008.

It is a daunting task to compare the definitions and properties of models of causal perception and ascription. A preliminary and useful step toward such an achievement, though, consists of illustrating the behavior of the models through a series of well-chosen examples. The examples should be realistic and relevant to the real world—but not so complex that they would no longer be manageable. They should strike the right balance between traditional, simplistic examples (the causal equivalents of the Tweety problem in default reasoning), and intractable scenarios such as the circumstances of Princess Diana's death. Traffic accidents reports offer an excellent source for such examples. They describe genuine events; they naturally lend themselves to causal analysis (in fact, they are often used for that very purpose); and they occur in a relatively self-contained micro-universe. We were able to gain access to a database of traffic accident reports submitted by drivers to insurance companies (the current sample consists of about one hundred reports of accidents that happened in France in recent years). We then submitted these reports to a battery of formal models (based on structural equations, nonmonotonic logics, graphs, or connectionism). Due to space limitations, we will restrict ourselves to one report:

Example 1 (Accident). We were at $***$, I was surprised by the person who braked in front of me, not having the option of changing lane and the road being wet, I could not stop completely in time.

All models will use the same common core of variables and pieces of knowledge. Variables are: Acc (occurrence of an accident), Wet (road being wet), $Brak$ (driver B brakes in front of driver A), $Reac$ (driver A brakes in reaction to driver B's braking), with variants $ReacS$ and $ReacL$ (driver A brakes shortly after B brakes, or with a longer delay), Ncl (A does not have the option of changing lane), Sur (A is surprised). Additional variables may be introduced in some models to display interesting variants of the example. Logical constraints exist among the variables: (1) $Reac \equiv ReacS \vee ReacL$, (2) $\neg ReacS \vee \neg ReacL$, (3) $\neg Reac \vee Brak$. The common core of knowledge is: (4) Accidents are abnormal, (5) Being surprised is abnormal, (6) $ReacL$ and Wet promote Acc, (7) $Brak$ and Ncl and Sur promote $ReacL$, (8) $Brak$ and Ncl and $\neg Sur$ promote $ReacS$. Each model will incorporate this common core of knowledge, up to its representational specificities (especially regarding the formalization of what 'abnormal' and 'promote' mean). Again, additional pieces of knowledge may be introduced to highlight interesting aspects of the models. The presentation of each model will follow the same structure: brief motivation, reminder of definitions, summary of characteristic features, treatment of the example, discussion. Although the original purpose of this paper is to compare the models mainly on the basis of formal considerations, their discussion will occasionally point to experimental data, when they exist.

2 Structural Equations Model

Halpern and Pearl [3] propose a model allowing identification of 'actual causes.' The model distinguishes between 'endogenous' and 'exogenous' variables. Assigned values of endogenous variables are governed by structural equations, whereas exogenous variables are assumed to be known and out of control. Only endogenous variables can be causes or be caused. Background knowledge in such model is given by the context and structural equations. A causal model is denoted by $M=(U,V,F)$ where U and V are sets of exogenous and endogenous variables. F is a function that assigns a value to each variable given each value of its parents. Each assignment of the exogenous variables $U = u$ determines a unique value x of each subset X of endogenous variables (i.e. $X \subseteq V$).

Definition 1. *The event $X = x$ is said to be an actual cause of an event ϕ if and only if:*

1. *$X(u) = x$ and $\phi(u)$ is true (when U takes the value u).*
2. *There exists a partition (Z, W) of V with $X \subseteq Z$ and some settings (x', w') of (X, W) such that if $Z(u) = z^*$ (z^* is the value assigned to Z when $U = u$), both of the following conditions hold:*
   a) *$\phi_{X \leftarrow x', W \leftarrow w'}(u)$ is false, namely, if X is set to x' and W is set to w' then ϕ becomes false.*
   b) *$\phi_{X \leftarrow x, W' \leftarrow [w'], Z' \leftarrow [z^*]}(u)$ is true for all $W' \subseteq W$ and for all $Z' \subseteq Z$. Namely, if X is set to x, W' is set to $[w']$ ($[w']$ is an instantiation of W' consistent with w'), and Z' is set to $[z^*]$ then ϕ remains true.*
3. *The subset X is minimal.*

Pearl and Halpern also proposed an extended causal model to deal with excluded settings. The extended version of Definition 1 consists of adding to the tuple (U, V, F) a set E that contains allowed settings of endogenous variables. E functions as some kind of integrity constraint. In our example, all settings are considered allowed, and the extended causal model collapses with Definition 1. The causal model described above can be represented using a graph, in which nodes are corresponding to variables in V and an edge from X to Y exists if the value of Y depends from the value of X. This graph is a directed acyclic graph (DAG) representing the relationships between variables which are fully specified by structural equations.

Example. We model the example presented in the introduction using only endogenous variables. Variables $Brak$, Ncl, Sur, Wet and Acc have the same meaning as previously given. The variable $Reac$ is a ternary variable taking its values in $\{ReacS, ReacL, NoReac\}$ where $NoReac$ stands for 'A does not brake'. For simplicity, we consider that all settings are allowed ($E = \emptyset$). The structural equations are given by:

- $Acc = \begin{cases} 1 \text{ if } wet = 1 \text{ and } Reac = ReacL \\ 0 \text{ otherwise} \end{cases}$
- $Reac = \begin{cases} NoReac \text{ if } Brak = 0 \text{ or } Ncl = 0 \\ ReacS \quad \text{if } Sur = 0 \text{ and } Brak = 1 \text{ and } Ncl = 1 \\ ReacL \quad \text{if } Sur = 1 \text{ and } Brak = 1 \text{ and } Ncl = 1 \end{cases}$

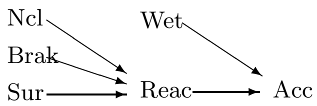

Fig. 1. A causal network

This model can be represented by the DAG given in Figure 1. Assume that the actual context is $Sur = 1$ and $Brak = 1$ and $Ncl = 1$ and $Reac = ReacL$ and $Wet = 1$ and $Acc = 1$. Now let us find causes of the event $Acc = 1$ in this context. We first check if $Ncl = 1$ is a cause of $Acc = 1$. Condition 1 holds since $Ncl = 1$ and $Acc = 1$ is true in the actual world. Given the partition $Z = \{Ncl, Reac, Acc\}$ and $W = \{Sur, Brak, Wet\}$, it is easy to check that maintaining the actual context ($w' = \{Sur = 1, Brak = 1, Wet = 1\}$) and changing the value of Ncl from true to false (i.e. $Ncl = 0$) is enough to change the value of Acc from true to false (i.e. $Acc = 0$). Condition 2a is satisfied for w'. Setting Ncl to true and setting all subsets $\hat{W}$ (e.g. $\{Sur = 1\}$) of W to their values $\hat{w}$ (consistent with w') is not enough to change the value of Acc which remains true (i.e. $Acc = 1$). Thus condition 2b is also satisfied. It is obvious that $Ncl = 1$ is minimal (condition 3). We conclude that $Ncl = 1$ is a cause of $Acc = 1$. Maintaining the same context and setting $W = w'$ we obtain that each event is a cause of $Acc = 1$.

Discussion. Despite the fact that this model allows to handle notorious case studies on causality, it still presents some limitations. Reasoning with structural equations means that all required information must be available (this makes sense in some physics applications where structural equations reflects physical laws among a limited set of variables). Unfortunately, this is not always the case, which may limit the scope of application. For example, rules (4) and (5) in the introduction cannot be easily represented [4], and using non-monotonic rules may be an interesting alternative. Besides, requiring that any assignment of exogenous variables uniquely determines the value of all endogenous variables is not always natural. The apparent lack of selective power of this model may also be considered a weakness, as an event is very easily designated as a cause of another. E.g., in our example, each event is a cause of $Acc = 1$. In order to select preferred causes, it may be interesting to assign 'weights' on the basis of levels of normality assigned to each cause according to its implication in making the event happening.

3 Nonmonotonic Logic Approaches

As discussed by philosophers of law [5], and experimentally checked by psychologists,'abnormal' facts are privileged when providing causal explanations [6]. Added to the insufficiency of material implication for representing causation, this naturally leads to consider nonmonotonic logic-based approaches for causal ascriptions. Relations between nonmonotonic inference and causality have already been emphasized by authors dealing with reasoning about actions and the frame problem [7,8]. ‘Causal rules’ are understood there 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 a modal operator. However, we are not interested in the following in the proper modeling of already established causality relations, but rather in the ascription of causality relations in a reported series of facts or events.

3.1 Nonmonotonic Consequence Approach

To reflect the fact that human agents cannot always couch their beliefs in precise probabilistic terms, Bonnefon et al. [9,10] offer a qualitative counterpart to probabilistic conceptions of causality. This approach is based on pieces of default knowledge, and privileges the role of abnormal events in a given context.

Definition 2. *Assume an agent learns of the sequence $\neg B_t$, A_t, B_{t+k}. Call K_t (the* context*) the conjunction of all other facts known by the agent at time t. Let $\mid\!\sim$ denote a nonmonotonic consequence relation. If the agent believes $K \mid\!\sim \neg B$ and $K \wedge A \mid\!\sim B$, the agent will perceive A to cause B in context K, denoted $A \rhd B$. If the agent believes that $K \mid\!\sim \neg B$, and $K \wedge A \not\mid\!\sim \neg B$ rather than $K \wedge A \mid\!\sim B$, then A is perceived as* facilitating *rather than causing B, denoted $A \blacktriangleright B$.*[1]

In the definitions of $\rhd$ and $\blacktriangleright$, $\mid\!\sim$ is a preferential entailment in the sense of Kraus et al. [11], and a rational closure entailment, respectively. This definition has noticeable features. E.g., causes and facilitations are abnormal in context: If $A \rhd B$ or $A \blacktriangleright B$ then $K \mid\!\sim \neg A$. Furthermore, causality is transitive only in particular cases: If A is the normal way of getting B in context K, i.e., $K \wedge B \mid\!\sim A$, and if $A \rhd B$ and $B \rhd C$, then $A \rhd C$. The practical significance of Def. 2 (including the distinction between causation and facilitation), as well as the restricted transitivity property, have been validated by behavioral experiments. Note that a facilitation is abnormal and is not a necessary condition for the effect, in contrast to an enabling condition (see below).

Example. The story that unfolds in the report reads: At any point in time, Wet is true. Initially, Ncl is true, and Acc, $Brak$, $Reac$, and Sur are false. Next, $Brak$ and Sur become true. Next, $ReacL$ becomes true. Finally, Acc becomes true. The formalization of the common core of knowledge is : (4) $\mid\!\sim \neg Acc$; (5) $\mid\!\sim \neg Sur$; (6) $ReacL \wedge Wet \mid\!\sim Acc$; (7) $Brak \wedge Ncl \wedge Sur \mid\!\sim ReacL$; (8) $Brak \wedge Ncl \wedge \neg Sur \mid\!\sim ReacS$.

From (4) and (6), we derive $ReacL \wedge Wet \rhd Acc$. The cause of the accident is the conjunction of braking late and the road being wet. Now let us consider a few additional plausible nonmonotonic rules. Assume that long-delay braking alone,

[1] K may be omitted in practice. Def. 1 corresponds to a basic scenario already considered by von Wright [1]: The falsity of B_t agrees with the piece of general knowledge $K \mid\!\sim \neg B$ and after A_t takes place B_{t+k} becomes true, although normally if A_t does not happen, $\neg B$ would have persisted.

although it does not make accidents normal, at least makes them not abnormal ($ReacL \not\mid\!\sim Acc$ together with $ReacL \not\mid\!\sim \neg Acc$). Adding this assumption, we can derive $ReacL \blacktriangleright Acc$; i.e., the long-delay braking alone facilitated the accident (but the cause of the accident is still the conjunction of late braking and the road being wet). Assume now that late braking is abnormal ($\mid\!\sim \neg ReacL$), and remains so in the context of others braking, and being unable to change lane ($Brak \wedge Ncl \mid\!\sim \neg ReacL$). Then it follows from (7) that being surprised caused the late braking ($Sur \rhd ReacL$). For the purpose of further illustration, let us assume that accidents remain abnormal even when roads are wet ($Wet \mid\!\sim \neg Acc$). Then it is possible to derive than late braking alone caused the accident ($ReacL \rhd Acc$). Note now that $Sur \rhd ReacL$ together with $ReacL \rhd Acc$. Surprise caused the late braking that itself caused the accident. Does it follow by transitivity that $Sur \rhd Acc$, i.e., that surprise caused the accident? Not necessarily so, for $\rhd$ is not generally transitive. If, however, we are ready to accept that $ReacL \mid\!\sim Sur$, i.e., a late braking is usually diagnostic of a surprised driver, then it follows from the restricted transitivity property of $\rhd$ that the surprise caused the accident. Finally, suppose that that we add to the story that some other car C hit B. Then, the nonmonotonic approach yields a disjunctive causal ascription 'car hitting OR late braking' caused the accident. Only a more detailed report may lead the approach to privilege one of the disjuncts.

Discussion. This approach relies on the beliefs about the 'normal' states and courses of the world. Such beliefs are agent-dependent, which explains that different individuals may have different readings of events. Since the inference engine based on System P is very cautious, many of these normal states must be explicitly coded rather than derived. Causality ascription is localized, thanks to a lack of general transitivity, but also because only events that are explicitly mentioned in the story can be detected as causes. Exceptional events are favored as potential causes, which help discriminating causes; in fact, the approach only exhibits causes that are abnormal events. A notion of 'necessary condition' (or enabling condition) [12] can be defined to deal with normal events without which nothing would have happened. Finally, this approach does not embed the notion of intervention and thus cannot readily distinguish spurious correlation from causation. See nonetheless the *Graphical Models and Interventions* section for an extension of the approach into that direction (both the current approach and graphical models can be encoded in a possibilistic setting).

3.2 Trajectory-Based Preference Relations

This proposal [13][2] starts with the idea that counterfactuality involves the computation of two kinds of evolutions of the world, namely extrapolation [14] and update [15]. If we want to know whether $Sur_{(2)}$ (being surprised at time point 2) is a counterfactual cause of $Acc_{(3)}$, given a scenario Σ ($Brak_{(1)} \wedge Sur_{(2)} \wedge Ncl_{(2)} \wedge Wet_{(2)} \wedge Acc_{(3)}$), we need to (i) compute the most normal evolutions of the world

[2] For the sake of brevity, this novel approach is only sketched in this paper.

(called trajectories) that correspond to the scenario $Sur_{(2)}$ and $Acc_{(3)}$. This computation is called extrapolation, it is a process of completing initial beliefs sets stemming from observations by assuming minimal 'abnormalities' in the evolution of the world with respect to generic knowledge. In our example, the preferred trajectories satisfying Σ do satisfy $Sur_{(2)}$ and $Acc_{(3)}$ (since they are mentioned in Σ). (ii) Compute what would have happened to Acc_3 if $Sur_{(2)}$ had not been true. This is done by updating the temporal formula representing the scenario by the formula $\neg Sur_{(2)}$. At this step, update aims at capturing a minimal change w.r.t. the initial scenario. The update operator proposed in [13] is based on a distance between trajectories that take into account the time point of the change and normality. Here, the trajectories that satisfy $\neg Sur_{(2)}$ that are closest to the previous preferred trajectories until the time of the change and that are the most normal satisfy $\neg Acc_{(3)}$. Hence the surprise can be considered as counterfactually causing the accident. One may consider that not all counterfactual causes are important; the lack of selectivity of counterfactuality is tackled here by using normality. Choosing among the 'normal' counterfactual causes, the most abnormal ones in context, would further increase selectivity.

3.3 Norm-Based Approach

This approach [16], too, rests on the idea that norms are crucial for people to find causes of events: if the event is considered normal, its cause is the norm itself; if abnormal, its cause is traced back to the violation of a norm.[3]

Principle. Searching for the cause of an abnormal event E occurring at time t basically amounts to finding an agent who should, according to some norm, adopt behavior b at a time $t' < t$, and actually adopted another behavior b', such that E appears as a normal consequence of b' (in that sense, for example, the lack of liability insurance is a norm violation but cannot usually be considered the cause of an accident, because it arguably does not normally have an accident as a consequence). Another condition must be checked, namely that, at t', the agent had the possibility to have the normal behavior b; otherwise, b' is only a derived anomaly and the search must be pursued to find a primary anomaly, occurring earlier than t' and explaining the impossibility of the agent to have the behavior b at t'. Whenever this search fails, i.e., when the privilege conferred to an 'interventionist' kind of cause gives no result, and only in this case, we look for some non agentive abnormal circumstance that could explain E.

Norm-based reasoning is intrinsically non monotonic, as norms are rules that apply by default. For this reason, in this approach, the knowledge necessary to causal ascription is expressed in a reified first-order logic augmented with default rules (in the sense of R. Reiter); the fact that property P holds for agent A at time t is written $holds(P, A, t)$. A discrete and linear model of time is sufficient, as only what really happened is represented. Two modalities are introduced to

[3] The word 'norm' is taken here in the 'normal' rather than 'normative' sense; but as we expect agents to respect their duties, the normative is seen as a special case of the normal.

express norm violations: $should(P, A, t)$ and $able(P, A, t)$ standing for: at time t, A should (resp. has the ability to) achieve P.

Testing this approach in the domain of road accidents requires to gather all the literals of the form $should(P, A, t)$ that are relevant for this domain. To this end, we examined 73 car-crash reports, used as a training sample among the 160 reports in our possession; the remainder being left for validation purposes. For the running example of this paper, we only need a few of these literals: By wet weather, one should reduce one's speed; having had an accident at time t entails that one had at time $t - 1$ the duty of avoiding some obstacle; and having this duty and being unable to change lane amounts to have the duty to stop. This is written (→ is the material implication):

(1) $Wet \rightarrow should(reduced\_speed, A, t)$
(2) $holds(Acc, A, t) \rightarrow should(avoid\_obs, A, t - 1)$
(3) $should(avoid\_obs, A, t) \wedge \neg able(ch\_lane, A, t)$
$\rightarrow should(stop, A, t)$

Expressed in this language, the cause of an abnormal event (the 'primary anomaly' P_ano) obtains as:

(4) $should(F, A, t) \wedge able(F, A, t) \wedge \neg holds(F, A, t + 1) \rightarrow P\_ano(F, A, t + 1)$

I.e., if at t an agent A should do F and was able to do F, while at $t+1$, F failed to be done, this failure is the cause looked for. Similarly, a 'derived anomaly' D_ano is detected by the rule:

(5) $should(F, A, t) \wedge \neg able(F, A, t) \rightarrow D\_ano(F, A, t)$

Assume as a default that agents having a duty are generally able to comply with it. Exceptions to this default mostly correspond to cases where the situation allows to prove the impossibility of actions known to produce the desired effect.

Example. With the notations adopted in this paper, the example is written: $holds(Brak, B, 0)$, $holds(Sur, A, 1)$, $holds(Ncl, A, 2)$, $holds(Reac, A, 2)$, $holds(Acc, A, 3)$, Wet. Ncl (inability to change lane) translates as:

(6) $holds(Ncl, A, t) \rightarrow \neg able(ch\_lane, A, T)$

Expressing that surprise entails a late brake is written as:

(7) $holds(Sur, A, t - 1) \wedge holds(Reac, A, t) \rightarrow holds(ReacL, A, t)$

Whether late braking entails or not an accident depends on the ability of the driver to stop the vehicle, i.e.:

(8) $holds(ReacL, A, t) \rightarrow [holds(Acc, A, t + 1) \leftrightarrow \neg able(stop, A, t)]$

Rule (2) and fact $holds(Acc, A, 3)$ yield $should(avoid\_obs, A, 2)$; (6) gives $\neg able(ch\_lane, A, 2)$, hence (3) deduces $should(stop, A, 2)$. From (7) with premises $holds(Reac, A, 2)$ and $holds(Sur, A, 1)$ we get $holds(ReacL, A, 2)$. So (8) shows that something abnormal occurred: agent A should have stopped at time 2 but was unable to. According to (5), this is a derived anomaly, so the search for the cause of the accident must go on. The ability to stop, under the circumstances, is expressed by (9) $able(stop, A, t) \leftrightarrow (\neg Wet \vee holds(reduced\_speed,\ A, t))$, which gives $\neg holds(reduced\_\ speed, A, 2)$. (1) shows that $should(reduced\_speed, A, t)$ for any t. Without proof to the contrary, the default 'agents who should do something are generally able to do it' yields $able(reduced\_speed, A, 1)$, and (4) tells that we have a primary anomaly, i.e., a cause of the accident: 'at time 1, A was able to reduce speed; because of the wetness of the road, A should have done so, but the occurrence of the accident at time 3 shows that he was still driving too fast at time 2.'

Discussion. In traffic accident examples, the norm-based approach views norms as normative duties. To generalize this approach to domains where norms are only what is normal (as opposed to mandatory), it is necessary to organize these norms in a hierarchy, and to conjecture that the most specific violated norm will be perceived as the cause of an abnormal event. Testing this conjecture requires to gather a reasonably complete set of norms for the domain, which is a hard task. This was achieved in the domain of traffic accidents, and the validation process for this domain is underway. We intuitively determined the causes of the 160 accidents in the corpus, translated the gathered norms in Smodels [17], and implemented a system translating natural language sentences into the language of the norm-based approach. This system [16] agrees with the researchers' intuitions in 95% of the training sample and 85% of the validation sample. Behavioral experiments are underway to check whether these intuitions are shared by a majority of subjects.

4 Graphical Models and Interventions

Intervention is a critical route to causation. Ascribing causality becomes easier when experimenting, then observing the effects of the manipulation on the system. Such changes cannot be deduced from a joint probability nor possibility distribution, even fully specified on the variables describing the system. Graphical causal models help make explicit the assumptions needed by allowing inference from interventions as well as observations. A causal Bayesian network is a Bayesian network where directed arcs of the graph are interpreted as elementary causal relations between variables. When there is an influence relation between two variables, intervention allows to determine the causality relation between these variables. In this case, arcs between variables should follow the direction of the causal process. Pearl [2] proposed an approach for handling interventions using causal graphs based on a 'do' operator. Note that causal relations expressed by graphs only concern variables, not complex events. Causal Bayesian networks organize causal knowledge in terms of a few basic mechanisms, each involving a relatively small number of variables. Each intervention entails local change at the level of only one parents-child relation.

This section summarizes manipulation methods for handling interventions in possibilistic causal networks. Indeed graphical models are compatible both with a probabilistic and a possibilistic modeling of uncertainty. The possibilistic setting [18] is adopted here. It is more qualitative, and allows us to more easily relate graphical models to nonmonotonic approaches. In fact, the ‘do’ operator has been first proposed within Spohn’s ordinal conditional functions framework which has strong relationships with possibility theory. The parents-child relation at the level of each variable A_i is governed by a local possibility distribution $\Pi(A_i|U_{A_i})$ where U_{A_i} is the parents set of A_i. The joint possibility distribution is computed using the chain rule: $\pi(A_1, ..., A_n) = \Diamond_{i=1,...,n}\Pi(A_i|U_{A_i})$, where $\Diamond$ is either equal to min or product. An intervention forcing a variable A_i to take the value a_i is denoted $do(A_i = a_i)$ or $do(a_i)$. This intervention consists of making A_i true independently from all its other direct causes (i.e. parents). Graphically, this modification is represented by the deletion of links from U_{A_i} pointing into A_i. The resulting graph is said to be mutilated and we have:

$$\pi(\omega|do(A_i = a_i)) = \pi_{mut}(\omega|A_i = a_i) \ = \begin{cases} \Diamond_{a_j:A_j\neq A_i}\Pi(a_j|u_{A_j}) & \text{if } \omega[A_i] = a_i \\ 0 & \text{otherwise} \end{cases}$$

where $\omega[A_i] = a_i$ means that ω is consistent with $A_i = a_i$, and π_{mut} is the joint possibility distribution given by the mutilated graph. Another approach [19] consists in adding a new variable denoted DO_{A_i} as a parent node of A_i. DO_{A_i} takes value $do_{A_i-noact}$ when no intervention is observed, and value do_{a_i} when an intervention occurs, forcing A_i to take value a_i (a_i belonging to the domain of A_i). The resulting graph is called augmented. In [20], we showed that the better option to compute the effect of interventions is using augmented graphs, since it allows to reuse existing propagation algorithms without any change.

Example. Let us consider the possibilistic causal network given in Fig. 1. The variables $Brak$, Ncl, Sur, Wet and Acc are binary variables with a domain $\{0, 1\}$ and have the same meaning as in above examples. However, $Reac$ is a ternary variable taking its values in $\{ReacS, ReacL, NoReac\}$ where $NoReac$ means ‘A does not brake’, $ReacS$ means ‘A brakes as soon as B brakes’ and $ReacL$ means ‘A brakes later after B braked’. For simplicity’s sake, we assume only three levels of normality: 1 (i.e. fully plausible) $> \beta > \alpha >$ 0 (i.e. impossible). Prior local possibility distributions are assumed to be: $\Pi(Sur = 0) = 1 > \Pi(Sur = 1) = \alpha$, which encodes rule (5) of the introduction; $\Pi(Brak = 0) = 1 > \Pi(Brak = 1) = \beta > \alpha$, $\Pi(Ncl = 0) = 1 > \Pi(Ncl = 1) = \alpha$, $\Pi(Wet = 0) = 1 > \Pi(Wet = 1) = \alpha$, which respectively express that normally: ’B does not brake’, ‘ there is no possibility to change lane’, and ’the road is not wet’. The local possibility distribution for $Reac$ (i.e. $\Lambda_1 = \Pi(Reac|Sur, Brak, Ncl)$) is given by (9):

$$\Lambda_1 = \begin{cases} 1 \text{ if } (Reac = NoReac \text{ and } (Brak = 0 \text{ or } Ncl = 0)) \\ \quad \text{or } (Reac = ReacS \text{ and } Sur = 0 \text{ and } Brak = 1 \text{ and } Ncl = 1) \\ \quad \text{or } (Reac = ReacL \text{ and } Sur = 1 \text{ and } Brak = 1 \text{ and } Ncl = 1) \\ \alpha \ otherwise \end{cases}$$

Rules (7) and (8) of the introduction are encoded. Indeed, for instance regarding rule (8) we have $\Pi(Reac = ReacL \mid Sur = 1, Ncl = 1, Brak = 1) = 1 > \Pi(Reac = ReacS|Sur = 1, Ncl = 1, Brak = 1) = \alpha$ (and $\Pi(Reac = ReacL \mid Sur = 1, Ncl = 1, Brak = 1) = 1 > \Pi(Reac = NoReac|Sur = 1,$

$Ncl = 1, Brak = 1) = \alpha$, which means that if 'when the driver B brakes, A is surprised and there is no a possibility to change lane' then it is more plausible that the driver A brakes with a longer delay than he does not brake or he brakes shortly after B brakes. Lastly, the local possibility distribution at the level of Acc (i.e. $\Lambda_2 = \Pi(Acc|Wet, Reac)$) is given by (10):

$$\Lambda_2 = \begin{cases} 1 \text{ if } (Acc = 1 \text{ and } Wet = 1 \text{ and } Reac = ReacL) \\ \quad\quad \text{or } (Acc = 0 \text{ and } (Wet = 0 \text{ or } Reac = NoReac \text{ or } Reac = ReacS)) \\ \alpha \ otherwise \end{cases}$$

Again, rule (6) is encoded since $\Pi(Acc = 1 \mid Wet = 1 \text{ and } Reac = ReacL) > \Pi(Acc = 1 \mid Wet = 0 \text{ or } Reac \neq ReacL)$. Note that rule (4) is not explicitly represented but is only derived. Indeed, after propagation of weights we obtain $\Pi(Acc = 0) = 1 > \Pi(Acc = 1) = \alpha$ which means that accidents are abnormal.

For binary variables, possibilistic graphical models can encode causality relations as defined by nonmonotonic logic approaches. $E \mid\!\sim F$ is interpreted by $\Pi(E \wedge F) > \Pi(E \wedge \neg F)$. This relation satisfies rational monotony in addition to System P, providing more causal relations. Besides, whereas only reported events can be causes as per Definition 1, unreported but strongly plausible events can be causes in the possibilistic frameworks. Lastly, graphical models provide a computational tool for causality ascriptions in presence of interventions. Recall that $\Pi(Acc = 0) = 1 > \Pi(Acc = 1)$, i.e. $Acc = 1$ is rejected in the initial context. Let us consider an external factor (say, an animal crossing the road) forcing the variable $Reac$ to take value $ReacL$. This intervention $do(Reac = ReacL)$ can be represented by mutilating or by augmenting the graph. Assume moreover that the road is wet. After computation, we have $\Pi(Acc = 1|do(Reac = ReacL), Wet = 1) = 1 > \Pi(Acc = 0|do(Reac = ReacL), Wet = 1)$. Namely, after intervention $do(Reac = ReacL)$ and observation $Wet = 1$, event $Acc = 1$ becomes accepted. We conclude that $do(Reac = ReacL)$ and $Wet = 1$ caused $Acc = 1$.

Discussion. Graphical models offer a natural representation of causal relations between elementary events (e.g. variables), thanks to the 'do' operator that models interventions. They can be viewed as complementing or extending nonmonotonic approaches. Indeed, Definition 1 can be naturally extended when reported events include interventions (as illustrated above). A graphical model goes beyond System P without recovering transitivity. It can be used to discriminate between possible causes by considering the most plausible ones, and allows causality ascription in presence of observations and interventions.

5 Theory of Explanatory Coherence (TEC)

Thagard's theory of explanatory coherence [21] and its connectionist implementation (ECHO) view causal ascriptions as attempts to maximize explanatory coherence between propositions. Although this model did not originate from the AI knowledge representation community, it addresses a similar concern to the other models with have reviewed, and it is as much implementable. In the accident example, maximizing coherence would lead to accept the most plausible hypotheses that explain the accident and reject the alternative hypotheses. If one proposition explains another, then there is a positive constraint between them. Negative constraints result from events that prevent or are inconsistent with other events. Maximizing coherence is generally considered to be computationally intractable. Nevertheless, good approximation algorithms are available, in particular connectionist algorithms such as ECHO. ECHO creates a network of units with explanatory and inhibitory non directional links and then makes inference by spreading activation through the network until all activations have reached stable values. Note that links can be excitatory or inhibitory and units can be positively or negatively activated. When units have settled, the acceptation and rejection of hypotheses depend on whether final activation is positive or negative. Some units can be given priority by linking them positively with a special unit whose activation is kept at 1. A coherence problem is defined as follows [22]. Let E be a finite set of elements $\{e_i\}$ and C be a set of constraints on E understood as a set $\{(e_i, e_j)\}$ of pairs of elements of E. C divides into C^+, the positive constraints on E, and C^-, the negative constraints on E. With each constraint is associated a number w, which is the weight (strength) of the constraint. The problem is to partition E into two sets, A and R, in a way that maximizes compliance with the following two coherence conditions:

1. if (e_i, e_j) is in C^+, then e_i is in A iff e_j is in A;
2. if (e_i, e_j) is in C^-, then e_i is in A iff e_j is in R.

Let W be the sum of the weights of the satisfied constraints. The coherence problem is then to partition E into A and R in a way that maximizes W. Let E, C, C^+, and C^- as defined above. ECHO runs as follows:

1. For every e_i of E, construct a unit u_i, a node in a network of units U;
2. For every positive (negative) constraint in C^+ (C^-) on elements e_i and e_j, construct an excitatory (inhibitory) link between the corresponding units u_i and u_j affected with the same positive (negative) weight.
3. Assign each unit u_i an equal initial activation. Update activation of all the units in parallel given current activations and the weights on links [23]
4. When units have settled, hypotheses acceptation and rejection depend on the sign of their final activation. Some units can be given priority by linking them positively with a special unit whose activation is kept at 1.

Example. In Figure 2 each node represents a variable. The three nodes on the left and the Wet node correspond to variables with priority; in this case, initial conditions at the beginning of the accident process. Dotted lines represent

Table 1. Final TEC activation values

	$Brak$	Wet	Sur	Ncl	$ReacL$	$ReacS$	Acc
$Brak$, Wet, Sur, Ncl initially set to 1	-.22	.72	.68	-.22	.23	-.70	.58
Only $Brak$ initially set to 1	.68	-.64	-.65	-.43	.46	.73	-.56

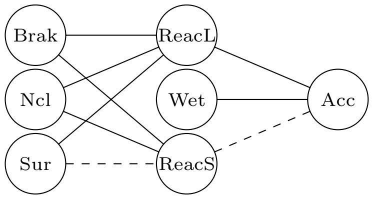

Fig. 2. Accident Example Network in ECHO

inhibitory links. Table 1 shows final activation values of the variables after the units have settled. When initial conditions are $Brak = Wet = Sur = Ncl = 1$ (*All* in Table 1), the hypothesis that the accident occurs is accepted ($Acc = .58$). Wet, Sur, and $ReacL$ are accepted, all other hypotheses are rejected. The most activated causes are Wet and Sur. When the only initial condition is $Brak$, the accident hypothesis is rejected: $Brak$ alone is not a sufficient condition for Acc.

Discussion. ECHO establishes an ordering between accepted causes, their final activation representing their causal power. It is of particular interest because previous experimental studies [9] have suggested that human distinction between facilitation and genuine causality is based on the strength of the relation between events. Inference in connectionist models like ECHO is not monotonic, not transitive, and can be forward or backward. Although the only central notion is coherence in TEC, questions of abnormality, temporality and intervention can be introduced in order to compute a more powerful causal inference. ECHO can be translated in Pearl's probabilistic networks [24], and has been used in diverse psychological domains in addition to the computation of causation [25].

6 General Discussion

Causality has always been the matter of hot debates. There exists no consensus about its very nature: is it a means by which the human mind makes sense of the world, or an objective property of the world? Is causation intrinsically deterministic, and only our ignorance makes it admissible to approach it by methods devoted to handle uncertain knowledge; or on the opposite, is its relation to uncertainty fundamental? The models we described are agnostic with respect to such debates—and this should be no surprise to the reader, as this paper is not about causation per se, but about how, under practical circumstances, agents prune among a huge number of potential causal factors.

Although the different models start with the same core of variables and pieces of knowledge (1–8), they rely on representation frameworks of different expressive power, and they may exploit additional pieces of knowledge that are not assumed to be available to other models. For example, the norm-based approach relies on a vast set of norms extracted from driving regulations, while for instance the graphical approach relies on probabilistic or possibilistic information. Although this introduces some heterogeneity in the treatment of the example,

Table 2. Comparison of the models, synthesis

	Structural Eq.	TEC	Norms	Trajectories	Nonmon. Consequences	Graphs
Selectivity	No	Yes	Yes	Yes	Yes	Yes
Abnormality	No	No	Yes	Yes	Yes	Yes
Temporality	No	No	Yes	Yes	Yes	Yes
Cause Present	No	No	No	No	Yes	Yes
Intervention	No	No	No	Yes	No	Yes
Agentivity	No	No	Yes	No	No	No
Backward Cause	Yes	Yes	No	No	No	No

this heterogeneity is irreducible if we want to compare a large span of approaches while respecting their specific modeling strategy.

Table 2 sums up some features of each model.[4] TEC and the structural equation model do not make explicit the temporal relation between the factors they deal with, e.g. braking occurs before stopping (**temporality**). All other models make this temporal link explicit. Accordingly, all models but TEC and the structural equation model assume that effects cannot precede their causes (**backward causation**). The structural equation approach is the least selective of all (but see caveat in section 2), in that sense that it delivers a set of factors that all reasonably have some relevant causal connection to the effect under consideration. All other models strive to select a smaller set of factors, apparently emulating human judgments (**selectivity**). These models privilege different aspects of information to select one event as the main cause. First, all these selective models make explicit the contrast between normal and abnormal states of affairs, to orient the search of causes of an abnormal event towards factors that make a normal course of events become abnormal (**abnormality**). Then, some of them (nonmonotonic consequences, graphical models) consider that the cause is bound to belong to the set of facts given in the description (**cause present**), whereas the other models are allowed to elicit causes among implicit elements derived from these facts, or including as background knowledge in the course of the modelling. Besides, one model (norm-based) privileges as causes events that are under the control of agents (**agentivity**). Finally, some models (trajectories, graphical model) can support explicit intervention-like manipulations, where a variable can be forced to take some value, regardless of what its normal value would be given the values of the other variables (**intervention**).

In addition to the criteria summarized in Table 2, let us note that only the structural equation model is deterministic, in the sense that there is (commonly) no uncertainty in the structural equations relating the variables representing the micro-universe under consideration. This could be seen as a guarantee of accurateness, as far as the description of the micro-universe is reasonably complete.

[4] **Transitivity** is not a built-in characteristic in any model we have considered. Depending on the specific setting of some parameters, though, some of them may take causation to be transitive. **Comptutational tractability** is not a truly discriminative criterion here either. All the formalisms underlying the approaches we have reviewed have already been implemented. Moreover, the formal complexity of all these frameworks has already been studied; and in any case, the treatment of traffic accident reports is unlikely to lead to any significant combinatorial explosion.

However, such accurateness would come with a price. Leaving aside the computational cost of ascription itself, a deterministic model is more costly in terms of the acquisition of information that is necessary prior to making any ascription. Furthermore, default, incomplete knowledge is arguably less unrealistic as a model of the kind of knowledge human agents bring to a causal ascription task.

Among the topics that we have not covered, the role of argumentation in causal ascription is worth mentioning. Argumentation is a dynamical process where arguments interact to assess a given claim (here, a causal claim), and processing of causal arguments requires a particular argumentation theory [26]. Agents may argue about where causation takes place in a sequence of events; they may use a weaker notion of causality than, e.g., Def. 2. But agents may also use argumentation in a self-serving way: in the case of a traffic accident, they may attempt to present events in a favorable way; to produce a 'biased description,' that remains respectful of the essential facts, but triggers inferences to conclusions that are in favor of the arguer. For example, one argumentation technique consists in suggesting a causal link between two facts, even if the causation is at best debatable. One typical case is to present the violation of a 'strong' norm as a consequence caused by the adversary's violation of a 'weak' norm, as in the example: 'At the stop sign, the driver on the main road delayed in entering the intersection; I proceeded.' The author wishes to convey that it is normal to overstep a stop sign, in case the vehicle having priority is hesitating. Identifying argumentative strategies may help to get a better understanding of reports, by detecting understatements, and reconstructing what is not explicitly said. In addition, further work will have to compare approaches in terms of their handling of preventative (negative) causation, and of their syntax sensitivity.

Acknowledgments

This research was supported by grant NT05-3-44479 from Agence Nationale de la Recherche. We thank L. Amgoud, D.J. Hilton, and F. Lévy for useful comments.

References

1. von Wright, G.H.: Norm and Action: A Logical Enquiry. Routledge, London (1963)
2. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
3. Halpern, J., Pearl, J.: Causes and explanations: A structural-model approach — part 1: Causes. British Journal for the Philosophy of Science 56, 843–887 (2005)
4. Hall, N.: Structural equations and causation. Philosophical Studies 132, 109–136 (2007)
5. Hart, H.L.A., Honoré, T.: Causation in the law. Oxford University Press, Oxford (1985)
6. Hilton, D.J., Slugoski, B.R.: Knowledge-based causal attribution: The abnormal conditions focus model. Psychological Review 93, 75–88 (1986)
7. Giunchiglia, E., Lee, J., McCain, N., Lifschitz, V., Turner, H.: Non-monotonic causal theories. Artificial Intelligence 153, 49–104 (2004)

8. McCain, N., Turner, H.: A causal theory of ramifications and qualifications. In: Proc. IJCAI 1995, vol. 95, pp. 1978–1984 (1995)
9. Bonnefon, J.F., Da Silva Neves, R.M., Dubois, D., Prade, H.: Background default knowledge and causality ascriptions. In: Proc. ECAI 2006, pp. 11–15 (2006)
10. Bonnefon, J.F., Da Silva Neves, R.M., Dubois, D., Prade, H.: Predicting causality ascriptions from background knowledge: Model and experimental validation. International Journal of Approximate Reasoning 48, 752–765 (2008)
11. Kraus, S., Lehman, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence 44, 167–207 (1990)
12. Prade, H.: Responsibility judgments: Towards a formalization. In: Proc. IPMU 2008, Malaga, June 22–27 (2008)
13. Dupin de Saint-Cyr, F.: Scenario update applied to causal reasoning. In: Proc. of KR 2008 (2008)
14. Dupin de Saint-Cyr, F., Lang, J.: Belief extrapolation (or how to reason about observations and unpredicted change). In: Proc. KR 2002 (2002)
15. Katsuno, H., Mendelzon, A.: On the difference between updating a knowledge base and revising it. In: Proc. KR 1991, pp. 387–394 (1991)
16. Kayser, D., Nouioua, F.: About norms and causes. International Journal on Artificial Intelligence Tools 1–2, 7–23 (2005)
17. Syrjaänen, T., Niemelä, I.: The Smodels systems. In: Eiter, T., Faber, W., Truszczyński, M. (eds.) LPNMR 2001. LNCS (LNAI), vol. 2173, pp. 434–438. Springer, Heidelberg (2001)
18. Dubois, D., Prade, H.: Possibility theory: Qualitative and quantitative aspects. In: Gabbay, D.M., Smets, P. (eds.) Quantified Representation of Uncertainty and Imprecision, pp. 169–226. Kluwer, Dordrecht (1998)
19. Pearl, J.: Comment: Graphical models, causality and intervention. Statistical Sciences 8 (1993)
20. Benferhat, S., Smaoui, S.: Possibilistic causal networks for handling interventions: A new propagation algorithm. In: Proc. AAAI 2007, pp. 373–378 (2007)
21. Thagard, P.: Explanatory coherence. Behavioral and Brain Sciences 12, 435–467 (1989)
22. Thagard, P., Verbeurgt, K.: Coherence as constraint satisfaction. Cognitive Science 22, 1–24 (1998)
23. McClelland, J., Rumelhart, D.: Explorations in parallel distributed processing. MIT Press, Cambridge (1989)
24. Thagard, P.: Probabilistic networks and explanatory coherence. Cognitive Science Quarterly 1, 91–114 (2000)
25. Read, S., Marcus-Newhall, A.: The role of explanatory coherence in the construction of social explanations. Journal of Personality and Social Psychology 65, 429–447 (1993)
26. Amgoud, L., Prade, H.: Arguing about potential causal relations. In: IAF 2007 (2007)