In. B. Nebel, C. Rich, and W. Swartout (Eds.), *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR`92)*, pp. 661-672, 1992.

Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions.

Moisés Goldszmidt < *moises@cs.ucla.edu* >
Judea Pearl < *judea@cs.ucla.edu* >
Cognitive Systems Laboratory, Computer Science Department,
University of California, Los Angeles, CA 90024

Abstract

We describe a ranked-model semantics for if-then rules admitting exceptions, which provides a coherent framework for many facets of evidential and causal reasoning. Rule priorities are automatically extracted form the knowledge base to facilitate the construction and retraction of plausible beliefs. To represent causation, the formalism incorporates the principle of *Markov shielding* which imposes a stratified set of independence constraints on rankings of interpretations. We show how this formalism resolves some classical problems associated with specificity, prediction and abduction, and how it offers a natural way of unifying belief revision, belief update, and reasoning about actions.

1 Introduction

This paper is a culmination of several attempts to give conditional knowledge bases (with exceptions) empirical semantics in terms of infinitesimal probabilities, to be regarded as qualitative abstractions of an agent's experience. This semantics can be described in terms of rankings on models, where higher ranked models stand for more surprising (or less likely) situations. At the heart of this formulation is the concept of *default priorities*, namely, a natural ordering of the conditional sentences which can be derived automatically from the knowledge base and which can be used to answer queries without computing explicit rankings of worlds or formulas. The result is a model-theoretic account of plausible beliefs that, as in classical logic, are qualitative and deductively closed and, as in probability, are subject to retraction and to varying degrees of firmness.

The first part of this paper (Section 2) gives a brief summary of this rank-based semantics and describes a query-answering system called system-Z^+ which embodies this semantics in effective computational procedures. The main thrust of the paper (Section 3) is the introduction, within the basic framework of ranking systems, of a simple mechanism called *stratification* for the representation of causal relationships, actions, and changes.

The lack of a mechanism for distinguishing causal relationships from other kinds of associations has been a serious deficiency in most nonmonotonic systems [28], the classical illustration of which is given by the now-famous Yale Shooting Problem (YSP) [20]. In its simplified version, the YSP builds the expectation that if a gun is loaded at time t_0 and Fred is shot with the gun at time t_1, Fred should be dead at time t_2, despite the normal tendency of *being alive* to persist. Many formulations — including circumscription [26], default logic [34], rational closure [23], and conditional entailment [13] — reveal an alternative, perfectly symmetrical version of reality, whereby somehow the gun got unloaded and Fred is alive at time t_2.

The inclination to choose the scenario in which Fred dies is grounded in notions of directionality and asymmetry that are particular to causal relationships. In this paper we show that these notions can be derived from one fundamental principle, *Markov shielding*, which can be embodied naturally in preferential model semantics using the device of stratified rankings. Informally, the principle can be stated as follows:

- Knowing the set of causes for a given effect renders the effect independent of all prior events.

In the YSP, given the state of the gun at time t_1, the effect of the shooting can be predicted with total disregard for the gun's previous history.

We propose a probabilistically motivated, ranked-model semantics for rules of the form "typically, if cause_1 and ... and cause_n, then effect", which incorporates the above principle under the assumption that "causes" precede their "effects". As a by-product, our semantics exhibits another feature characteristic of causal organizations: *modularity*. Informally,

- Adding rules that predict future events cannot invalidate beliefs concerning previous events.

This is analogous to a phenomena we normally associate with causal mechanisms such as logical gates in electrical circuits, where connecting the inputs of a new gate to an existing circuit does not alter the circuit's behavior [7].

Although several remedies were proposed for the YSP within conventional nonmonotonic formalisms [35, 13, 37, 2, 24], the formalism we explore in this paper seeks to uncover remedies systematically from basic probabilistic principles [29, pp. 509–516]. We show that incorporating such principles in the qualitative context of world ranking yields useful results on several frontiers. In prediction tasks (such as the YSP), our formalism prunes the undesirable scenarios, without the strong commitment displayed by *chronological minimization* [35] and without the addition of *external* causal operators to the conditional interpretation of the rules [13]. In abduction tasks (such as when Fred is seen alive at t_2), our formalism yields plausible explanations for the facts observed (e.g., similar to [37], the gun must have been unloaded sometime before the shooting at t_1). This suggests that the principle of Markov shielding, by being grounded in probability theory (hence in empirical reality), can provide a coherent framework for the many facets of causation found in commonsense reasoning. Moreover, given the connection formed among causation, defaults, and probability, we can now ask not merely how to reason with a given set of causal assertions but also whether those assertions are compatible with a given stream of observations.

In the last part of this paper (Section 4) we demonstrate how rank-based systems can embody and unify the theories of belief revision [1] and belief updating [21], two theories of belief change that have been developed independently of research in default and causal reasoning. Basically, theories of belief change seek general principles for constraining the process by which a rational agent ought to incorporate a new piece of information ϕ into an existing set of beliefs ψ, regardless of how the two are represented and manipulated. Belief revision deals with new information obtained through new observations in a static world, while belief update deals with tracing changes in an evolving world, such as that subjected to the external influence of actions.

We show that system-Z^+ offers a natural embodiment of the principles of belief revision as formulated by Alchourrón, Gärdenfors and Makinson (AGM) [1], with the additional features of enabling the absorption of new conditional sentences and the verification of counterfactual sentences and nested conditionals. We then show that the addition of stratification to system-Z^+, by virtue of representing actions and causation, also provides the necessary machinery for embodying belief updates consistent with the principles proposed by Katsuno and Mendelzon (KM) [21].

2 Rankings and System-Z^+: Review

We assume throughout a finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$ of atomic propositions. The greek letters $\varphi, \psi, \sigma, \varphi$ will denote well–formed formulas (wff) built from the elements in $\mathcal{X}$. A *possible world* ω is a truth assignment to the propositions in $\mathcal{X}$. The satisfaction of a wff φ by an world ω is defined as usual and denoted by $\omega \models \varphi$. If ω satisfies φ then we say that ω is a model for φ.

A *defeasible conditional* or *default* is a formula "$\varphi \xrightarrow{\delta} \psi$", where φ and ψ are wffs (built from $\mathcal{X}$), "$\rightarrow$" is a new binary connective, and δ is a non-negative integer. The intended reading of $\varphi \xrightarrow{\delta} \psi$ is "typically, if φ then *expect* ψ (with strength δ)".[1] The connective "$\rightarrow$" imposes preferences among the possible worlds ω, requiring that if $\varphi \rightarrow \psi$, then ψ must be true in all the most preferred models for φ. In order to represent these preferences, we introduce ranking functions on the set Ω of possible worlds.

Definition 1 (Rankings) A ranking function κ is an assignment of non-negative integers to the elements in Ω, such that $\kappa(\omega) = 0$ for at least one $\omega \in \Omega$. We extend this definition to induce rankings on wffs:

$$\kappa(\varphi) = \begin{cases} \min_{\omega \models \varphi} \kappa(\omega) & \text{if } \varphi \text{ is satisfiable} \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

Similarly, for a pair of wffs φ and ψ we define the conditional ranking $\kappa(\psi|\varphi)$ as

$$\kappa(\psi|\varphi) = \begin{cases} \kappa(\psi \wedge \varphi) - \kappa(\varphi) & \text{if } \kappa(\varphi) \neq \infty \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

Preferences are associated with lower κ, and *surprise* or *abnormality* with higher κ. Thus, $\kappa(\psi) < \kappa(\varphi)$ if ψ is preferred to φ in κ, or equivalently, if φ is more abnormal (surprising) than ψ in κ. Intuitively, $\kappa(\psi|\varphi)$ stands for the degree of *surprise* or *abnormality* associated with finding ψ to be true, given that we already know φ. The inequality $\kappa(\neg\psi|\varphi) > \delta$ means that, given φ it would be surprising (i.e., abnormal) by at least $\delta + 1$ ranks to find $\neg\psi$, and it is equivalent to $\kappa(\psi \wedge \varphi) + \delta < \kappa(\neg\psi \wedge \varphi)$ which is precisely the constraint on worlds we attribute to $\varphi \xrightarrow{\delta} \psi$.

Definition 2 (Consistency) A ranking κ is said to be **admissible** relative to a given Δ, iff

$$\kappa(\varphi_i \wedge \psi_i) + \delta_i < \kappa(\varphi_i \wedge \neg\psi_i) \quad (3)$$

(equivalently $\kappa(\neg\psi_i|\varphi_i) > \delta_i$) for every rule $\varphi_i \xrightarrow{\delta_i} \psi_i \in \Delta$. A set Δ is **consistent** iff there exists an admissible ranking κ relative to Δ.

[1] The special case of $\delta = \infty$ corresponds to a *strict* conditional, to be denoted by $\Rightarrow$.

Consistency can be decided in $O(|\Delta|^2)$ satisfiability tests on the *material counterparts*[2] of the defaults in Δ, and it is independent of the δ-values assigned to the rules in Δ [15]. Eq. 3 echoes the usual interpretation of *defaults*, according to which ψ holds in all *minimal* models for φ. In our case, minimality is reflected in having the lowest rank. If we say that ω *falsifies* or *violates* a rule $\varphi \xrightarrow{\delta} \psi$ whenever $\omega \models \varphi \wedge \neg\psi$, the parameter δ can be interpreted as the minimal degree of surprise (or abnormality) associated with finding the rule $\varphi \xrightarrow{\delta} \psi$ violated, given that we know φ. In probabilistic terms, consistency guarantees that for every $\varepsilon > 0$, there exists a probability distribution P such that if $\varphi_i \xrightarrow{\delta_i} \psi_i \in \Delta$, then $P(\psi_i|\varphi_i) \geq 1 - c\varepsilon^{\delta_i}$ (see [17]).

2.1 The most normal ranking: κ^+

Given a set Δ, each admissible ranking κ induces a consequence relation $\mid\!\sim_\kappa$, where $\phi \mid\!\sim_\kappa \sigma$ iff $\kappa(\sigma \wedge \phi) < \kappa(\neg\sigma \wedge \phi)$. A straightforward way to declare σ as a plausible conclusion of Δ given ϕ would be to require $\phi \mid\!\sim_\kappa \sigma$ in all κ admissible with Δ. This leads to an entailment relation called ε-semantics [29], 0-entailment [31], and r-entailment [23], which is recognized as being too conservative. The approach we take here, following [31, 15, 23], is to select a distinguished admissible ranking, in our case κ^+, and declare σ as a plausible conclusion of Δ given ϕ, written $\phi \mid\!\sim_+ \gamma$, iff $\kappa^+(\phi \wedge \sigma) < \kappa^+(\phi \wedge \neg\sigma)$.[3] The distinguished ranking κ^+ assigns to each world the lowest possible rank permitted by the admissibility constraints of Eq. 3 (Def. 2), thus reflecting the assumption that, unless we are forced to do otherwise, each world is considered as normal (likely) as possible.

Definition 3 (The ranking κ^+) Let $\Delta = \{r_i \mid r_i = \varphi_i \xrightarrow{\delta_i} \psi_i\}$ be a consistent set of rules. κ^+ is defined as an admissible ranking function that is minimal in the following sense: Any other admissible ranking function must assign a higher ranking to at least one world and a lower ranking to none.

Theorem 4 ([17]) *Any consistent Δ has a unique minimal ranking κ^+ given by*

$$\kappa^+(\omega) = \begin{cases} 0 & \text{if } \omega \text{ does not falsify any rule in } \Delta, \\ \max_{\omega \models \varphi_i \wedge \neg\psi_i}[Z^+(r_i)] + 1 & \text{otherwise,} \end{cases} \quad (4)$$

*where $Z^+(r_i)$ is a set of integers defined on rules (*priorities*) which can be computed from Δ.*

Thus, the default rule priorities Z^+ constitute an economical way of encoding the ranking κ^+, linear in the size of Δ, from which the κ^+ of any world can be computed according to Eq. 4. In [17] we present an effective procedure, Procedure Z_rank, for computing Z^+, as well as answering queries. In the special case of a *flat* Δ, that is all δ's $= 0$, the procedure is as follows: We first identify all rules $r_i : \varphi_i \rightarrow \psi_i$ in Δ for which the formula

$$\varphi_i \wedge \psi_i \bigwedge_{j \neq i, r_j \in \Delta} \varphi_j \supset \psi_j \quad (5)$$

is satisfiable. Next we assign to these defaults priority $Z^+ = 0$, remove them from Δ, and repeat the process, assigning to the next set of defaults the priority $Z^+ = 1$, then $Z^+ = 2, \ldots$ and so on. Once Z^+ is known, the rank κ^+ of any wff ϕ is given by $\kappa^+(\phi) =$ minimum i such that

$$\phi \bigwedge_{j: Z^+(r_j) \geq i} \varphi_j \supset \psi_j \quad (6)$$

is satisfiable

Theorem 5 ([17]) *Given a consistent Δ, the computation of the Z^+ priorities requires $O(|\Delta|^2 \times \log|\Delta|)$ satisfiability tests. Moreover, given the Z^+ priorities, determining the ranking κ^+ of a wff ψ and the strength δ with which an arbitrary query σ is confirmed, given the information ϕ, that is $\phi \mid\!\sim^\delta_+ \sigma$, requires $O(\log|\Delta|)$ satisfiability tests.*

Another important result implied by Eqs. 5 and 6 gives a method of constructing a propositional theory $Th(\phi)$ that implies all the conclusions γ that plausibly follow from a given evidence ϕ, i.e., $\phi \mid\!\sim_+ \gamma$. Such a theory is given by the formula

$$Th(\phi) = \bigwedge_{i: Z^+(r_i) \geq \kappa^+(\phi)} \varphi_i \supset \psi_i \quad (7)$$

Clearly, if the rules in Δ are of Horn form, computing the priority ranking Z^+, κ^+ of a given ψ, and deciding the plausibility δ of queries $(\phi \mid\!\sim^\delta_+ \sigma)$ can be done in polynomial time [9]. The resulting system for default reasoning based on κ^+ and Z^+ is called system-Z^+ [15, 17].

System-Z^+ can also be used to reason with *soft* evidence or imprecise observations such as when the context ϕ of a query is not given with absolute certainty, and all we have is a testimony saying that "ϕ is supported to a degree n." In [17] we establish two strategies processing such reports. The first strategy, named J-conditionalization, is based on *Jeffrey's Rule of Conditioning* [30]. It interprets the report as specifying that "all things considered," the new degree of disbelief for $\neg\phi$ should be $\kappa'(\neg\phi) = n$. The second strategy, named L-conditionalization, is based on the *virtual evidence* proposal described in [29]. It interprets the report as specifying the desired *shift* in the degree of belief in ϕ, as warranted by that report alone and

[2] The material counterpart of $\varphi \xrightarrow{\delta} \psi$ is the wff $\varphi \supset \psi$.

[3] If we are concerned with the strength δ with which the conclusion is endorsed, then $\phi \mid\!\sim^\delta_+ \gamma$ iff $\kappa^+(\phi \wedge \sigma) + \delta < \kappa^+(\phi \wedge \neg\sigma)$.

"nothing else considered". Both interpretations yield semi-tractable procedures (i.e., polynomial for Horn theories) for assessing the plausibility of σ, free from the computational difficulties that plague most non-monotonic systems.

Section 4.1 demonstrates how the computational procedures of system-Z^+ can be employed in the context of belief revision. Next we strengthen the admissibility condition with an additional requirement which gives a causal character to the defaults in Δ.

3 Stratified Rankings

Let $c_1, \ldots, c_m$ and e be literals over the elements of $\mathcal{X}$. A *rule* is defined as the default $c_1 \wedge \ldots \wedge c_m \rightarrow e$,[4] where the conjunction "$c_1 \wedge \ldots \wedge c_m$" is called the antecedent of the rule and "e" its consequent.[5]

Given $\mathcal{X}$ and a set Δ of rules, the *underlying characteristic graph* for $\langle \mathcal{X}, \Delta \rangle$, is the directed graph $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ such that there is a node v_i for each $x_i \in \mathcal{X}$, and there is a directed edge from v_i to v_j iff there is a rule R in Δ where x_i (or $\neg x_i$) is part of the antecedent of R, and x_j (or $\neg x_j$) is the consequent of R. We say that Δ is a *causal network* (or *network* for short) if $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ is acyclic (i.e., $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$ is a DAG). If $v_r, \ldots, v_s$ are the parents of v_t in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$, then the set $\{x_r, \ldots, x_s\}$ is called the *parent set* of x_t and the set $\{x_r, \ldots, x_s\} \cup \{x_t\}$ is called a *family*. Intuitively, the parent set of an event e represents all the known causes for e. A network Δ induces a strict partial order "$\prec$" on the elements of $\mathcal{X}$ where $x_i \prec x_j$ iff there is a directed path from v_i to v_j in $\Gamma_{\langle \mathcal{X}, \Delta \rangle}$. We will use $\mathcal{O}(\mathcal{X})$ to denote any total order on the elements of $\mathcal{X}$ satisfying $\prec$.[6] Intuitively, $\prec$ represents a natural order on events where causes precede their effects.

Definition 6 (Stratified Rankings.) Given a network Δ, an admissible ranking κ, and an ordering $\mathcal{O}(\mathcal{X})$; let X_i $(1 \leq i \leq n)$ denote a literal variable taking values from $\{x_i, \neg x_i\}$, and let Par_{X_i} denote the conjunction $X_r \wedge \ldots \wedge X_s$ where $\{X_r, \ldots, X_s\}$ is the parent set of x_i. We say that κ is **stratified** for Δ under $\mathcal{O}(\mathcal{X})$, if for $2 \leq i \leq n$, and for any instantiation of the variables $X_1, \ldots, X_i$, we have

$$\kappa(X_i | X_{i-1} \wedge \ldots \wedge X_1) = \kappa(X_i | Par_{X_i}) \qquad (8)$$

Eq. 8 says that in a stratified ranking the degree of (ab)normality of an event x_i given all its prior events must be equal to the degree of (ab)normality of x_i given just the set of events constituting its parent set. This condition of stratification is closely related to the Markovian independence conditions embodied in Bayes Networks (BN) [29]. A BN is a pair (D, P) where D is a DAG and P is a probability distribution. Each node v_i in D corresponds to a variable X_i in P, and P decomposes into the product:

$$P(X_n, \ldots, X_1) = \prod_{i=1}^{i=n} P(X_i | Par_{X_i}) \qquad (9)$$

which, similarly to Eq. 8, incorporates the assumption that the parent set of any given variable X_i renders X_i probabilistically independent of all its predecessors (in the given ordering). Causal networks can in fact be regarded as an order of magnitude abstraction of BN's, where exact numerical probabilities are replaced by integer-valued levels of surprise (κ), addition is replaced by min, and multiplication is replaced by addition (see [17, 36, 32]). Note that Eq. 8 can be rewritten to mirror Eq. 9 as:[7]

$$\kappa(X_n, \ldots, X_1) = \sum_{i=1}^{i=n} \kappa(X_n | Par_{X_n}) \qquad (10)$$

We shall show that this requirement augments admissible rankings with the properties of Markov shielding and modularity (see Theorems 8 and 9 below), that we normally attribute to causal organizations.

The following theorem states that the stratification criteria (Eq. 8) does not depend on the specific ordering $\mathcal{O}(\mathcal{X})$. This implies that in order to test whether a given ranking κ is stratified relative to a network Δ, it is enough to test Eq. 8 against *any* ordering $\mathcal{O}(\mathcal{X})$.

Theorem 7 *Given a network Δ, let $\mathcal{O}_1(\mathcal{X})$ and $\mathcal{O}_2(\mathcal{X})$ be two orderings of the elements in $\mathcal{X}$ according to Δ. If κ is stratified for Δ under $\mathcal{O}_1(\mathcal{X})$, then κ is stratified for Δ under $\mathcal{O}_2(\mathcal{X})$.*

3.1 c-entailment

Similar to the case of defaults (Sec 2.1), given a network Δ each stratified ranking κ defines a consequence relation $|\!\sim_\kappa$ where $\phi \,|\!\sim_\kappa \sigma$ iff $\kappa(\sigma \wedge \phi) < \kappa(\neg\sigma \wedge \phi)$ or if $\kappa(\phi) = \infty$. A consequence relation is said to be

[4] We only consider flat causal rules in this paper.

[5] The form $c_1 \wedge \ldots \wedge c_m \rightarrow e$ does not restrict the development of this paper but it clarifies the exposition. A causal rule may take on the general form $\alpha(c_1, \ldots, c_m) \rightarrow \beta(e_1, \ldots, e_n)$ where α and β are any Boolean formulae. Any $\alpha(c_1, \ldots, c_m)$ can be simulated by a set of simpler rules, each containing a conjunction of atomic antecedents. Moreover, any rule $\alpha(c_1, \ldots, c_m) \rightarrow \beta(e_1, \ldots, e_n)$ can be represented by the following set of rules: $\alpha(c_1, \ldots, c_m) \rightarrow e'$, $\beta(e_1, \ldots, e_n) \Rightarrow e'$, and $\neg\beta(e_1, \ldots, e_n) \Rightarrow \neg e'$, where e' is a dummy variable and $\Rightarrow$ is a *strict conditional*.

[6] Note that, in particular, any ordering $\mathcal{O}(\mathcal{X})$ induced by a topological sort on the nodes of $\Gamma_{\langle \mathcal{X}\Delta \rangle}$, where $x_i < x_j$ if v_i precedes v_j in the topological sort, satisfies $\prec$.

[7] An even coarser abstraction of Eq. 9 in the context of relational databases can be found in [7], where the stratification condition is imposed on relations and then used in finding backtrack free solutions for constraint satisfaction problems.

proper for $\phi \mid\!\!\sim_\kappa \sigma$ iff $\kappa(\phi) \neq \infty$. A network Δ **c-entails** σ given ϕ, written $\phi \mid\!\!\sim_\Delta \sigma$, iff $\phi \mid\!\!\sim_\kappa \sigma$ in every κ stratified for Δ, which is proper for $\phi \mid\!\!\sim_\kappa \sigma$. In other words, given Δ, we can expect σ from the evidence ϕ, iff the preference constraint conveyed by $\phi \rightarrow \sigma$ is satisfied by every stratified ranking for Δ. We remark that c-entailment is not to be interpreted as stating that ϕ is believed to *cause* σ. Rather, it expresses an *expectation* to find σ true in the context of ϕ, having given a causal character to the rules in Δ.

Since the set of stratified rankings is a subset of the admissible rankings, by the results in [22], all the inference rules that are sound for cumulative logics [25] and ε-semantics [13] are also sound for c-entailment [18]. These inference rules however, are known to be too weak to constitute a full account of plausible reasoning. The next two theorems provide additional inference power (reflecting the stratification condition) which emanates from the causal structure of Δ. They establish conditions under which these inference rules can be applied modularly to subsets $\Delta' \subset \Delta$ with the guarantee that the resulting inferences will hold in Δ.

Theorem 8 *Let Δ be a network, and let $\{p_r, \ldots, p_s\}$ be a set of literals corresponding to the parent set $\{x_r, \ldots, x_s\}$ of x_t (each p_i, $r \leq i \leq s$, is either x_i or $\neg x_i$). Let e_{x_t} denote a literal built on x_t, and let $\mathcal{Y} = \{y_1, \ldots, y_m\}$ be a set of atomic propositions such that no $y_i \in \mathcal{Y}$ is a descendant of x_t in $\Gamma_{(\mathcal{X},\Delta)}$. Let ϕ_Y be any wff built only with elements from $\mathcal{Y}$ such that $\phi_Y \wedge p_r \wedge \ldots \wedge p_s$ is satisfiable. If $p_r \wedge \ldots \wedge p_s \mid\!\!\sim_\Delta e_{x_t}$ then $\phi_Y \wedge p_r \wedge \ldots \wedge p_s \mid\!\!\sim_\Delta e_{x_t}$.*

Theorem 9 *Let $\Delta' \subset \Delta$ and $\mathcal{X}' \subset \mathcal{X}$ such that if $x' \in \mathcal{X}'$ then all the rules in Δ with either x' or $\neg x'$ as their consequent are also in Δ'. Let φ and ψ be two wffs built with elements from $\mathcal{X}'$. If $\varphi \mid\!\!\sim_{\Delta'} \psi$ then $\varphi \mid\!\!\sim_\Delta \psi$.*

These theorems confirm that stratified rankings exhibit the properties of Markov shielding and modularity. As a corollary to Theorem 9 it is easy to see that c-entailment is insensitive to *irrelevant* propositions, moreover, given two networks with no causal interaction, their respective sets of plausible conclusions will be independent of each other. To obtain a complete proof theory for c-entailment the four axioms of graphoids [29, Chapter 3] need to be invoked.[8] However, Theorems 8 and 9 cover the essence of these axioms and are sufficiently powerful for the purposes of this paper.

To demonstrate the behavior of the proposed formalism, and the usefulness of Theorems 8 and 9 as inference rules, consider the following example:[9]

[8] The conditional independence defined by $\kappa(X_3|X_2, X_1) = \kappa(X_3|X_2)$ is clearly a graphoid since κ represents infinitesimal probabilities.

[9] This example is isomorphic to the YSP [13].

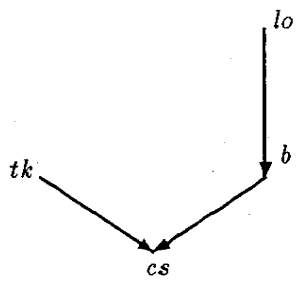

Figure 1: Underlying graph for the causal rules in Example 1

Example 1 (Dead battery) The network $\Delta = \{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, lo \rightarrow bd\}$ encodes the information that "typically if I turn the ignition key the car starts", "typically if I turn the ignition key and the battery is dead the car will not start", and "typically if I leave the head lights on all night the battery is dead". The underlying graph for this network is depicted in Figure 1. Given Δ, and the fact the we left the head lights on all night, we don't expect the car engine to start once we turn the ignition key (i.e., $lo \wedge tk \mid\!\!\sim_\Delta \neg cs$). As in the case of YSP, an unintended scenario exists, in which the car engine actually starts and the battery is not dead after all. Table 1 contains an example of a stratified ranking for Δ, from which we can conclude that $lo \wedge tk \mid\!\!\sim_\Delta \neg cs$ as intended. A formal derivation of this conclusion is given in [18]. The key intermediate steps in this derivation rely on

κ	worlds
0	$(\neg lo, \neg bd, tk, cs)$, $(\neg lo, \neg bd, \neg tk, \neg cs)$
1	$(lo, bd, tk, \neg cs)$, $(lo, bd, \neg tk, \neg cs)$, $(\neg lo, bd, tk, \neg cs)$, $(\neg lo, bd, \neg tk, \neg cs)$
2	$(lo, \neg bd, tk, cs)$, $(lo, \neg bd, \neg tk, \neg cs)$
3	Rest of the ω's

Table 1: Stratified ranking for $\{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, lo \rightarrow bd\}$.

Theorems 8 and 9:

- $tk \wedge lo \mid\!\!\sim_\Delta bd$. This follows from the proposition tk and applying Theorem 9 to the sub-network Δ' containing only the rule $lo \rightarrow bd$.
- $tk \wedge bd \mid\!\!\sim_\Delta \neg cs$ and $tk \wedge bd \wedge lo \mid\!\!\sim_\Delta \neg cs$. The former follows directly from ε-semantics, and the latter from applying Theorem 8 to the rule $tk \wedge bd \rightarrow \neg c$, and the proposition lo.

The next example presents a simple abduction (or backward projection) problem. We contrast the behavior of c-entailment with that of chronological minimization [35].

Example 2 (Unloading the gun.) Consider $\Delta = \{l_0 \rightarrow l_1, l_1 \rightarrow l_2, \ldots, l_{n-1} \rightarrow l_n\}$ standing for the various instances of "typically, if a gun is loaded at time t_i, then it is expected to remain loaded at time t_{i+1}" $(0 \leq i < n)$. We say that a rule $l_i \rightarrow l_{i+1}$ is *falsified* by ω iff $\omega \models l_i \wedge \neg l_{i+1}$; a stratified ranking κ relative to Δ can be constructed as follows:

$$\kappa(\omega) = \text{number of rules in } \Delta \text{ falsified by } \omega \qquad (11)$$

Given that the gun is loaded at t_0 and that it is found unloaded at time t_n (i.e., $l_0 \wedge \neg l_n$ is true), the scheme of chronological minimization will favor the somewhat counterintuitive inference that the gun remained loaded until t_{n-1} (i.e., $l_1 \wedge \ldots \wedge l_{n-1}$ is true). c-entailment on the other hand, only yields the weaker conclusion that the gun must have been unloaded any time within t_1 and t_{n-1} (i.e., $\neg(l_1 \wedge \ldots \wedge l_n)$), but the exact instant where the "unloading" of the gun occurs remains uncertain.

A full account of explanation and abduction using stratified rankings can be found in [18].

c-entailment and chronological minimization are expected to yield the same conclusions in problems of pure prediction, since enforcing ignorance of future events is paramount to the principle of modularity, which is inherent to c-entailment. They differ however in tasks of abduction, as demonstrated in Example 2. In this respect, c-entailment is closer to both *motivated action theory* [37] and *causal entailment* [13]. However, contrary to the motivated action theory, c-entailment automatically enforces specificity-based preferences, which are natural consequences of the conditional interpretation of rules.[10]

We end this section by discussing the *strict* version of a causal rule denoted by $\Rightarrow$, which will be useful in representing non-defeasible causal influences in Section 4. Semantically, strict rules impose the following constraints on the admissibility condition (Def. 2): for each $\varphi \Rightarrow \psi$ in the knowledge base,

$$\kappa(\psi \wedge \varphi) < \kappa(\neg\psi \wedge \varphi) = \infty, \quad \text{and } \kappa(\varphi) < \infty. \qquad (12)$$

Intuitively, a strict conditional voids interpretations that render its antecedent true and its consequent false by assigning them the lowest possible preference; a rank κ equal to infinity. The following are two properties of strict rules:

Theorem 10 *Let* $C_1 \wedge \ldots \wedge C_n \Rightarrow E \in \Delta$

1. **(Contraposition)** *If there exists a stratified ranking for* Δ *where* $\kappa(\neg E) < \infty$ *then* $\neg E \mid\!\!\sim_\Delta \neg(C_1 \wedge \ldots \wedge C_n)$
2. **(Transitivity)** *If* $\varphi \mid\!\!\sim_\Delta \psi$ *and* $\psi \models (C_1 \wedge \ldots \wedge C_n)$ *then* $\varphi \mid\!\!\sim_\Delta E$

These properties mirror the behavior of the material implication "$\supset$", however, they are not entirely identical to those governing a wff of propositional logic. In order for *contraposition* to hold, there is the additional consistency requirement that the negation of the consequent of the rule must be "possible" in at least one ranking. The precondition for *transitivity* to hold is governed by $\varphi \mid\!\!\sim_\Delta \psi$ and not by $\varphi \models \psi$. The semantic difference though between a strict rule $c \Rightarrow e$ and the wff $c \supset e$ is that the former expresses necessary hence permanent constraints while the latter expresses information bound to the current situation. Thus, the former participates in constraining the admissible rankings while the latter is treated as an "observation" formula $\neg c \vee e$, and can affect conclusions only by entering the antecedents of queries.[11]

3.2 c-consistency

Parallel to the notion of admissibility (Def. 2), we can define a notion of *c-consistency* as follows:

Definition 11 A network Δ is **c-consistent** iff there exists at least one stratified ranking κ for Δ

An example of a c-inconsistent network is the following: $\Delta = \{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, tk \rightarrow x, x \rightarrow bd\}$.[12] The lack of an appropriate causal representation for this set of rules is not surprising. If we accept that tk causes cs, we should expect $\neg bd$ to hold by default when tk is true. On the other hand if there is a *causal path* from tk to bd, we should expect bd to hold in the context of tk. Note that this trouble case is admissible as shown by the ranking in Table 2.[13] This ranking depicts a situation in which the act of predicting the consequences of turning the key seems to protect the battery against the damage inflicted by x and such a flow of events is indeed contrary to the common understanding of causation.

Another admissible yet c-inconsistent set is $\Delta = \{a \Rightarrow c, b \Rightarrow \neg c\}$ which might possibly arise when we physically connect the outputs of two logic gates with conflicting functions. A stratified ranking for Δ would imply that

$$\kappa(a \wedge b) = \kappa(a) + \kappa(b) \qquad (13)$$

In other words the abnormality of a should be independent of the abnormality of b. However, if each time we

[10] We remark that the formalism in [37] deals with a much richer time ontology than the formalism presented here, and with a first-order language.

[11] See [14] for further discussion of strict rules vs. material implication.

[12] This is the network used in Example 1 augmented with two rules $tk \rightarrow x$ and $x \rightarrow bd$.

[13] This ranking is not stratified for Δ since $\kappa(bd \wedge x \wedge tk) = 2$, but $\kappa(bd|x) + \kappa(x|tk) + \kappa(tk) = 1$ which contradicts Eqs. 8 and 10.

κ	worlds
0	$(\neg tk, x, bd, \neg cs)$
1	$(tk, x, \neg bd, cs)$
2	$(tk, x, bd, \neg cs)$
3	Rest of the ω's

Table 2: Admissible ranking for $\{tk \rightarrow cs, tk \wedge bd \rightarrow \neg cs, tk \rightarrow x, x \rightarrow bd\}$.

observe a we should expect c, but each time we observe b we should expect not c, a and b must be mutually exclusive, hence negatively correlated events. Indeed, since $\kappa(a \wedge b \wedge c) = \kappa(a \wedge b \wedge \neg c) = \infty$ then $\kappa(a \wedge b) = \infty$ and Eq. 13 cannot be satisfied unless either a or b is permanently false, thus defying the "possible antecedent" requirement for strict rules (Eq. 12).

3.3 The most normal stratified ranking

In Section 2.1 we showed that the constraints provided by the default rules need to be supplemented with an additional assumption, so as to obtain a unique preferred ranking. The incorporation of this "most normal" assumption in the context of stratified rankings, results in a substantial increase of expressiveness as discussed in [18].

Note however, that contrary to the case of system-Z^+ (see Theorem 4), the most-normal stratified ranking may not be unique; for example the network $\Delta = \{a \rightarrow c, b \rightarrow \neg c\}$ has two minimal rankings [18]. Thus, we now need to define entailment in minimal rankings, denoted by $|\!\sim_\Delta^*$, with respect to the consequence relations of all most-normal stratified rankings.

4 Belief Revision and Updating

In this section we demonstrate how the semantics of model ranking, together with the syntactic machinery developed for processing queries, can be applied to manage the tasks of belief revision and belief update. In both tasks we seek to incorporate a new piece of information ϕ into an existing set of beliefs ψ. In belief revision ϕ is assumed to be a piece of evidence while in update ϕ is treated as a change occurring by external intervention. We first apply the evidence handling capability of system-Z^+ to belief revision and then use stratified ranking and its representation of actions and causal relations to govern the dynamics of belief update.

4.1 Belief revision

AGM have advanced a set of postulates that have become a standard against which proposals for belief revision are tested [1]. The AGM postulates model epistemic states as deductively closed sets of (believed) sentences and characterize how a rational agent should change its epistemic states when new beliefs are added, subtracted, or changed. The central result is that the postulates are equivalent to the existence of a complete preordering of all propositions according to their degree of *epistemic entrenchment* such that belief revisions always retain more entrenched propositions in preference to less entrenched ones. Although the AGM postulates do not provide a calculus with which one can realize the revision process or even specify the content of an epistemic state [3, 10, 27], they nevertheless imply that a rational revision must behave as though propositions were ordered on some scale.

Spohn [36] has shown how belief revision conforming to the AGM postulates can be embodied in the context of ranking functions. Once we specify a single ranking function κ on possible worlds, we can associate the set of beliefs, with those propositions β for which $\kappa(\neg\beta) > 0$. It follows then that the models for the theory ψ representing our beliefs (written $Mods(\psi)$) consist of those worlds ω for which $\kappa(\omega) = 0$. To incorporate a new belief ϕ, one can raise the κ of all models of $\neg\phi$ relative to those of ϕ, until $k(\neg\phi)$ becomes (at least) 1, at which point the newly shifted ranking defines a new set of beliefs. This process of belief revision, which Spohn named α-conditioning (with $\alpha = 1$ for this particular case), represents the ranking equivalent of Jeffrey's rule of probability kinematics [29] and was shown to comply with the AGM postulates [12]. It follows then that the process of revising beliefs in all three forms of conditioning also obey the AGM postulates. Ordinary conditioning amounts to setting $\alpha = \infty$, J-conditioning amounts to $\alpha = J$, while L-conditioning calls for shifting the models of ϕ relative to those of $\neg\phi$ by L units of surprise. If we denote by $\kappa_\phi(\omega)$ the revised ranking after conditioning (with $\alpha = \infty$), then the dynamics of belief is governed by the following equation:

$$\kappa_\phi(\omega) = \begin{cases} \kappa(\omega) - \kappa(\phi) & \text{if } \omega \models \phi, \\ \infty & \text{otherwise.} \end{cases} \tag{14}$$

Accordingly, testing whether a given sentence β is believed after revision amounts to testing whether $\kappa_\phi(\neg\beta) > 0$ or, equivalently, whether $\kappa(\neg\beta|\phi) > 0$.

The unique feature of the system described in this paper is that the above test can be performed by purely syntactic terms, involving only the rules in Δ [17]. These computations are demonstrated in the following example.

Example 3 (Working students) The set $\Delta = \{s \rightarrow \neg w, s \rightarrow a, a \rightarrow w\}$ stands for "typically students don't work", "typically students are adults", and "typically adults work", respectively.[14] The Z^+ priorities on the rules (computed according to Eq. 5) are: $Z^+(a \rightarrow w) = 0$ and $Z^+(s \rightarrow \neg w) = Z^+(s \rightarrow a) = 1$,

[14] Note that all δ_i's are 0 for this example

from which the initial κ^+ ranking can be computed (Eq. 4), as depicted in Table 3. The rankings in Tables 4 and 5 show the revised rankings after observing an adult (κ_a) and a student (κ_s) respectively.

κ^+	Possible worlds
0	$(\neg s, a, w)$, $(\neg s, \neg a, w,)$, $(\neg s, \neg a, \neg w,)$
1	$(\neg s, a, \neg w)$, $(s, a, \neg w,)$
2	(s, a, w), $(s, \neg a, \neg w)$, $(s, \neg a, w)$

Table 3: Initial ranking for the student triangle in Example 3

κ_a^+	Possible worlds
0	$(\neg s, a, w)$
1	$(\neg s, a, \neg w)$, $(s, a, \neg w,)$
2	(s, a, w)

Table 4: Revised ranking after observing an adult

κ_s^+	Possible worlds
0	$(s, a, \neg w,)$
1	(s, a, w), $(s, \neg a, \neg w)$, $(s, \neg a, w)$

Table 5: Revised ranking after observing a student

The beliefs associated with these rankings can be computed from the worlds residing in $\kappa^+ = 0$. Thus, in κ_a^+ "an adult works", whereas in κ_s^+ "a student is an adult that does not work". These beliefs can be computed more conveniently by syntactic analysis of the rules and their Z^+ priorities, either by using Eq. 6, or by extracting from Δ a propositional theory that is maximally consistent with the observation using Eq. 7. For example, the beliefs associated with observing a student s are given by the theory $\{s, s \supset a, s \supset \neg w\}$. These two implications mirror the rules $s \rightarrow \neg w$ and $s \rightarrow a$ which are the unique set of rules that are maximally consistent with s.

4.2 Discussion and related work

There are several computational and epistemological advantages to basing the revision process on a finite set of conditional rules, and not on the beliefs, or on the rankings or the expectations that emanate from those rules. The number of propositions in one's belief set is astronomical, as is the number of worlds, while the number of rules is usually manageable.

This computational necessity has been recognized by several researchers. Nebel [27] adapted the AGM theory so that finite sets of *base* propositions mediate revisions. The basic idea in these syntax-based systems is to define a (total) priority order on the set of base propositions and to select revisions to be maximally consistent relative to that order as exemplified in the nonmonotonic systems of Brewka [5] and Poole [33] and in Example 3. Nebel has shown that such a strategy, can satisfy almost all the AGM axioms. Boutilier [3] has further shown that, indeed, the priority function Z^+ corresponds naturally to the epistemic entrenchment ordering of the AGM theory.[15]

Unfortunately, even Nebel's theory does not completely succeed at formalizing the practice of belief revision, as it does not specify how the priority order on the base propositions is to be determined. Although one can imagine, in principle, that the knowledge author specify this priority order in advance, such specification would be impractical, since the order might (and, as we have seen, should) change whenever new rules are added to the knowledge base. By contrast, system-Z^+ extracts both beliefs and rankings of beliefs automatically from the content of Δ; no outside specification of belief orderings is required.

Finally, and perhaps most significantly, system-Z^+ is capable of responding not merely to empirical observations but also to linguistically transmitted information such as conditional sentences (i.e., if-then rules). For example, suppose someone tells us that leaving the radio on also tends to render the battery dead; we add this new rule to our knowledge base (verifying first that the addition is admissible), recompute Z^+, and are prepared to respond to new observations or hearsay. In Spohn's system, where revisions are limited to α-conditioning, one cannot properly revise beliefs in response to conditional statements. The AGM postulates, too, are inadequate for describing revision due to incorporation of new conditionals.[16]

The ability to adopt new conditionals (as rules) also provides a simple semantics for interpreting nested conditionals (e.g., "If you wear a helmet whenever you ride a motorcycle, then you won't get hurt badly if you fall"[17]). Nested conditionals cease to be a mystery once we permit explicit references to default rules. The sentence "If $(a \rightarrow b)$ then $(c \rightarrow d)$" is interpreted as

> "If I add the default $a \rightarrow b$ to Δ, then the conditional $c \rightarrow d$ will be satisfied by the consequence relation $\mid\!\sim_{\Delta'}$ of the resulting knowledge base $\Delta' = \Delta \cup \{a \rightarrow b\}$".

[15] The proof in [3] considers the priorities Z^+ resulting from a *flat* set of rules as in system-Z [31]. Boutilier [4] also shows that an entrenchment ordering obeying the AGM framework obtains from the Z priorities of the negations of the material counterpart of rules.

[16] Gärdenfors [12] attempts to devise postulates for conditional sentences, but finds them incompatible with the Ramsey test (page 156-160).

[17] An example due to Calabrese (personal communication).

which is clearly a proposition that can be tested in the language of default-based ranking systems. Note the essential distinction between having a conditional sentence $a \rightarrow b$ explicitly in Δ versus having a conditional sentence $a \rightarrow b$ *satisfied* by the consequence relation of Δ. This distinction gets lost in systems that do not acknowledge defaults as the basis for ranking and beliefs.[18]

4.3 Belief update

The introduction of stratified ranking adds the capability for implementing a new type of belief changes, named *update* by Katsuno and Mendelzon, which result from external influences, and which act differently from those reflecting new evidence. Katsuno and Mendelzon [21] have shown that the AGM postulates are inadequate for describing changes caused by updates, for which they have proposed new sets of postulates. The basic difference between revision and update is that the latter permits changes in each possible world independently, as was proposed by Winslett [38].[19]

This type of belief change can be embodied in a stratified ranking system using the following device: For each instruction to "update the knowledge base by ϕ" we add a set of rules that simulates the action "$do(\phi)$, leaving everything else constant (whenever possible)", and then condition κ on the truth of $do(\phi)$. The following set of causal rules embody the intent of this action, where ϕ and ϕ' stand for "ϕ holds at t" and "ϕ holds at $t' > t$", respectively:[20]

$$\phi \rightarrow \phi' \tag{15}$$

$$\neg\phi \rightarrow \neg\phi' \tag{16}$$

$$do(\phi) \Rightarrow \phi'. \tag{17}$$

The following example (adapted from Winslett [38]) demonstrates how this device differentiates between update and revision.

Example 4 (XOR-gate) A XOR Boolean gate $c = XOR(a, b)$ is examined at two different times. At time t, we observe the output $c = true$ and conclude that one of the inputs a or b must be true, but not both. At a later time t' we learn that b' is true (primed letters denote propositions at time t'), and we wish to change our beliefs (in a and a') accordingly. Naturally, this change should depend on how the truth of

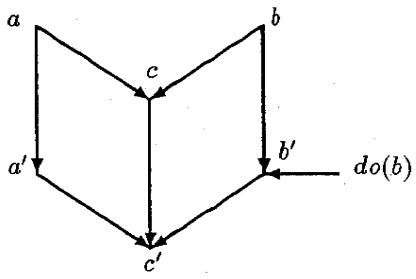

Figure 2: Graph depicting the causal dependencies in Example 4

b' is learned. If we learn b' by measuring the voltage on the b terminal of the gate, then we have a belief revision process on our hands, and we expect a' to be false. On the other hand, if we learn that b' is true as a result of physically connecting the b terminal to a voltage source, we no longer expect a' to be false, since we have no reason to believe that the output c has retained its truth value in the process.

In the stratified ranking formulation, the knowledge base corresponding to this example will consist of three components:

1. The functional description of the XOR gate at times t and t',

$$a \wedge b \Rightarrow \neg c \quad ; \quad \neg a \wedge \neg b \Rightarrow \neg c \tag{18}$$

$$a \wedge \neg b \Rightarrow c \quad ; \quad \neg a \wedge b \Rightarrow c, \tag{19}$$

   and an equivalent set of rules for a', b', c'.

2. The persistence rules: For every x in $\{a, b, c\}$,

$$x \rightarrow x' \quad ; \quad \neg x \rightarrow \neg x'. \tag{20}$$

3. The action $do(b)$, which represents the external influence on b':

$$do(b) \Rightarrow b'. \tag{21}$$

The underlying graph for the network Δ corresponding to this knowledge base is depicted in Figure 2.

Initially, after observing c, our evidence consists only of c. The minimal stratified ranking κ_c for a Δ consisting of rules in Eqs. 18-21 is depicted in Table 6. To represent belief revision, we add b' to our evidence set and query whether $c \wedge b' \mid\!\sim^*_\Delta \neg a'$.[21] In contrast, to represent belief update, we add $do(b)$ to our evidence set and query whether $(c \wedge b' \wedge do(b)) \mid\!\sim^*_\Delta \neg a'$.

It can be shown that the first query is answered in the affirmative, as the second in the negative. The left-hand side of Table 7 shows the ranking resulting from

[18] Belief revision systems proposed in the database literature [11, 6] suffer from the same shortcoming. In that context-defaults represent integrity constraints with exceptions.

[19] In the language of Bayesian networks, the difference between updates and revisions parallels the distinction between causal and evidential information [28].

[20] The two persistence rules, Eqs. 15 and 16, are presumed to apply between any two atomic propositions at two successive times.

[21] Recall that $\mid\!\sim^*_\Delta$ denotes the consequence relation of the minimal stratified ranking for Δ (see Sec. 3.3), which is unique for this example.

κ_c	$\neg do(b)$	$do(b)$
0	$(\neg a, b, \neg a', b'), (a, \neg b, a', \neg b')$	
1	$(\neg a, b, a', b'), (a, \neg b, \neg a', \neg b'), (a, \neg b, a', b')$	$(\neg a, b, \neg a', b'), (a, \neg b, a', b')$
2	$(\neg a, b, a', \neg b'), (a, \neg b, \neg a', b')$,	$(\neg a, b, a', b'), (a, \neg b, \neg a', b')$
∞	models for $\neg c$	models for $\neg c$

Table 6: Minimal stratified ranking for Example 4 after c is observed

κ_c	Revision $\kappa_c(\omega\|b')$	Update $\kappa_c(\omega\|do(b))$
0	$(\neg a, b, \neg a', b')$	$(\neg a, b, \neg a', b'), (a, \neg b, a', b')$
1	$(\neg a, b, a', b'), (a, \neg b, a', b')$	$(\neg a, b, a', b'), (a, \neg b, \neg a', b')$
2	$(a, \neg b, \neg a', b')$,	
∞	models for $\neg b'$	models for $\neg do(b)$

Table 7: Rankings after observing b, and after "doing" b

the revision of the ranking in Table 6 by b' (first query), while the right-hand side shows the ranking of worlds after updating by $do(b)$ (second query). Note that in the revised ranking the only world in the zero rank is a model for $\neg a'$, while the updated ranking shows an additional world which is a model for a' (the state of the output c in this world changed as a consequence of the action). The action $do(b)$ establishes the truth of b' but has no effect on what we believe about the second input a'. Since neither a nor $\neg a$ were believed at t, they remain unbelieved at t'.

4.4 The dynamics of belief update

The example above demonstrates that, given a ranking κ and a network Δ, it is possible to predict a system's behavior under external interventions. For example, if we wish to inquire whether event e will hold true after we force some variable A to become true, we simply add to Δ the rule $do(a) \Rightarrow a$,[22] recompute the resulting stratified ranking κ', and compute $\kappa'(e|do(a))$. It can be shown [16] that there is a simple relation between $\kappa(e|a)$ and $\kappa'(e|do(a))$, which is best represented as a transformation between two ranking functions, $\kappa(\omega)$ and $\kappa'(\omega)$, the latter being an abbreviation of $\kappa'(w|do(a))$. We simply replace the term $\kappa(a|Par_A)$ in the sum of Eq. 10 with the term $\kappa(a|do(a))$, representing the new influence $do(a)$ that now governs a:

$$\kappa'(\omega) = \begin{cases} \kappa(w) - \kappa(a|Par_A(w)) & \text{if } \omega \models a. \\ \infty & \text{if } \omega \models \neg a. \end{cases} \quad (22)$$

In other words, the κ of each world ω satisfying a is reduced by an amount equal to the degree of surprise of finding $A = true$, given the realization of Par_A in ω (the κ of each world falsifying a is of course ∞). Such independent movement from world to world is shown in Example 4, where $\kappa(\omega)$ is depicted on the left-hand side of Table 6 and $\kappa'(\omega)$ is depicted on the right hand side of Table 7. If A has no *parents* (direct causes), then κ' is obtained by shifting the κ of each $\omega \models a$ by a constant amount $\kappa(a)$, as in ordinary conditioning, and $\kappa'(\omega)$ would be equal to $\kappa(\omega|a)$, as expected. However, when the manipulated variable has direct causes Par_A, the amount of shift would vary from world to world, depending on how surprising it would be (in that world) to find a happening naturally (without external intervention). For instance, if A is governed by persistence rules, $a(t-1) \rightarrow a(t), \neg a(t-1) \rightarrow \neg a(t)$, then worlds in which $a(t-1)$ hold will shift less than those in which $a(t-1)$ is false, because $a(t)$ is expected to hold in the former and not in the latter. Note that the amount of shift subtracted from $\kappa(\omega)$ is equal precisely to the fraction of surprise $\kappa(a|Par_A(\omega))$ that $A = true$ contributes to $\kappa(\omega)$ and that now becomes *explained away* (hence excusable) by the action $do(a)$.

4.5 Relation to KM postulates

It can be shown [16] that when the update by a formula ϕ is given as a conjunction of literals (representing concurrent or sequential actions), then the movement of worlds toward $\kappa = 0$ will yield a set of updated beliefs consistent with the KM postulates.[23] More specifically, for every world $\omega \models \neg\phi$ that is currently in

[22] We use lowercase to denote the instantiation of variable A to a truth value.

[23] Updates involving disjunctions require special treatment. If they are to be interpreted as a license to effect any change satisfying the disjunction, then the final state of belief is the union, taken over all disjuncts, of worlds that drift to $\kappa = 0$. In this interpretation, the instruction "make sure the box is painted either blue or white" will leave the box color unknown, even knowing that the box was white initially (contrary to the postulate (U2) of KM). However, if the intention is to effect no change as long as the disjunctive condition is satisfied, then the knowledge base should be augmented with an observation-dependent strategy "$do(\phi)$ when ϕ is not satisfied", instead of using the pure action $do(\phi)$. Conditioning on such a strategy again yields a belief set consistent with the KM postulates. The first interpretation is useful for discrediting earlier observations, for example, "I am not sure the em-

$\kappa = 0$ there is at least one image world $\omega^* \models \phi$, having $\kappa(\omega^*) > 0$ that will end up at $\kappa'(\omega^*) = 0$ according to Eq. 22. In an image world ω^*, every term $\kappa(x_i|Par_{X_i}(\omega^*)) > 0$ represents a violation of expectation that would be totally excusable were it caused by an external intervention such as ϕ. Intuitively, the image world corresponds to a scenario in which all the unexpected events are attributed to the intervention of ϕ but otherwise the world follows its natural, unperturbed course as dictated by the prediction of the causal theory.

That updates resulting from Eq. 22 comply with the KM postulates can be seen by the following consideration.[24] KM have shown that their axioms are equivalent to the existence of a function mapping each possible interpretation world ω to a partial preorder $\leq_\omega$, such that for any interpretation ω', if $\omega \neq \omega'$ then $\omega <_\omega \omega'$. Then the set of models for the update of a formula ψ (representing our current beliefs) by a formula ϕ, written $\psi \diamond \phi$, is found by taking the union of the minimal models for ϕ, with respect to each one of the pre-orders defined by the models for ψ:

$$Mods(\psi \diamond \phi) = \bigcup_{\omega \in Mods(\psi)} \min(Mods(\phi), \leq_\omega). \quad (23)$$

It is not hard to show that the image ω^* as described above is indeed a minimal element in the order $\leq_\omega$, defined as follows:

Definition 12 (World orderings) Let $\mathcal{O} = x_1, x_2, \ldots, x_n$ be any order of the variables that is consistent with the dag $\Gamma_{(\mathcal{X},\Delta)}$. Given three worlds ω, ω_1, and ω_2, we say that $\omega_1 \leq_\omega \omega_2$ iff the following conditions hold:

1. ω disagrees with ω_2 on a literal that is earlier (in $\mathcal{O}$) than any literal on which ω disagrees with ω_1.
2. If a tie occurs, then $w_1 \leq_\omega \omega_2$ if $\kappa(\omega_1) \leq k(w_2)$.

4.6 Related work

The connection between belief update and theories of action was noted by Winslett [38] and has been elaborated more recently by del Val and Shoham [8] using the situation calculus.

Unlike del Val and Shoham [8], we would not claim that "the KM-postulates need not be postulated at all, but can instead be derived analytically". While the KM postulates can indeed be derived from our formulations of actions, persistence, and causation, the interesting power of these postulates is that they cover a wide variety of such formulations, from a simple theory such as ours to the intricate machinery of the situation calculus. Our analysis offers the KM postulates an intuitive, model-theoretic support that is well grounded in probability theory, where the distinction between observations and actions can be formulated naturally and tractably. It also offers a simple unification of revision and update, since both are embodied in a conditioning operator, the former by conditioning on *observations* and the latter by conditioning on *actions*.

Grahne et. al. [19] showed that revision could be expressed in terms of an update operator in a language of introspection (intuitively, *observing* a piece of evidence has the same effect as *causing* the observer to augment her beliefs by that very evidence). Our analysis shows that the converse is also true: belief updates can be expressed in terms of a *conditioning* operator, which is normally reserved for belief revision. The intuition is that *acting* to produce a certain effect yields the same beliefs as *observing* that action performed. This translation is facilitated by the special status that the added *action* $\Rightarrow$ *effect* rules enjoy in stratified ranking, where actions are always represented as root nodes, independent of all other events except their consequences. This ensures that the immediate effects of those actions are explained away and do not reflect back on other events in the past. It is this stratification that produces the desired distinction between observing an action produce an effect and observing the effect without the action.[25]

Acknowledgements

This work was supported in part by NSF grant #IRI-9200918, AFOSR grant #900136, and MICRO grant #91-124. We thank C. Boutilier, D. Lehmann, and two referees for comments on an earlier draft of this paper.

References

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] A. B. Baker. Nomonotonic reasoning in the framework of situation calculus. *Artificial Intelligence*, 49:5–23, 1991.

[3] C. Boutilier. Conditional logics for default reasoning and belief revision. Ph.D. dissertation, University of Toronto, 1992.

[4] C. Boutilier. What is a default priority? In *Proceedings of CCAI-92*, Vancounver, 1992.

ployee's salary is $50K$; it could be anywhere between $40K$ and $60K$".

[24] A formal proof can be found in [16].

[25] Note that update cannot be expressed in terms of the AGM operators of revision and contraction, because it is impossible to simulate with these operators the acceptance of a new conditional $do(\phi) \Rightarrow \phi$, so that the acceptance of $do(\phi)$ is treated differently than the acceptance of ϕ. Similarly, update cannot be formulated in Spohn's system, because the identity of the image world ω^* cannot be described in terms of the initial ranking alone; it requires the causal theory Δ.

[5] G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of IJCAI-89*, Detroit, 1989.

[6] M. Dalal. Investigations into a theory of knowledge base revision: Preliminary report. In *Proceedings of AAAI-88*, pages 475–479, 1988.

[7] R. Dechter and J. Pearl. Directed constraint networks: A relational framework for causal modeling. In *Proceedings of IJCAI-91*, Australia, 1991.

[8] A. del Val and Y. Shoham. Deriving properties of belief update from theories of acion. In *Proceedings of AAAI-92*, pages 584–589, San Jose, California, 1992.

[9] W. Dowling and J. Gallier. Linear-time algorithms for testing the satisfiability of propositional Horn formulae. *Journal of Logic Programming*, 3:267–284, 1984.

[10] J. Doyle. Rational belief revision (preliminary report). In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 163–174, Cambridge, Massachusetts, 1991.

[11] R. Fagin, J. D. Ullman, and M. Vardi. On the semantics of updates in databases. In *Proceedings of the 2^{nd} ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 352–365, 1983.

[12] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, 1988.

[13] H. A. Geffner. Default reasoning: Causal and conditional theories. MIT Press, Cambridge, 1991.

[14] M. Goldszmidt and J. Pearl. On the consistency of defeasible databases. *Artificial Intelligence*, 52:121–149, 1991.

[15] M. Goldszmidt and J. Pearl. System Z^+: A formalism for reasoning with variable strength defaults. In *Proceedings of AAAI-91*, pages 394–404, Anaheim, CA, 1991.

[16] M. Goldszmidt and J. Pearl. Dynamics of belief update. Technical Report TR-190, University of California Los Angeles, Cognitive Systems Lab., Los Angeles, 1992, (In preparation).

[17] M. Goldszmidt and J. Pearl. Reasoning with qualitative probabilities can be tractable. In *Proceedings of the 8^{th} Conference on Uncertainty in AI*, Stanford, 1992.

[18] M. Goldszmidt and J. Pearl. Stratified rankings for causal relations. In *Proceedings of the Fourth International Workshop on Nonmonotonic Reasoning*, Vermont, 1992.

[19] G. Grahne, A. Mendelzon, and R. Reiter. On the semantics of belief revision systems. In *Proceedings of TARK-92*, pages 132–142, Monterrey, CA, 1992.

[20] S. Hanks and D. McDermott. Non-monotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[21] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 387–394, Boston, 1991.

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

[23] D. Lehmann. What does a conditional knowledge base entail? In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 212–222, Toronto, 1989.

[24] V. Lifschitz. Formal theories of action. In M. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*, pages 410–432. Morgan Kaufmann, San Mateo, 1987.

[25] D. Makinson. General theory of cummulative inference. In M.Reinfrank, J. de Kleer, M. Ginsberg, and E. Sandewall, editors, *Non-monotonic Reasoning.* Springer-Verlag, Lecture Notes on Artificial Intelligence 346, Berlin, 1989.

[26] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[27] B. Nebel. Belief revision and default reasoning: Syntax-based approaches. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 417–428, Cambridge, Massachusetts, 1991.

[28] J. Pearl. Embracing causality in formal reasoning. *Artificial Intelligence*, 35:259–271, 1988.

[29] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[30] J. Pearl. Jeffrey's rule, passage of experience and neo–Bayesianism. In H. Kyburg, R. Loui, and G. Carlson, editors, *Defeasible Reasoning and Knowledge Representation*, pages 121–135. Kluwer Publishers, San Mateo, 1990.

[31] J. Pearl. System Z: A natural ordering of defaults with tractable applications to default reasoning. In R. Parikh, editor, *Proceedings of TARK-90*, pages 121–135. Morgan Kaufmann, San Mateo, CA, 1990.

[32] J. Pearl. Epsilon-semantics. In *Encyclopedia of Artificial Intelligence*, pages 468–475. Wiley Interscience, New York, 1992. Second Edition.

[33] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[34] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[35] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, Mass., 1988.

[36] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, pages 105–134. Reidel, Dordrecht, Netherlands, 1987.

[37] L. Stein and L. Morgenstern. Motivated action theory: A formal theory of causal reasoning. In *Proceedings of AAAI-88*, pages 518–523, 1988.

[38] M. Winslett. Reasoning about action using a possible worlds approach. In *Proceedings of AAAI-88*, pages 89–93, 1988.