Artificial Intelligence 238 (2016) 96–118

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Making the right exceptions ☆,☆☆

Harald Bastiaanse, Frank Veltman *

Institute for Logic, Language and Computation, Universiteit van Amsterdam, P.O. Box 94242, 1090 GE Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form 26 May 2016
Accepted 30 May 2016
Available online 2 June 2016

Keywords:
Circumscription
Defaults
Nonmonotonic logic
Inheritance networks

ABSTRACT

This paper is about the logical properties of sentences of the form *S's are normally P*, and starts from the idea that any logical theory for such sentences should meet the following simple requirement:

> If the only available information about some object *x* is that *x* has property *S*, it must be valid to infer by default that *x* has all the properties *P* that objects with property *S* normally have.

We investigate how this requirement can be met by theories developed within the framework of circumscription, and specify a constraint – the *exemption principle* – that must be satisfied to do so. This principle determines in cases of conflicting default rules which objects are *exempted* from which rules, and, as such, is the main source for the capricious logical behavior of the sentences we are interested in.
To facilitate comparison (and implementation) we supply an algorithm for inheritance networks and prove that arguments that can be expressed in both frameworks are valid on the circumscriptive account if and only if the inheritance algorithm has a positive outcome.

1. Introduction

Discussions often end before the issues that started them have been resolved. In the 1980s and 1990s default reasoning was a hot topic in the field of logic and AI. The result of this discussion was not one single theory that met with general agreement, but a collection of alternative theories, each with its merits, but none entirely satisfactory. This paper aims to give a new impetus to this discussion.

The issue is the logical behavior of sentences of the form

S's are normally P

Such sentences function as default rules. What they mean is roughly this: when you are confronted with an object with property *S*, and you have no evidence to the contrary, you are legitimized to assume that this object has property *P*.

☆ The research for this paper was partly financed by the Netherlands Organisation for Scientific Research (project NWO-360-20-200), whose support is gratefully acknowledged.

☆☆ This paper has a long history. Successive drafts have been presented between 2009 and 2015 in Amsterdam, Groningen, Gent, Göttingen, Westport, Ann Arbor, Beijing, and Guangzhou. We thank the audiences for their valuable feedback. We also want to thank the referees, whose suggestions led to significant improvements both in content and in presentation.

* Corresponding author.
E-mail addresses: bastiaanse_harald@hotmail.com (H. Bastiaanse), f.veltman@uva.nl (F. Veltman).

http://dx.doi.org/10.1016/j.artint.2016.05.005
0004-3702/

The 'evidence to the contrary' can vary. Sometimes it simply consists in the empirical observation that the object concerned is in fact an exception to the rule. On other occasions the evidence may be more indirect. Consider:

premise 1	*A's are normally E*
premise 2	*S's are normally not E*
premise 3	*S's are normally A*
premise 4	*c is A and c is S*
by default	*c is not E*

This is a case of conflicting defaults.[1] At first sight one might be tempted to draw both the conclusion that *c is E* (from premises 1 and 4) and that *c is not E* (from premises 2 and 4), and maybe on second thought to draw neither. But the third premise states that objects with the property *S* normally have the property *A* as well. So, apparently, *normal S*'s are *exceptional A*'s, as the rule that *A's are normally E* does not hold for them. In other words, only the *S*-defaults apply to *c*. So, presumably, *c* is not *E*.

Default reasoning has been formalized in various ways, and within each of the existing theoretical frameworks a number of strategies have been proposed to deal with conflicting defaults. In the following we will focus on two of these frameworks, Circumscription (McCarthy [1,2]), and Inheritance Networks (Horty et al. [3]), and implement a new strategy to deal with conflicting rules in each of these.

2. Naive circumscription

Within the circumscriptive approach a sentence of the form *S's are normally P* is represented by a formula of the form

$$\forall x((Sx \wedge \neg Ab_{SxPx}\, x) \rightarrow Px).$$

Here $Ab_{SxPx}\, x$ is a one place predicate. The subscript '$SxPx$' serves as an index, indicating the rule concerned. If an object a satisfies the formula $Ab_{SxPx}\, x$, this means that a is an *abnormal* object with respect to this rule.

More generally, let $\mathcal{L}_0$ be a language of *monadic* first order logic. With each pair $\langle \varphi(x), \psi(x) \rangle$,[2] we associate a new one-place predicate $Ab_{\varphi(x)\psi(x)}$, thus obtaining the first order language $\mathcal{L}$.

A *default rule* is a formula of $\mathcal{L}$ of the form

$$\forall x((\varphi(x) \wedge \neg Ab_{\varphi(x)\psi(x)}\, x) \rightarrow \psi(x)).$$

Here, $\varphi(x)$ and $\psi(x)$ must be formulas of $\mathcal{L}_0$ that are quantifier-free and in which no individual constant occurs. The formula $\varphi(x)$ is called the *antecedent* of the rule, $Ab_{\varphi(x)\psi(x)}\, x$ is its *abnormality clause*, and $\psi(x)$ its *consequent*. Again, the index $\varphi(x)\psi(x)$ is there just to indicate that it concerns the abnormality predicate of the rule with antecedent $\varphi(x)$ and consequent $\psi(x)$. When it is clear which variable is at stake we will write $Ab_{\varphi\psi}$ rather than $Ab_{\varphi(x)\psi(x)}$. And often we will shorten '$\forall x((\varphi(x) \wedge \neg Ab_{\varphi\psi}\, x) \rightarrow \psi(x))$' further to

$$\forall x(\varphi(x) \rightsquigarrow \psi(x)).$$

Since it is clear from the antecedent and the consequent of a default rule what the abnormality clause is, this should not cause confusion.[3]

In ordinary logic, for an argument to be valid, the conclusion must be true in *all* models in which the premises are true. The basic idea underlying circumscription is that not all models of the premises matter but only the most normal ones – only the ones in which the extension of the abnormality predicates is inclusion-wise minimal given the information at hand. Formally:

Definition 2.1.

(i) Let $\mathcal{L}$ be a language as described above, and let $\mathfrak{A} = \langle \mathcal{A}, \mathcal{I} \rangle$ and $\mathfrak{A}' = \langle \mathcal{A}', \mathcal{I}' \rangle$ be two models for $\mathcal{L}$ with the following properties:
  (a) $\mathcal{A} = \mathcal{A}'$;
  (b) for all individual constants c, $\mathcal{I}(c) = \mathcal{I}'(c)$;

[1] If a concrete example is wanted, substitute '*adult*' for *A*, '*employed*' for *E*, and '*student*' for *S*.

[2] Notation: we write $\varphi(x)$ to denote a formula φ of $\mathcal{L}_0$ in which (at most) the variable x occurs freely.

[3] Some readers may not like the fact that in this set up the formulas $\forall x(Sx \rightsquigarrow Px)$ and $\forall y(Sy \rightsquigarrow Py)$ are not logically equivalent, because they contain different abnormality predicates. We could remedy this defect by introducing the same abnormality predicate $Ab_{\varphi(\cdot)\psi(\cdot)}$ for all pairs $\langle \varphi(x), \psi(x) \rangle$, independent of the free variable x occurring in $\varphi(x)$ and $\psi(x)$. Here '$\cdot$' refers to a symbol that does not belong to the vocabulary of $\mathcal{L}_0$, and by $\varphi(\cdot)$, we mean the expression that one obtains from $\varphi(x)$ by replacing each free occurrence of x by an occurrence of $\cdot$.

Some readers may insist that on top of this we should enforce that whenever $\varphi(x)$ is logical equivalent to $\chi(x)$, and $\psi(x)$ to $\theta(x)$, $\forall x(\varphi(x) \rightsquigarrow \psi(x))$ gets equivalent to $\forall x(\chi(x) \rightsquigarrow \theta(x))$. This can be done by stipulating that we are only interested in models that assign the same extension to $Ab_{\varphi(\cdot)\psi(\cdot)}$ and $Ab_{\chi(\cdot)\theta(\cdot)}$ if $\varphi(x)$ is logical equivalent to $\chi(x)$ and $\psi(x)$ to $\theta(x)$. However, for our purposes, we can keep things simple.

(c) for all abnormality predicates $Ab_{\varphi\psi}$, $\mathcal{I}(Ab_{\varphi\psi}) \subseteq \mathcal{I}'(Ab_{\varphi\psi})$.
Then $\mathfrak{A}$ *is at least as normal as* $\mathfrak{A}'$.

(ii) Let $\mathfrak{C}$ be a class of models. Then $\mathfrak{A} = \langle \mathcal{A}, \mathcal{I} \rangle$ is an *optimal* model in $\mathfrak{C}$ iff $\mathfrak{A} \in \mathfrak{C}$ and there is no model in $\mathfrak{C}$ that is more normal than $\mathfrak{A}$.

(iii) Let Δ be a set of sentences. Then $\Delta \models_c \varphi$ iff φ is true in all optimal models of Δ.

Notice that in (i) of this definition nothing is said about the interpretation of ordinary predicates. $\mathfrak{A}$ can be at least as normal as $\mathfrak{A}'$, while for all $P \in \mathcal{L}_0$, the interpretations $\mathcal{I}(P)$ and $\mathcal{I}'(P)$ are totally different. However, in practice we are always looking for the most normal models within a given class $\mathfrak{C}$, and it may very well happen that within $\mathfrak{C}$ the interpretation of the ordinary predicates is heavily constrained or even fixed.

If $\Delta \models_c \varphi$, we say that φ follows by circumscription from Δ. Here is an example.

premise 1	*Adults normally have a bank account*
premise 2	*Adults normally have a driver's license*
premise 3	*John is an adult*
premise 4	*John does not have a driver's license*
by default	*John is an adult with a bank account*

This can be formalized as

premise 1	$\forall x((Ax \wedge \neg Ab_{AB}\, x) \rightarrow Bx)$
premise 2	$\forall x((Ax \wedge \neg Ab_{AD}\, x) \rightarrow Dx)$
premise 3	Aj
premise 4	$\neg Dj$
by circumscription	Bj

It is easy to check that the conclusion Bj follows by circumscription from the premises.

This example illustrates why the abnormality predicates have a double index referring to both the antecedent and the consequent of the rule, rather than a single one referring to just the antecedent. It is not sufficient to distinguish between normal and abnormal A's, and formalize a sentence like *Adults normally have a bank account* as $\forall x((Ax \wedge \neg Ab_A\, x) \rightarrow Bx)$. The distinction has to be more fine grained. An object with the property A can be a normal A in some respects and an abnormal A in other. Even though John is an abnormal adult in not having a driver's license, he is a normal adult in having a bank account, or at least we want to be able to conclude by default that he is. If we had formalized the argument in the following way, we would not have gotten very far.[4]

premise 1	$\forall x((Ax \wedge \neg Ab_A\, x) \rightarrow Bx)$
premise 2	$\forall x((Ax \wedge \neg Ab_A\, x) \rightarrow Dx)$
premise 3	Aj
premise 4	$\neg Dj$

Let us now look at the case of conflicting defaults introduced at the end of section 1. The formalized version looks like this:

premise 1	$\forall x(Ax \rightsquigarrow Ex)$
premise 2	$\forall x(Sx \rightsquigarrow \neg Ex)$
premise 3	$\forall x(Sx \rightsquigarrow Ax)$
premise 4	$Ac \wedge Sc$
by circumscription	$\neg Ec$

Unfortunately, in this simple set up the conclusion $\neg Ec$ does not follow from the premises. We find two kinds of optimal models: in some the sentences $\neg Ab_{SA}\, c$, $\neg Ab_{S\neg E}\, c$, and $Ab_{AE}\, c$ hold, which is fine, but in the other the sentences $\neg Ab_{SA}\, c$, $Ab_{S\neg E}\, c$, and $\neg Ab_{AE}\, c$ are true.

Recall that in the informal discussion of this example it was suggested that the three default rules involved together imply that objects with property S are *exceptional* A's; normal A's have the property E, but normal S's don't, even though normal S's do have property A.

In the next section we will see how one can enforce that in all models in which these three defaults hold, also the formula $\forall x(Sx \rightarrow Ab_{AE}\, x)$ will be true. Once we have this, the only optimal models will be models in which $\neg Ab_{SA}\, c$, $\neg Ab_{S\neg E}\, c$, and $\neg Ab_{AE}\, c$ are true, which means that the conclusion follows.

[4] The first to point this out was John McCarthy in [2].

The circumscriptive theory developed below differs from other circumscriptive theories in various ways. We will illustrate these differences by comparing our set up with the set up of the theory developed in [4] and [5] by Bonatti et al. for description logic.[5]

Bonatti et al. introduce abnormality predicates with only one index, and write for example $\mathtt{Whale} \sqsubseteq \mathtt{Ab}_{\mathtt{mammal}}$ to indicate that whales are abnormal mammals. As we pointed out above, this way it gets difficult, if not impossible, to deal properly with different – independent – default properties of the same kind of objects. Unfortunately, the paper does not discuss this problem.[6]

Secondly, the strategy Bonatti et al. use to deal with cases of conflicting defaults differs from the strategy we will use. They introduce a priority order $\prec$ on the set of abnormality predicates, the idea being that if $\mathtt{Ab}_{\mathtt{A}} \prec \mathtt{Ab}_{\mathtt{B}}$, minimizing $\mathtt{Ab}_{\mathtt{A}}$ has preference over minimizing $\mathtt{Ab}_{\mathtt{B}}$. It is not really clear how this priority ordering comes about. "The user can specify priorities between minimized predicates" is one thing Bonatti et al. say about this, but they add that these priorities normally reflect the *Specificity Principle*: if $\mathtt{A}$ is a more specific concept than $\mathtt{B}$, then $\mathtt{Ab}_{\mathtt{A}} \prec \mathtt{Ab}_{\mathtt{B}}$.

The 'if' in the last sentence cannot be an 'if and only if'. Specificity is not all that matters in deciding which defaults are applicable in a given situation. Consider for example

premise 1	$\forall x(Sx \rightsquigarrow Ax)$
premise 2	$\forall x(Ax \rightsquigarrow \neg Sx)$
premise 4	Sc
by default	?

Read 'Sx as 'x is a student' and 'Ax as 'x is an adult'. It is impossible that both $\neg Ab_{SA}\, c$ and $\neg Ab_{A\neg S}\, c$ hold. Intuitively, minimizing Ab_{SA} should have priority in this case, enabling the conclusion that Ac is presumably true. But S is not more specific than A – not at least in the sense that Bonatti et al. give to this phrase. On their account, concept $\mathtt{S}$ is more specific than $\mathtt{P}$ if the extension of $\mathtt{S}$ is a subset of the extension of $\mathtt{P}$ in all relevant models (i.e. all models the knowledge base allows). But in the example Sx does not strictly imply Ax.

Unlike Bonatti et al., we think that priority questions cannot be left to the user. It is not a pragmatic matter to decide which defaults apply in which circumstances. It's not something to decide *ad hoc*. It is a matter of semantics. That some of the arguments discussed in this paper are valid and other arguments are not, is because '*normally*' means what it means. We think of '*normally*' as a logical constant next to '*not*', '*necessarily*', '*sometimes*' etc., but differing from these in that its properties can only be described in a non-monotonic system.

Our strategy in the following fits in better with the second suggestion of Bonatti et al. We will specify an alternative to the specificity principle, and investigate the logic it generates.

3. Exemption and inheritance

There are two kinds of rules, rules that allow for exceptions and rules that do not allow for exceptions. So far we talked only about the first kind, but we also want to discuss the second kind. In order to do so, sentences of the form $\forall x(\varphi(x) \rightarrow \psi(x))$ can get a special status as *strict rules*. These strict rules are to be distinguished from universal sentences that are only accidentally true, and they will be treated differently.[7]

The general set up will be this: Let Σ be a finite set of default rules and strict rules and Π be a set of sentences. Think of $I = \langle \Sigma, \Pi \rangle$ as the *information* of some agent at some time, where Σ is the set of rules the agent is acquainted with, and Π the agent's factual information. We correlate with I a pair $\langle \mathcal{U}_I, \mathcal{F}_I \rangle$, and call this the (information) *state generated by* I. $\mathcal{U}_I$ is called the *universe* of the state. The elements of $\mathcal{U}_I$ are models of Σ, but not all models of Σ are allowed. The universe $\mathcal{U}_I$ must satisfy some additional *constraint* that will be discussed below. $\mathcal{F}_I$ consists of all models in $\mathcal{U}_I$ that are models of Π.

In this set up validity is defined as follows:

$$\Sigma, \Pi \models_d \varphi \text{ iff for all optimal models } \mathfrak{A} \in \mathcal{F}_I, \mathfrak{A} \models \varphi.$$

Read '$\Sigma, \Pi \models_d \varphi$' as '$\varphi$ follows by default from Σ and Π'.

To explain the constraint, it is necessary to introduce some technical notions.

[5] The expressive power of the languages of description logic differs from the expressive power of monadic first order languages, but the differences are not relevant for our discussion.

[6] Dealing properly with independent default properties is also a problem for theories that analyze the default implication *if …, then normally …* as a variable strict conditional. Here Delgrande [6], Asher and Morreau [7], and Boutilier [8] can serve as examples. The only modal theory we know of that gets this right is the one presented in Veltman [9].

[7] It is tempting to introduce a necessity operator in the object language to distinguish rules from accidental statements. We resist this temptation, making the distinction only at a meta-level, because we want to stay as closely as possible to the original circumscriptive framework.

Definition 3.1 *(Complying).*

(i) Suppose $\mathfrak{A} \models \forall x(\varphi(x) \rightsquigarrow \psi(x))$, and let d be an element of the domain[8] of $\mathfrak{A}$. Then d *complies with* $\forall x(\varphi(x) \rightsquigarrow \psi(x))$ (in $\mathfrak{A}$) iff d does not satisfy $Ab_{\varphi\psi}x$.
(ii) Let Σ be a set of rules, and let d be some element of the domain of some model $\mathfrak{A}$ of Σ. Then d *complies with* Σ (in $\mathfrak{A}$) iff d complies with all the default rules in Σ.

So, if an object satisfying $\varphi(x)$ complies with $\forall x(\varphi(x) \rightsquigarrow \psi(x))$, it will also satisfy $\psi(x)$. But notice that the definition allows for the following situations:

- The object d complies with $\forall x(\varphi(x) \rightsquigarrow \psi(x))$, but d does not satisfy $\varphi(x)$.
- The object d satisfies $\varphi(x)$ and $\psi(x)$, but d does not comply with $\forall x(\varphi(x) \rightsquigarrow \psi(x))$.

We will present examples later on. For now, just take 'comply' as a technical term.

3.1. The exemption principle

The constraint we will impose on $\mathcal{U}_I$ is motivated by the following minimal requirement.

If the *only* information about some object is that it has property P, it must be valid to infer by default that this object complies with all the default rules for objects with property P.[9]

What would be the use of these rules if they would not at least allow this inference?
It may seem easy to satisfy this requirement, but it is not.

Definition 3.2.

(i) An exemption clause is a formula of the form $\forall x(\varphi(x) \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta x)$.
Here we write 'Ab_δ' to refer to the abnormality predicate of the default rule δ. Δ can be any finite set of default rules. So, $\bigvee_{\delta\in\Delta} Ab_\delta x$ is a disjunction of a finite number of abnormality predicates.[10]
(ii) Let Σ be a set of rules. $\Sigma^{\varphi(x)}$ is the set of rules in Σ with antecedent $\varphi(x)$.
(iii) The exemption clause $\forall x(\varphi(x) \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta x)$ is an exemption clause *for* Σ iff $\Sigma \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta\in\Delta\cup\Sigma^{\varphi(x)}} Ab_\delta x)$.[11]

To see how these definitions work, consider again

$$\Sigma = \{\forall x(Ax \rightsquigarrow Ex), \forall x(Sx \rightsquigarrow \neg Ex), \forall x(Sx \rightsquigarrow Ax)\}.$$

Here $\Sigma^{Sx} = \{\forall x(Sx \rightsquigarrow Ax), \forall x(Sx \rightsquigarrow \neg Ex)\}$. Let $\Delta = \{\forall x(Ax \rightsquigarrow Ex)\}$. Clearly, there is no model such that some object in its domain satisfies Sx and complies with $\Delta \cup \Sigma^{Sx}$. So,

$$\Sigma \models \forall x(Sx \rightarrow \bigvee_{\delta\in\Delta\cup\Sigma^{Sx}} Ab_\delta x).$$

By *(iii)* above this means that $\forall x(Sx \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta x)$, i.e. $\forall x(Sx \rightarrow Ab_{AE}x)$, is an exemption clause for Σ, the idea being that objects with property S are, so to speak, *exempted* from the rule that A's are normally E.

The word 'exempted' suggests that default rules are some kind of normative rules. Indeed, often it is helpful to think of them that way. The use of the word 'normally', already suggests that we are dealing with a kind of norms here. To count as a normal S, S's must be A, and to count as a normal A, A's must be E, but here an exception is made for the S's. S's must be A, but they do not have to be E, they are not subjected to this rule. Actually, for them the opposite holds, to count as a normal S they must be not E.

'Being exempted from a rule' does not mean 'being an exception to the rule': The S's don't have to be E, but this does not mean they are, in fact, not E.

In the following definition it is made explicit for any set of rules Σ which kinds of objects are exempted from which rules in Σ.

---

[8] Let $\mathfrak{A} = \langle \mathcal{A}, \mathcal{I}\rangle$ be a model. Usually, the set $\mathcal{A}$ is called the universe of $\mathfrak{A}$, but since that phrase is already in use for something else we refer to $\mathcal{A}$ as the *domain* of $\mathfrak{A}$.

[9] The earliest place we know where this requirement is explicitly stated is on page 63 in Geffner [10]: "Given the evidence $E = \{p\}$ we can apply a default $p \rightarrow q$ *even in the presence of sets of defaults in conflict with* $p \rightarrow q$."

[10] By definition, if $\Delta = \emptyset$, $\bigvee_{\delta\in\Delta} Ab_\delta x = \bot$.

[11] Here and elsewhere '$\Gamma \models \psi$' means that Γ entails ψ in classical logic.

Definition 3.3. Let Σ be a set of rules, and let Π be an arbitrary set of formulas.

(i) The *exemption extension* $\Sigma^{\in}$ of Σ is given by

$$\Sigma^{\in} = \bigcup_{n=0}^{\infty} \Sigma_n^{\in}$$

where $\Sigma_0^{\in} = \Sigma$ and $\Sigma_{n+1}^{\in} = \Sigma_n^{\in} \cup \{\varphi \mid \varphi \text{ is an exemption clause for } \Sigma_n^{\in}\}$.

(ii) The state generated by $I = \langle \Sigma, \Pi \rangle$ is the state $\langle \mathcal{U}_I, \mathcal{F}_I \rangle$ given by

(a) $\mathfrak{A} \in \mathcal{U}_I$ iff $\mathfrak{A}$ is a model of $\Sigma^{\in}$;

(b) $\mathcal{F}_I$ consists of all models in $\mathcal{U}_I$ that are models of Π.

Notice that $\Sigma^{\in}$ has the following property, which we will call the *Exemption Principle*.

$$\text{If } \Sigma^{\in} \models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)}} Ab_\delta x), \text{ then } \Sigma^{\in} \models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta} Ab_\delta x).$$

$\Sigma^{\in}$ is defined inductively. By adding exemption clauses to a set of rules a new set of rules is created which may generate new exemption clauses, etc.[12] Given that Σ is finite, there are only finitely many possible exemption clauses. So, after finitely many steps a fixed point is reached. $\Sigma^{\in}$ is the weakest extension of Σ for which the exemption principle holds.

On the face of it, the exemption principle is not very strong. It just says that if the rules for objects with the property expressed by $\varphi(x)$ are incompatible with the rules in some set Δ, then every object with the property expressed by $\varphi(x)$ is exempted from at least one of the rules in Δ.

Without the exemption principle the minimal requirement cannot be met. If the principle does not hold, then there exists an optimal model $\mathfrak{A}$, a property $\varphi(x)$, and an entity d in the domain of $\mathfrak{A}$ such that d satisfies $\varphi(x)$, but d does not comply with at least one of the rules in $\Sigma^{\varphi(x)}$, because instead d complies with a rule that is not compatible with the rules in $\Sigma^{\varphi(x)}$.

With the exemption principle this cannot happen.

Proposition 3.4 *(Minimal requirement). Suppose* $\forall x(\varphi(x) \rightsquigarrow \psi(x)) \in \Sigma$. *Then* $\Sigma, \{\varphi(c)\} \models_d \psi(c)$.

Proof. To determine whether $\Sigma, \{\varphi(c)\} \models_d \psi(c)$, we have to look at the state $\langle \mathcal{U}_I, \mathcal{F}_I \rangle$ generated by $I = \langle \Sigma, \{\varphi(c)\} \rangle$ and check that every optimal model in $\mathcal{F}_I$ has the property that the object named c complies with $\Sigma^{\varphi(x)}$.

Recall that $\mathcal{U}_I$ contains the models of $\Sigma^{\in}$, and that $\mathcal{F}_I$ contains all models in $\mathcal{U}_I$ in which $\varphi(c)$ is true.

Now, consider any model $\mathfrak{A} = \langle \mathcal{A}, \mathcal{I} \rangle$ in $\mathcal{F}_I$ in which the object $\mathcal{I}(c)$ does not comply with $\Sigma^{\varphi(x)}$. We will show that $\mathfrak{A}$ is not optimal.

Let Δ be the set of defaults in $\Sigma^{\in}$ with which $\mathcal{I}(c)$ complies. Apparently, $\Sigma^{\in} \not\models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta} Ab_\delta x)$. By the exemption principle this means that $\Sigma^{\in} \not\models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)}} Ab_\delta x)$. Hence, there exists a model $\mathfrak{A}' = \langle \mathcal{A}', \mathcal{I}' \rangle$ in $\mathcal{U}_I$ such that some element d_0 in $\mathcal{A}'$ satisfies $(\varphi(x) \wedge \neg \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)}} Ab_\delta x)$.

Now, let $\mathfrak{A}'' = \langle \mathcal{A}'', \mathcal{I}'' \rangle$ be defined as follows:

- $\mathcal{A}'' = \mathcal{A}$;
- For individual constants a, $\mathcal{I}''(a) = \mathcal{I}(a)$;
- For P an ordinary predicate or an abnormality predicate,
  if $d \neq \mathcal{I}(c)$, then $d \in \mathcal{I}''(P)$ iff $d \in \mathcal{I}(P)$, and
  if $d = \mathcal{I}(c)$, then $d \in \mathcal{I}''(P)$ iff $d_0 \in \mathcal{I}'(P)$.

Consider any quantifier-free formula $\theta(x)$ in which no individual constant occurs. Clearly, if $d \neq \mathcal{I}''(c)$, then d satisfies $\theta(x)$ in $\mathfrak{A}''$ iff d satisfies $\theta(x)$ in $\mathfrak{A}$, while $\mathcal{I}''(c)$ satisfies $\theta(x)$ in $\mathfrak{A}''$ iff $I'(c)$ satisfies $\theta(x)$ in $\mathfrak{A}'$.

Given that all sentences of $\Sigma^{\in}$ are of the form $\forall x \theta(x)$ with θ as described, $\mathfrak{A}''$ will be a model of $\Sigma^{\in}$. And clearly, $\mathfrak{A}''$ is more normal than $\mathfrak{A}$. Therefore, $\mathfrak{A}$ is not optimal. □

Example. Applying Proposition 3.4 we find that

$$\forall x(Ax \rightsquigarrow Ex), \forall x(Sx \rightsquigarrow \neg Ex), \forall x(Sx \rightsquigarrow Ax), Sc \models_d Ac \wedge \neg Ec.$$

[12] An example is given below, when the inheritance principle is discussed.

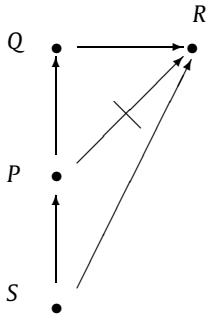
3.2. The inheritance principle

The exemption principle has some surprising consequences, one of which is the *inheritance principle*, which in its simplest form runs as follows:

> Let Σ be a set of rules. Suppose that $\Sigma^{\in} \models \forall x(\varphi(x) \rightsquigarrow \psi(x))$ and $\Sigma^{\in} \models \forall x(\psi(x) \rightarrow Ab_{\chi\theta}\, x)$. Then $\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow Ab_{\chi\theta}\, x)$.

To see how this works, consider the theory Σ consisting of the following five rules

$\forall x(Q\,x \rightsquigarrow R x)$
$\forall x(P x \rightsquigarrow Q\,x)$
$\forall x(P x \rightsquigarrow \neg R x)$
$\forall x(S x \rightsquigarrow P x)$
$\forall x(S x \rightsquigarrow R x)$



Consider the first three rules, and notice that the exemption principle enforces that $\forall x(P x \rightarrow Ab_{Q\,R}\, x) \in \Sigma^{\in}$. Now, it follows in one step from the inheritance principle that $\forall x(S x \rightarrow Ab_{Q\,R}\, x) \in \Sigma^{\in}$.

Without an appeal to the inheritance principle the proof is a bit longer. We give it here because it illustrates why Definition 3.3 (i) is inductive. Note first that $\forall x(P x \rightarrow Ab_{Q\,R}\, x) \in \Sigma_1^{\in}$. Given that $\models \forall x(S x \rightarrow (P x \vee \neg P x))$, it follows that $\Sigma_1^{\in} \models \forall x(S x \rightarrow (Ab_{Q\,R} x \vee Ab_{S\,P} x))$. But then $\forall x(S x \rightarrow Ab_{Q\,R}\, x) \in \Sigma_2^{\in}$.

By applying the exemption principle to the last three rules we find that $\forall x(S x \rightarrow Ab_{P\neg R}\, x) \in \Sigma^{\in}$. So, the S's, which in optimal circumstances are both P and Q, are exempted both from the rule that P's are normally not R and from the rule that Q's are normally R. They are normally R for independent reasons. As such, the example illustrates the fact that it is possible for an object not to comply with a rule whereas both the antecedent and the consequent of the rule hold for it. Objects with the property S do not comply with the rule $\forall x(Q\,x \rightsquigarrow R x)$, but in optimal circumstances they will have both the properties Q and R.

In its general form the inheritance principle runs as follows.

Proposition 3.5 *(Inheritance principle). Let $\langle \mathcal{U}_I, \mathcal{F}_I \rangle$ be the state generated by the information $I = \langle \Sigma, \Pi \rangle$. Let $\Delta \subseteq \Sigma$ be a set of default rules.*

Suppose

$$(a)\ \Sigma^{\in} \models \forall x(\varphi(x) \rightsquigarrow \psi(x))\ \textit{and}\ (b)\ \Sigma^{\in} \models \forall x(\psi(x) \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta\, x).$$

Then

$$\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta\, x).$$

Proof. By first-order logic alone, it is trivially true that

$$\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow (\psi(x) \vee \neg\psi(x))).$$

Given (a), objects that satisfy $\varphi(x)$ and $\neg\psi(x)$ will also satisfy $Ab_{\varphi\psi}\, x$. Thus, the above statement remains true when $\neg\psi(x)$ is replaced by $Ab_{\varphi\psi}\, x$. Similarly, given (b) the formula above remains true when $\psi(x)$ is replaced by $\bigvee_{\delta\in\Delta} Ab_\delta\, x$. Hence,

$$\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow (\bigvee_{\delta\in\Delta} Ab_\delta\, x \vee Ab_{\varphi\psi}\, x)).$$

Given the exemption principle this means

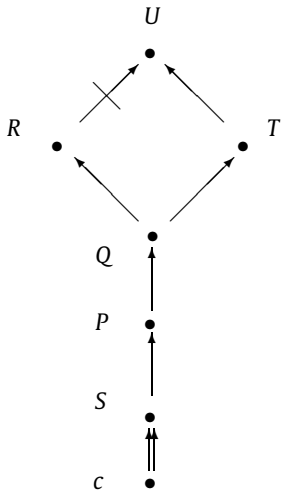
$$\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta\in\Delta} Ab_\delta\, x). \qquad \square$$

3.3. Some more examples

3.3.1. The kite

Given the inheritance principle it is easy to see why the following argument is valid.

$\forall x(Rx \rightsquigarrow \neg Ux)$
$\forall x(Tx \rightsquigarrow Ux)$
$\forall x(Qx \rightsquigarrow Tx)$
$\forall x(Qx \rightsquigarrow Rx)$
$\forall x(Px \rightsquigarrow Qx)$
$\forall x(Sx \rightsquigarrow Px)$
Sc
$\therefore Rc \wedge Tc$



Looking at the first four rules, we see that the exemption principle enforces that $\forall x(Qx \rightarrow (Ab_{R\neg U}\, x \vee Ab_{TU}\, x)) \in \Sigma^{\in}$. By applying the inheritance principle twice we see that $\forall x(Sx \rightarrow (Ab_{R\neg U}\, x \vee Ab_{TU}\, x)) \in \Sigma^{\in}$. So in all relevant models either $Ab_{R\neg U}\, c$ or $Ab_{TU}\, c$ is true. From this it follows that in all optimal models $\neg Ab_{S\neg P}$, $\neg Ab_{PQ}$, $\neg Ab_{QR}$, and $\neg Ab_{QT}$ are true, which enables the default conclusion that Pc, Qc, Rc and Tc.

Notice that on the naive account from section 2 none of these can be concluded. It would not even be possible to make the first step upwards from Sc to Pc. Here we can not only make this first step but also a second to Qc and further up to Rc and Tc. Only when we hit a direct conflict do we need to stop. By having the upper abnormalities propagate downward, we do not have to take into account potential abnormalities at the lower levels.

3.3.2. Defeasible modus tollens

Both *Defeasible Modus Ponens* and *Defeasible Modus Tollens* are valid.

$\forall x(Sx \rightsquigarrow Px)$	$\forall x(Sx \rightsquigarrow Px)$
Sc	$\neg Pc$
$\therefore Pc$	$\therefore \neg Sc$

The latter shows that an object need not have property S to count as an object that complies with the rule $\forall x(Sx \rightsquigarrow Px)$. Intuitively, if the object c had property S, it would be an abnormal S. So, assuming that the object c is normal and *complies* with the rule, it will not have property S.

3.3.3. Defeasible modus ponens beats defeasible modus tollens

Now, consider the following premises

premise 1 $\forall x(Sx \rightsquigarrow Px)$
premise 2 $\forall x(Px \rightsquigarrow \neg Sx)$
premise 3 Sc

At first sight one might be tempted to conclude Pc by *Defeasible Modus Ponens* and $\neg Pc$ by *Defeasible Modus Tollens*, but in fact the exemption principle enforces that $\forall x(Sx \rightarrow Ab_{P\neg S}\, x) \in \Sigma^{\in}$. This means that the only default conclusion to be drawn is Pc.

The reason we bring this up is that several authors have questioned the validity of *Defeasible Modus Tollens* with putative counterexamples like the following:

premise 1 *Men normally don't have a beard*
premise 2 *John has a beard*
by default *John is not a man*

However, the only thing this example shows is that one has to be very careful in providing 'intuitive' counterexamples when dealing with default arguments. One must be sure that the premises faithfully represent *all* one knows about the matter at issue.

In this case we know in fact more than the premises state. We know, for instance, that people with a beard normally are men. That's is why the conclusion sounds weird in the first place.

Now, if we state this explicitly as a third premise we get:

premise 1 *People with a beard normally are men*
premise 2 *Men normally don't have a beard*
premise 3 *John has a beard*

And as we saw, *Defeasible Modus Ponens* beats *Defeasible Modus Tollens*, so the only conclusion to be drawn is that *John is a man*.

Modus Tollens is closely related to *Contraposition*, but it is not the same. Modus Tollens, in the defeasible form discussed here, says: If all you know is the rule $\forall x(Sx \rightsquigarrow Px)$ and the fact $\neg Pa$, then it follows by default that $\neg Sa$. This is much weaker than saying that the rule $\forall x(Sx \rightsquigarrow Px)$ implies the rule $\forall x(\neg Px \rightsquigarrow \neg Sx)$. What we have at best is this: if all you know is the rule $\forall x(Sx \rightsquigarrow Px)$, it follows by default that $\forall x(\neg Px \rightsquigarrow \neg Sx)$. But it is important to realize that this conclusion expresses just an *accidental* truth, not a rule.[13] And notice that $\forall x(Sx \rightsquigarrow Px), Sa, \neg Pa \not\models_d \forall x(\neg Px \rightsquigarrow \neg Sx)$, so this accidental conclusion is easily defeasible, whereas in the same context we can still apply defeasible Modus Tollens: $\forall x(Sx \rightsquigarrow Px), Sa, \neg Pa, \neg Pb \models_d \neg Sb$.

3.4. Coherence

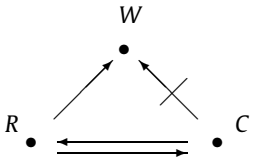
Every set Σ of default rules is consistent in the strict logical sense of the word.[14] This does not mean that every such set is okay. Here are some examples.

Consider

$$\Sigma = \{\forall x(Sx \rightsquigarrow Px), \forall x(Sx \rightsquigarrow \neg Px)\}.$$

Within our framework a theory of this form is of no use. Note that $\Sigma \models \forall x(Sx \rightarrow (Ab_{SP} \vee Ab_{S\neg P}))$. We can apply the exemption principle (take $\Delta = \emptyset$ and $\varphi(x) = Sx$) to find that $\forall x(Sx \rightarrow \bot)$ is an exemption clause for Σ. So, $\Sigma^{\in} \models \neg\exists x Sx$.

A more complicated example is this one:



$\forall x(Rx \rightsquigarrow Cx)$
$\forall x(Cx \rightsquigarrow Rx)$
$\forall x(Rx \rightsquigarrow Wx)$
$\forall x(Cx \rightsquigarrow \neg Wx)$

'Rainy days normally are cold', 'Cold days normally are rainy', 'On rainy days the wind is normally west', 'On cold days the wind is normally not west'. Something is wrong with this theory. By the exemption principle no such days exist: $\Sigma^{\in} \models \neg\exists x Rx$. Proof: note first that $\forall x(Cx \rightarrow Ab_{RW}x) \in \Sigma^{\in}$. By the inheritance principle it follows that $\forall x(Rx \rightarrow Ab_{RW}) \in \Sigma^{\in}$. Applying the exemption principle once more yields $\forall x(Rx \rightarrow \bot) \in \Sigma^{\in}$.

A third example is given by

$$\Sigma = \{\forall x(Sx \rightsquigarrow Px), \forall x((Sx \wedge Qx) \rightsquigarrow \neg Px), \forall x((Sx \wedge \neg Qx) \rightsquigarrow \neg Px)\}.$$

Again, this does not sound like an acceptable theory. Too many exceptions are being made. Note that $\Sigma \models \forall x((Sx \wedge Qx) \rightarrow (Ab_{(Sx\wedge Qx)\neg Px}x) \vee Ab_{SxPx}x)$. Hence, by the exemption principle $\forall x((Sx \wedge Qx) \rightarrow Ab_{SxPx}x) \in \Sigma^{\in}$; similarly, $\forall x((Sx \wedge \neg Qx) \rightarrow Ab_{SP}x) \in \Sigma^{\in}$. Hence, $\Sigma^{\in} \models \forall x(Sx \rightarrow Ab_{SP}x)$. But then $\forall x(Sx \rightarrow \bot) \in \Sigma^{\in}$.

The above leads to the following definition.

Definition 3.6. A set of rules Σ is coherent iff for every $\varphi(x)$ which is the antecedent of some rule in Σ, $\Sigma^{\in} \cup \{\exists x\varphi(x)\}$ is consistent.

A set of rules is incoherent if it is logically impossible to satisfy the minimal requirement. In such a case there is some property such that no object with this property can comply with all the rules for objects with this property. Given the exemption principle, no such objects are allowed.

As will become clear in due course, for inheritance networks we can give an exact syntactic characterization of the sets of rules that are incoherent.

3.5. Is the exemption principle all there is to it?

We have presented the exemption principle as a minimal constraint – a constraint that *must* be imposed on a circumscriptive theory to meet the minimal requirement. We have seen that this principle has many, sometimes surprising

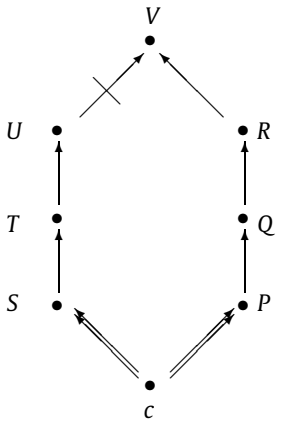
[13] Here, it would help if we had introduced a necessity operator $\Box$ in the object language. Then we could just say $\Box\forall x(Sx \rightsquigarrow Px) \models_d \forall x(\neg Px \rightsquigarrow \neg Sx)$, but $\Box\forall x(Sx \rightsquigarrow Px) \not\models_d \Box\forall x(\neg Px \rightsquigarrow \neg Sx)$.

[14] To see why this is so, consider a model in which all objects are abnormal in all respects.

consequences, and deals successfully with a lot of examples. So, naturally the question arises if this constraint is all one needs to settle questions of priority between conflicting defaults.

One might think that the next example shows that the answer is 'no'.[15] Consider the following set of premises.

$\forall x(Ux \rightsquigarrow \neg Vx)$
$\forall x(Rx \rightsquigarrow Vx)$
$\forall x(Tx \rightsquigarrow Ux)$
$\forall x(Sx \rightsquigarrow Tx)$
$\forall x(Qx \rightsquigarrow Rx)$
$\forall x(Px \rightsquigarrow Qx)$
$Sc \wedge Pc$



Here, one might want to conclude first that Tc and Qc – after all these are immediate default consequences of Sc and Pc – and then make a second step further up to Uc and Rc – why stop before we hit a direct conflict? This is how we reasoned when we discussed the kite example in section 3.3.1. Why wouldn't this kind of forward chaining go through here?

There is, in fact, a huge difference between this example and the example in section 3.3.1, which is best brought out by comparing the pictures. In example 3.3.1 we have one path – one line of argument – leading up from Sc to the conclusions Rc and Tc. Here we have for each of the properties P, Q, R, S, T, U, V two different paths – two completely independent lines of argument – one leading to the conclusion that c has the property concerned, the other leading to the conclusion that c does not have the property concerned.[16] There is no reason to think, given the information at hand, that the one line of argument leads to a more normal situation than the other. This means that no conclusion can be drawn here. Indeed, on our account it is not valid to conclude that Uc or Rc; it is not even valid to conclude that Tc or Qc.

So, we don't think this example shows that the exemption principle is too weak. But of course, the possibility is not excluded that somebody will come up with a more convincing example. Besides, also in the absence of such an example, it might be enlightening – perhaps only from a mathematical point of view – to investigate possible strengthenings of the exemption principle.

We have to leave this to another occasion. For now, we just content ourselves setting a new minimum – the old one being the specificity principle – to principles regulating the priority ordering between conflicting defaults.

4. Networks

Inheritance networks are, simply put, the kind of structures we have been picturing to illustrate the examples. Thus, an inheritance network is a directed graph in which the nodes represent properties and the arcs represent rules. More precisely:

Definition 4.1. An *inheritance network* is a pair $\langle V, \Sigma \rangle$, where each element of Σ is a combination of an ordered pair of elements of V and a *polarity* which may be positive, negative, strict positive or strict negative.

Elements of V are called nodes and elements of Σ are called arcs. We will refer to an arc from u to v as 'uv' if this arc is positive, as 'uv^-' if it is negative, as 'uv^*' if it is strict positive and as 'uv^{*-}' if it is strict negative.

All nodes except one represent properties. This one node stands for an object. There are no arcs to this node and only strict arcs from this node to other nodes. The object is supposed to have the properties represented by these nodes.

The definition above does not distinguish between nodes representing properties and the node representing an object. The difference is purely a matter of interpretation.

To describe inferences in these networks, the notion of a *path* is crucial.

Definition 4.2. Let $\langle V, \Sigma \rangle$ be an inheritance network, with $a, b \in V$.

(i) A *positive path* from a to b is a subset $\{\alpha_1, \ldots, \alpha_n\} \subseteq \Sigma$ such that there exist $v_1, \ldots, v_{n-1} \in V$ such that:

[15] The example was brought to our attention by one of the referees.

[16] For example, prima facie one wants to conclude Uc from Sc via $\forall x(Sx \rightsquigarrow Tx)$, and $\forall x(Tx \rightsquigarrow Ux)$, and $\neg Uc$ from Pc via $\forall x(Px \rightsquigarrow Qx)$, $\forall x(Qx \rightsquigarrow Rx)$, $\forall x(Rx \rightsquigarrow Vx)$, and $\forall x(Ux \rightsquigarrow \neg Vx)$.

- α_1 is a positive (or strict positive) arc from a to v_1;
- α_i is a positive (or strict positive) arc from v_{i-1} to v_i, where $1 < i < n$;
- α_n is a positive (or strict positive) arc from v_{n-1} to b.

Moreover, the empty set is considered a positive path from any $v \in V$ to itself.

(ii) $X \subseteq \Sigma$ is a *negative path* from a to b if there are X_1, X_2, a', b', α such that

- $X = X_1 \cup \{\alpha\} \cup X_2$;
- X_1 is a positive path from a to a';
- X_2 is a positive path from b to b';
- α is a negative (or strict negative) arc from a' to b', or from b' to a'.[17]

If there exists a positive (negative) path from a to b, this serves as prima facie evidence that objects with property a have (do not have) property b. Of course, in interesting examples we have prima facie evidence for both b and not b, which brings us to the next key notion: the *conflicting set*.

Definition 4.3. Where $\langle V, \Sigma\rangle$ is an inheritance network and $a \in V$, a subset $X \subseteq \Sigma$ is a *conflicting set relative to a* iff there is some $b \in V$ such that X contains both a positive and a negative path from a to b.

A conflicting set X is a *minimal* if no proper subset of X is a conflicting set relative to a.

Here 'minimal' does not mean having the least possible number of elements. It means that if more arcs were taken out, the set would no longer be conflicting.

4.1. Making inferences in inheritance networks

Let $\langle V, \Sigma\rangle$ be an inheritance network, and $u, v \in V$. We will write '$u \rightleftharpoons v$' to indicate that there is both a positive path from u to v and a positive path from v to u.

Definition 4.4. Where $\langle V, \Sigma\rangle$ is an inheritance network and $a \in V$, let $Ess_\Sigma(a)$ be set of arcs defined by

$$\begin{aligned}\alpha \in Ess_\Sigma(a) \text{ iff } \alpha \in \Sigma \text{ and (i) } &\alpha \text{ is strict,}\\ \text{or (ii) } &\alpha = uv \text{ for some } u \rightleftharpoons a,\\ \text{or (iii) } &\alpha = uv^- \text{ for some } u \rightleftharpoons a.\end{aligned}$$

For a given property a, the set $Ess_\Sigma(a)$ contains the rules that are essential for a, the rules from which the objects with property a cannot be exempted. No object can be exempted from any strict rule; the objects with property a cannot be exempted from any rule for objects with property a, and more generally, the objects with property a cannot be exempted from any rule for objects with a property u that is "default equivalent" to a.

Definition 4.5. Where $\langle V, \Sigma\rangle$ is an inheritance network and $a \in V$, let $d(a)$ be defined by

$$Y \in d(a) \text{ iff } Y = X - Ess_\Sigma(a) \text{ for some minimal conflicting set } X \text{ relative to } a.$$

The intuition is that the objects with property a are exempted from at least one rule in every set $Y \in d(a)$.

The inheritance principle comes in by letting the d function propagate backwards along positive paths, collecting d-sets in the D function defined below.

Definition 4.6. Where $\langle V, \Sigma\rangle$ is an inheritance network and $a \in V$, let $D(a)$ be defined by

$$Y \in D(a) \text{ iff } Y \in d(b) \text{ for some } b \text{ for which there is a positive path from } a \text{ to } b.$$

We are now ready to define the consequence relation. This will be done in terms of *exception sets*, sets of arcs representing rules to which an exception must be made.

Definition 4.7. Let $\langle V, \Sigma\rangle$ an inheritance network and $a \in V$.

$X \subseteq \Sigma$ is an *acceptable exception set for a* iff $X \cap Y \neq \emptyset$ for all $Y \in D(a)$.

Such an X is *minimal* if every proper subset of X is not an acceptable exception set for a.

[17] Note that it's possible that $a = a', b = b'$ and X_1 and X_2 are empty.

Each minimal exception set represents a way to make as few exceptions as possible. A given conclusion b follows from a in a network if b can be reached from a under each of these ways.

Definition 4.8. Let $\langle V, \Sigma \rangle$ be an inheritance network. Let $a, b \in V$.

- $a \vdash_\Sigma b$ iff for every minimal exception set X for a there is a positive path Y from a to b such that $X \cap Y = \emptyset$.
- $a \vdash_\Sigma \neg b$ iff at least one of the following is true:
  (i) For every minimal exception set X for a there is a negative path Y from a to b such that $X \cap Y = \emptyset$.
  (ii) No minimal exception set X for a is also an acceptable exception set for b.

We did not prepare the reader for the second clause of negative entailment. It is there for the special case in which there is no path from a to b. In such a case it may happen that objects with property b are so abnormal that one can safely assume that objects with property a[18] do not have property b.

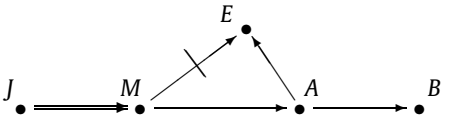
To see how this works, notice first that if there is a path from a to b, $D(b) \subseteq D(a)$. So in that case every minimal exception set for a is an acceptable exception set for b.[19] Now, if there is no path from a to b, and some minimal exception set for a is an acceptable exception set for b, this means that one has to reckon with the possibility that objects with property b are in all respects at least as normal as objects with property a. However, if no minimal exception set is an acceptable exception set for b, this means that the a's are normal in some respect in which the b's are exceptional. Assuming that the a's are as normal as possible, one may in such a case by default infer that the a's don't have the property b.[20]

4.2. Examples

4.2.1. As a first example, we consider the following desirable inference.

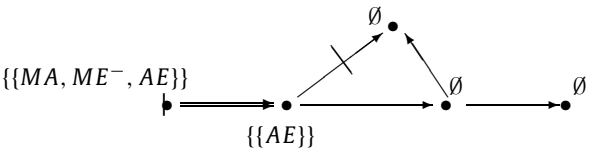
premise 1	*Adults normally have a bank account*
premise 2	*Master students are normally adults*
premise 3	*Master students are normally not employed*
premise 4	*Adults are normally employed*
premise 5	*John is a master student*
by default	*John is an adult with a bank account, but he is not employed*

Rendered as an inheritance network, the premises look like this.



Our first step is to determine the d function. Since there are no conflicting sets relative to A, B, and E, we have $d(A) = d(B) = d(E) = \emptyset$. The conflicting sets relative to master student are $\{MA, ME^-, AE\}$ and $\{AB, MA, ME^-, AE\}$. Only the first of these is minimal. Since $Ess_\Sigma(M) = \{MA, ME^-, JM\}$, we obtain $d(M) = \{\{AE\}\}$.

Similarly, there is a single minimal conflicting set relative to J: the set $\{MA, ME^-, AE, JM\}$. We have $Ess_\Sigma(J) = \{JM\}$, so $d(J) = \{\{MA, ME^-, AE\}\}$.



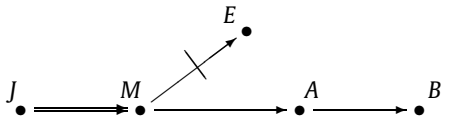
We can now determine $D(J)$. Since there is a positive path from J to every other node, $D(J)$ is the union of all the d's. Only two are non-trivial, so $D(J) = \{\{MA, ME^-, AE\}, \{AE\}\}$.

Since $\{AE\} \in D(J)$, every acceptable exception set for John will contain arc AE. Since $\{AE\}$ is itself an acceptable exception set, this makes it the only minimal exception set. Thus, a conclusion is acceptable iff there is a path from J to it that does not use arc AE. That is, if there is a path in the following network.

[18] a could be the individual node, but this does not change the story. (It's the special case in which there is only one object with property a.)

[19] The converse does not hold. For example, in a network without conflicts $\emptyset$ is a minimal exception set – actually, the only one – for every node.

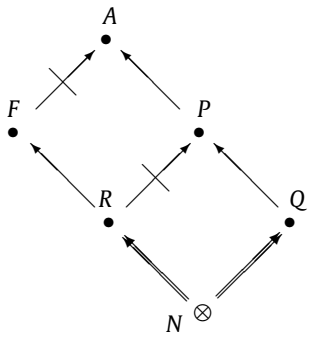
[20] Admittedly, we might not have come up with this second clause, if we had not been looking for an algorithm covering the circumscriptive validities. In all the inheritance algorithms we know of it is a necessary condition for $b/\neg b$ to follow from a that there is a positive/negative path from a to b.



Therefore as desired we obtain $J \vdash_\Sigma \neg E$, $J \vdash_\Sigma A$, $J \vdash_\Sigma B$.

4.2.2. The Double Diamond

The following network is a well-known extension of the Nixon Diamond, generally referred to as the Double Diamond.



premise 1 *Nixon is a Republican and a Quaker*
premise 2 *Quakers are normally Pacifist*
premise 3 *Republicans are normally not Pacifist*
premise 4 *Republicans are normally Football fans*
premise 5 *Pacifists are normally Anti-military*
premise 6 *Football fans are normally not Anti-military*

The question is whether Nixon is Anti-military.

In preemption based approaches (notably Horty et al. [3]), the positive path from N to A is disabled by the negative path from N to P, so that $\neg A$ may be concluded. This outcome is considered counterintuitive since the negative path to A is itself disabled by its positive counterpart, which is why such paths are referred to as *zombie paths*. (See Makinson & Schlechta [11].)

Since our own approach is not based on this kind of preemption, we can do a bit better here.

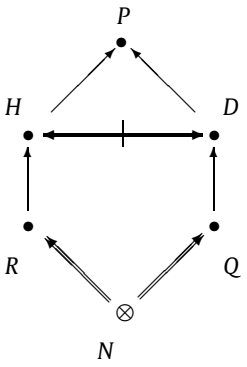
The first thing to notice is that there are no pairs of conflicting paths starting at P, F, A, R, or Q. Therefore, all of them have empty d, and $D(N) = d(N)$. We subsequently find that $D(N) = \{\{Q\,P, RP^-\}, \{Q\,P, RF, PA, FA^-\}\}$. It is important to keep in mind that 'minimal exception set' does not mean 'exception set with the smallest amount of elements' – $\{Q\,P\}$ is not the only minimal exception set (of N) here. The others are $\{RP^-, RF\}$, $\{RP^-, PA\}$ and $\{RP^-, FA^-\}$.

We trivially obtain $N \vdash_\Sigma R$, $N \vdash_\Sigma Q$. But as to the other properties, nothing can be concluded, not even $N \vdash_\Sigma F$.

4.2.3. A floating conclusion

Also the next example is much discussed in the literature on inheritance networks.[21]

premise 1 *Nixon is a Republican and a Quaker*
premise 2 *Quakers are normally Doves*
premise 3 *Republicans are normally Hawks*
premise 4 *Nobody is both a hawk and a dove*
premise 5 *Hawks normally are politically motivated*
premise 6 *Doves normally are politically motivated*



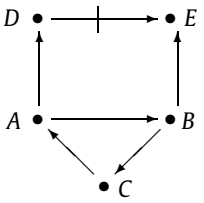
Does it follow that Nixon is politically motivated? According to the theory presented here, the answer to this question is 'yes'.[22] It is easy to see that $D(N) = d(N)$ and that $d(N) = \{\{RH, Q\,D\}\}$. This means there are two minimal exception sets of N, namely $\{RH\}$ and $\{Q\,D\}$. The exception set $\{RH\}$ does not contain any element of the rightmost path from N to P, and the exception set $\{Q\,D\}$ does not contain any element of the leftmost path from N to P. Thus, for each minimal exception set there is a positive path from N to P which does not contain any element of that set. Therefore, $N \vdash_\Sigma P$.

[21] The example has been around since 1987, and is due to Ginsberg, but he did not publish it in print until 1993 in [12].

[22] This is what most people working in this field want. Horty [13] provides a counterexample, but it concerns normative rules rather than defaults. See Prakken [14] for an insightful discussion.

4.2.4. Closed loops

The theory covers inheritance networks with cyclic paths. Here is an example.



premise 1	*A*'s are normally *B*
premise 2	*B*'s are normally *C*
premise 3	*C*'s are normally *A*
premise 4	*A*'s are normally *D*
premise 5	*D*'s are normally not *E*
premise 6	*B*'s are normally*E*
premise 7	*x* is *A*
by default	*x* is *E*

At first glance, one might expect $d(A) = \{\{DE^-, BE\}\}$. However, this is not the case.

Since all points of the loop must be taken into account, we have $Ess_\Sigma(A) = \{AD, AB, BE, BC, CA\}$. Therefore the conflicting set $X = \{AB, AD, DE^-, BE\}$ leads to the inclusion of $X - Ess_\Sigma(A) = \{DE^-\}$ in $d(A)$, rather than the inclusion of $\{DE^-, BE\}$. Thus, E may be validly concluded when starting at A, B or C.

4.3. Completeness

We have been using pictures of networks to illustrate the examples in the sections on circumscription. So it will come as no surprise that the inheritance networks can be translated into the monadic first order language we discussed there. More surprising, and a lot less trivial to show, is the fact that this translation preserves logical validity: the consequence relation $\vdash$ specified for the network-based approach, covers exactly (for what it can express) the validity notion $\models_d$ of the circumscriptive framework.[23]

We provide the translation and the formal statement here. For the proof, see Appendix A.

Definition 4.9. Let $N = \langle V, \Sigma\rangle$ be an inheritance network, and suppose $V = \{v_1, \ldots, v_n\}$. We associate with every $v_i \in V$ a predicate P_i, and with every arc $\alpha \in \Sigma$ a rule given by

$$
\begin{aligned}
v_i v_j^{\uparrow} &= \forall x(P_i x \rightsquigarrow P_j x)\\
v_i {v_j}^{-\uparrow} &= \forall x(P_i x \rightsquigarrow \neg P_j x)\\
v_i v_j^{*\uparrow} &= \forall x(P_i x \rightarrow P_j x)\\
v_i v_j^{*-\uparrow} &= \forall x(P_i x \rightarrow \neg P_j x)
\end{aligned}
$$

We will call $\Sigma^\uparrow = \{\alpha \mid \alpha \in \Sigma\}$ the *lift* of N.[24]

Theorem 4.10 *(Soundness–completeness theorem). Let $N = \langle V, \Sigma\rangle$ and $\Sigma^\uparrow$ be as in the definition. Suppose $\Sigma^\uparrow$ is coherent. Then $v_i \vdash_\Sigma v_j$ if and only if $\Sigma^\uparrow, \{P_i c\} \models_d P_j c$, and $v_i \vdash_\Sigma \neg v_j$ if and only if $\Sigma^\uparrow, \{P_i c\} \models_d \neg P_j c$.*

The theorem only holds for coherent theories. Therefore it is desirable know what the notion of coherence amounts to in terms of inheritance networks. This is where the following theorems come in. Again, the proofs can be found in Appendix A.

Theorem 4.11. *Let $\Sigma^\uparrow$ be the lift of the inheritance network $\langle V, \Sigma\rangle$. Then $\Sigma^\uparrow$ is coherent if and only if there is no $v \in V$ with $\emptyset \in d(v)$.*

The following definition and proposition describe what a network looks like when $\emptyset \in d(v)$ for some $v \in V$.

Definition 4.12. The node x *semi-strictly implies* (*semi-strictly refutes*) y if there is a positive (negative) path from x to y where every arc after the first is strict.

[23] In this paper the starting point was the notion $\models_d$ and we have been looking for a consequence relation $\vdash$ on networks matching this notion. It would be interesting to go the other way around and represent, say, the network algorithm developed in Thomason, Horty & Touretzky [15] in circumscriptive terms. We have tried to do so, but so far failed. Still, we think it should be possible. In [16] Bochman develops a theory in the framework of Reiter's Default Logic [17] matching the network algorithm developed in Thomason et al. [15] (as presented in Horty [18]). In Qian & Irani [19] a policy is given for representing default theories developed in Reiter's framework in a circumscriptive set up. So, maybe combining the insights of these two papers will do the job.

[24] Since networks do not distinguish between individuals and properties, the lift will represent individuals with predicates. A premise like '*John is an adult*', which in the circumscription framework could be represented as Aj, is represented in an inheritance network as a strict arc from J to A, the lift of which is $\forall x(Jx \rightarrow Ax)$. This is clumsy, but, fortunately, it does not lead to any real problems.

Proposition 4.13. *Let $\langle V, \Sigma\rangle$ be an inheritance network. Then $\emptyset \in d(v)$ iff there is some z and some $y \rightleftharpoons v, x \rightleftharpoons v$ such that y semi-strictly implies z and x semi-strictly refutes z.*

5. Conclusion

In the above we have studied the logical properties of defaults, or more particularly of sentences of the form *S's are normally P*. We have shown that most, if not all of their capricious logical behavior can be explained on the basis of one simple underlying principle that determines in cases of conflicting defaults which objects are *exempted* from which rules. We have built two theories based on this principle, one within a circumscriptive framework, and the other in terms of inheritance networks. In the appendix we will prove a completeness theorem showing that arguments that can be expressed in both systems are valid on the one account iff they are valid on the other.

Despite the length of this paper, we have only taken the first steps developing these systems. Undoubtedly, a more systematic *model theoretic* study of the circumscriptive part will result in a shorter and more elegant proof of the completeness theorem. We also think that on the *algorithmic* side further investigations may yield simplifications. For example, it will become clear in the appendix that things get a lot less complicated if the networks do not have cycles. Finally, a study like this should be complemented by a study answering the question under which conditions a set of default rules can be safely adopted as a guiding line for taking decisions. Maybe this is a question for methodologists rather than for logicians, but the answer is important to everybody interested in common sense reasoning.

Appendix A. Completeness of the network algorithm

When defining the d and D functions we suggested that d models the effect of a weakened version of the exemption principle and D the effect of the inheritance principle. Before starting with the completeness proof proper, we will prove this explicitly.

Things are more complicated for networks with cycles than for networks without cycles. Therefore, we first concentrate on the latter.

A.1. Reformulating the principles

Definition A.1. The *weak* exemption extension Σ^w of Σ is given by

$$\Sigma^w = \Sigma \cup \{\varphi \mid \varphi \text{ is an exemption clause for } \Sigma\}.$$

Note that $\Sigma^w = \Sigma^{\in}_1$. The weak exemption extension uses only one step of the definition of the regular exemption extension.

Definition A.2.

(i) The clause $\forall x(\varphi(x) \to \bigvee_{\delta\in\Delta} Ab_\delta x)$ is an inherited clause for Σ iff there is some ψ such that $\forall x(\varphi(x) \rightsquigarrow \psi(x)) \in \Sigma$ and $\Sigma \models \forall x(\psi(x) \to \bigvee_{\delta\in\Delta} Ab_\delta x)$.

(ii) The inheritance extension Σ^I of Σ is given by

$$\Sigma^I = \bigcup_{n=0}^{\infty} \Sigma^I_n$$

where $\Sigma^I_0 = \Sigma$ and $\Sigma^I_{n+1} = \Sigma^I_n \cup \{\varphi \mid \varphi \text{ is an inherited clause for } \Sigma^I_n a\}$.

Since the exemption principle implies the inheritance principle, the following holds.

Proposition A.3. $\Sigma^{\in} \models \Sigma^{wI}$.

How about $\Sigma^{wI} \models \Sigma^{\in}$? We doubt this holds for every Σ, but it does hold for the special case that Σ is the lift of a cycle-free inheritance network. Before we turn to the proof of this statement some more observations are needed.

The rules and exemption clauses figuring in the sets $(\Sigma^{\uparrow})^{wI}$ have a very specific syntactic form, which gives us a lot of freedom constructing models of such sets. For example, all the sentences concerned are universal, so every $(\Sigma^{\uparrow})^{wI}$ is preserved under submodels. Moreover, if the only difference between two models $\mathfrak{A}$ and $\mathfrak{A}'$ is that $\mathfrak{A}'$ has more abnormalities than $\mathfrak{A}$, then $\mathfrak{A}'$ will be a model of $(\Sigma^{\uparrow})^{wI}$ if $\mathfrak{A}$ is. This also holds if for some predicates P_i that do not occur in the consequent of any rule in $(\Sigma^{\uparrow})^{wI}$, the extension of P_i in $\mathfrak{A}'$ is a subset of the extension of P_i in $\mathfrak{A}$. More precisely:

Lemma A.4. *Let $\Sigma^{\uparrow}$ be the lift of an inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_m\}$. Let Γ consist of sentences of the form $\forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x)$.*

Let $\mathfrak{A} = \langle \mathcal{A}, \mathcal{I}\rangle$ and $\mathfrak{A}' = \langle \mathcal{A}', \mathcal{I}'\rangle$ be two models with the following properties:

(a) $\mathfrak{A} \models \Sigma^{\uparrow} \cup \Gamma$.
(b) $\mathcal{A} = \mathcal{A}'$.
(c) *For all individual constants c, $\mathcal{I}(c) = \mathcal{I}'(c)$.*
(d) *For all predicates P_i, the following holds:*
  (da) *If P_i does not occur in the consequent of any rule in $\Sigma^{\uparrow}$, then $\mathcal{I}'(P_i) \subseteq \mathcal{I}(P_i)$.*
  (db) *Otherwise, $\mathcal{I}'(P_i) = \mathcal{I}(P_i)$.*
(e) *For all P_i and P_j, $\mathcal{I}(Ab_{P_i P_j}) \subseteq \mathcal{I}'(Ab_{P_i P_j})$ and $\mathcal{I}(Ab_{P_i \neg P_j}) \subseteq \mathcal{I}'(Ab_{P_i \neg P_j})$.*

Then $\mathfrak{A}' \models \Sigma^{\uparrow} \cup \Gamma$.

Proof. Left to the reader.

On the way to the completeness theorem, we will often be looking for correspondences between notions that play a role in inheritance networks and notions in the circumscription framework. One such notion is the notion of a *path*.

Clearly, if there is a positive path X from v_i to v_j in the network $\langle V, \Sigma\rangle$, then $\Sigma^{\uparrow} \models \forall x((P_i x \wedge \bigwedge_{\alpha\in X} \neg Ab_\alpha x) \to P_j x)$.[25] For coherent theories the converse is also true. This follows immediately from the following more general proposition.

Lemma A.5. *Let $\Sigma^{\uparrow}$ be the lift of an inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_m\}$. Let Γ consist of sentences of the form $\forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x)$. Let $\varphi(x)$ be a quantifier-free formula in which all predicates are abnormality predicates, and such that for some P_i, $\Sigma^{\uparrow} \cup \Gamma \cup \{\exists x(P_i x \wedge \varphi(x))\}$ is consistent. If $\Sigma^{\uparrow} \cup \Gamma \models \forall x((P_i x \wedge \varphi(x)) \to P_j x)$, then there is a positive path from v_i to v_j.*

Proof. Let $\mathfrak{A}$ be a model of $\Sigma^{\uparrow} \cup \Gamma$ in which $P_i(c) \wedge \varphi(c)$ is true for some c and $Ab_\delta c$ is true for every δ except those where $\varphi(c) \models \neg Ab_\delta c$. Given Lemma A.4 such a model must exist.

Create $\mathfrak{A}'$ by adjusting $\mathfrak{A}$ so as to make $P_k c$ false whenever there is no positive path from v_i to v_k. This does not affect the truth of $P_i(c) \wedge \varphi(c)$.

Claim: $\mathfrak{A}'$ is still a model of $\Sigma^{\uparrow} \cup \Gamma$. This means that $P_j c$ is true in $\mathfrak{A}'$. Hence, there is a path from v_i to v_j.

The claim is proved by induction on the number of distinct consequents of rules in $\Sigma^{\uparrow}$.

Case $n = 0$: If there are no rules in $\Sigma^{\uparrow}$, then none of the predicates we make false is the consequent of such a rule. Therefore, $\mathfrak{A}'$ is a model of $\Sigma^{\uparrow} \cup \Gamma$ by Lemma A.4.

Induction hypothesis: If the number of distinct consequents occurring in the rules of $\Sigma^{\uparrow}$ is at most n, then $\mathfrak{A}'$ is a model of $\Sigma^{\uparrow} \cup \Gamma$.

Case $n + 1$: Assume that $\Sigma^{\uparrow}$ have $n + 1$ distinct consequents. Suppose towards contradiction that $\mathfrak{A}'$ is not a model of $\Sigma^{\uparrow} \cup \Gamma$, and consider some v_l such that $\mathfrak{A} \models P_l c$, $\mathfrak{A}' \not\models P_l c$.

Let $\Sigma^{\uparrow}_{-l}$ be $\Sigma^{\uparrow}$ with all rules where P_l is a consequent removed. By the Induction Hypothesis, $\mathfrak{A}'$ is a model of $\Sigma^{\uparrow}_{-l} \cup \Gamma$. Therefore, $\mathfrak{A}'$ contradicts some rule in $(\Sigma^{\uparrow} \cup \Gamma) - (\Sigma^{\uparrow}_{-l} \cup \Gamma)$. More specifically, there must be some k such that $\mathfrak{A}' \models P_k c, \neg P_l c$ and $\forall x(P_k x \rightsquigarrow P_l x) \in \Sigma^{\uparrow}$.

Since $\mathfrak{A}' \models P_k c$, by construction there is a positive path from v_i to v_k. The rule $\forall x(P_k x \rightsquigarrow P_l x)$ is the lift of an arc from v_k to v_l. Hence, there is a path from v_i to v_l.

But then $P_l c$ should not have been made false in constructing $\mathfrak{A}'$. This contradiction shows that $\mathfrak{A}'$ is a model of $\Sigma^{\uparrow} \cup \Gamma$. □

The next lemma shows that when Σ is the lift of a network, Σ^{wI} is 'closed' with regards to exemption clauses. That is: if Σ^{wI} entails some exemption clause, then this exemption clause or a stronger one is an element of Σ^{wI}.

Lemma A.6. *Let $\Sigma^{\uparrow}$ be the lift of an inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_m\}$. Consider $(\Sigma^{\uparrow})^{wI} = \Sigma^{\uparrow} \cup \Gamma$, and assume $(\Sigma^{\uparrow})^{wI} \cup \{\exists x P_i x\}$ is consistent. If $(\Sigma^{\uparrow})^{wI} \models \forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x)$, then there is some $\Delta' \subseteq \Delta$ such that $\forall x(P_i x \to \bigvee_{\delta\in\Delta'} Ab_\delta x) \in \Gamma$.*

Proof. By contraposition. Suppose there is no such Δ'. Then

$$\Sigma^{\uparrow} \not\models \forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x).$$

[25] We are a bit sloppy here. We should have written '$Ab_{\alpha^{\uparrow}}$' instead of 'Ab_α', because it concerns the abnormality predicate of the lift $\alpha^{\uparrow}$ of the arc α.

Take

$$\varphi(x) = \left(\bigwedge_{\delta\in\Delta} \neg Ab_\delta x\right) \wedge \left(\bigwedge_{\delta\notin\Delta} Ab_\delta x\right)$$

and let $\mathfrak{A}$ be a model with a single element c, where $P_i c$ and $\varphi(c)$ are true and $P_j c$ is true for only those P_j where $\Sigma^\uparrow \models \forall x((P_i x \wedge \varphi(x)) \to P_j x)$.

By construction $\mathfrak{A} \not\models \forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x)$, so it suffices to show that $\mathfrak{A}$ is a model of $(\Sigma^\uparrow)^{wI}$.

We will first show that $\mathfrak{A}$ is a model of $\Sigma^\uparrow$. Consider any rule $\forall x(P_k x \rightsquigarrow P_l x)$. Two possibilities occur. (i) $\Sigma^\uparrow \not\models \forall x((P_i x \wedge \varphi(x)) \to P_k x)$. In this case $\mathfrak{A} \not\models P_k c$, and since the object named c is the only element of the domain, trivially $\mathfrak{A} \models \forall x(P_k x \rightsquigarrow P_l c)$. (ii) $\Sigma^\uparrow \models \forall x((P_i x \wedge \varphi(x)) \to P_k x)$. Now, either $\varphi(x) \models Ab_{P_k P_l} c$, in which case trivially $\mathfrak{A} \models \forall x(P_k x \rightsquigarrow P_l x)$. Or else $\varphi(x) \models Ab_{P_k P_l} c$, in which case we find that $\Sigma^\uparrow \models \forall x((P_i x \wedge \varphi(x)) \to P_l x)$. So, by construction $\mathfrak{A} \models P_l x$, and this together with the facts that $\mathfrak{A} \models P_k c$ and $\mathfrak{A} \models Ab_{P_k P_l} c$ yields that $\mathfrak{A} \models \forall x(P_k x \rightsquigarrow P_l c)$.

Now suppose $\mathfrak{A}$ is not a model of $(\Sigma^\uparrow)^{wI}$. Then Γ contains a formula of the form $\forall x(P_j x \to \bigvee_{\delta\in\Delta'} Ab_\delta x)$ such that $\mathfrak{A} \models P_j c$ and $\Delta' \subseteq \Delta$. Since $\mathfrak{A} \models P_j c$, by construction $\Sigma^\uparrow \models \forall x((P_i x \wedge \varphi(x)) \to P_j x)$. Therefore by Lemma A.5 there is a positive path from v_i to v_j. But then by definition Γ should include the exemption clause $\forall x(P_i x \to \bigvee_{\delta\in\Delta'} Ab_\delta x)$. This contradicts our assumption of there being no such Δ', so $\mathfrak{A}$ must be a model of $(\Sigma^\uparrow)^{wI}$. □

Theorem A.7. *Let $\Sigma^\uparrow$ be the lift of a cycle-free inheritance network $\langle V, \Sigma\rangle$ with $V = \{v_1, \ldots, v_m\}$. If $\Sigma^\uparrow$ is coherent, $(\Sigma^\uparrow)^{wI} \models (\Sigma^\uparrow)^{\in}$.*

Proof. It suffices to show that $(\Sigma^\uparrow)^{wI}$ satisfies the exemption principle. So, let θ, θ' be any clauses of the form below:

$$\theta = \forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x),$$

$$\theta' = \forall x(P_i x \to \bigvee_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_i x}} Ab_\delta x).$$

We have to prove that whenever such a θ' is implied by $(\Sigma^\uparrow)^{wI}$, so is θ. Suppose $(\Sigma^\uparrow)^{wI} \models \theta'$. By Lemma A.6, $\theta' \in (\Sigma^\uparrow)^{wI}$.[26] There are two cases, depending on where θ' was added.

Case (i): $\theta' \in (\Sigma^\uparrow)^{w}$. In this case

$$\Sigma^\uparrow \models \forall x(P_i x \to \bigvee_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_i x}\cup(\Sigma^\uparrow)^{P_i x}} Ab_\delta x).$$

This means that $\Sigma^\uparrow \models \theta'$, and therefore, $\theta \in (\Sigma^\uparrow)^{w}$.

Case (ii): $\theta' \in (\Sigma^\uparrow)^{wI} - (\Sigma^\uparrow)^{w}$. In this case there is a positive path from v_i to some v_j such that

$$(\Sigma^\uparrow)^{w} \models \forall x(P_j x \to \bigvee_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_i x}} Ab_\delta x).$$

By Lemma A.6 we may assume that $(\Sigma^\uparrow)^{w}$ contains this exemption clause, and therefore

$$(\Sigma^\uparrow) \models \forall x(P_j x \to \bigvee_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_i x}\cup(\Sigma^\uparrow)^{P_j x}} Ab_\delta x).$$

Claim: the above can be strengthened to

$$\Sigma^\uparrow \models \forall x(P_j x \to \bigvee_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_j}} Ab_\delta x).$$

To prove the claim, let χ be the strengthened formula and χ' the 'un-strengthened' one. Assume $\Sigma^\uparrow \not\models \chi$.

Define $\mu(x)$ as follows:

$$\mu(x) = \left(\bigwedge_{\delta\in\Delta\cup(\Sigma^\uparrow)^{P_j}} \neg Ab_\delta x\right) \wedge \left(\bigvee_{\delta\in(\Sigma^\uparrow)^{P_i}} Ab_\delta x\right).$$

[26] Strictly speaking a stronger clause than θ' might be included instead, but we may assume without loss of generality that θ' is maximally strong to begin with.

$\Sigma^{\uparrow} \cup \{\exists x(P_j x \wedge \mu(x))\}$ is consistent. (Since $\Sigma^{\uparrow}$ is coherent, $\Sigma^{\uparrow} \cup \{\exists x P_j x\}$ is consistent. Therefore this follows directly from $\Sigma^{\uparrow} \models \chi'$, $\Sigma^{\uparrow} \not\models \chi$.) Since there is no path from v_j to v_i (otherwise there would be a cycle), contraposition of Lemma A.5 tells us it cannot be the case that $\Sigma^{\uparrow} \models \forall x((P_j x \wedge \mu(x)) \rightarrow P_i x)$. Therefore there is a model of $\Sigma^{\uparrow}$ with some element d satisfying $P_j x \wedge \neg P_i x \wedge \mu(x)$.

Adjust this model such that for no δ in $(\Sigma^{\uparrow})^{P_i}$ d satisfies $Ab_{\delta} x$. Since d does not satisfy $P_i x$, this adjusted model is still a model of $\Sigma^{\uparrow}$ (given that there is no path as above). However, this model does not make χ' true. Contradiction.

Given that $\Sigma^{\uparrow} \models \chi$, it follows that $\forall x(P_j x \rightarrow \bigvee_{\delta \in \Delta} Ab_{\delta} x) \in (\Sigma^{\uparrow})^{w}$. Since there is a path from v_i to v_j, this in turn leads to $\theta \in (\Sigma^{\uparrow})^{wI}$. □

A.1.1. Adding cycles

In order to properly deal with lifts of networks that include cycles, we first introduce a notion of equivalence that (for lifts) corresponds to being in the same cycle.

Definition A.8. Let Σ be a set of rules. The formulas φ and ψ are *equivalent in* Σ iff there are $\varphi_1, \ldots, \varphi_m$, $\psi_1, \ldots, \psi_n$ such that $\varphi_m = \psi = \psi_1$, $\psi_n = \varphi = \varphi_1$ and for all $1 \leq i < m$, $1 \leq j < n$

$\Sigma \models \forall x(\varphi_i(x) \rightsquigarrow \varphi_{i+1}(x))$, and $\Sigma \models \forall x(\psi_j(x) \rightsquigarrow \psi_{j+1}(x))$.

If φ and ψ are equivalent in Σ, we write $\varphi \approx_{\Sigma} \psi$, or simply $\varphi \approx \psi$ if no confusion is possible.

Instead of the weak exemption extension Σ^{w} we will utilize the *expanded* weak exemption extension Σ^{W}.

Definition A.9.

(i) The clause $\forall x(\varphi(x) \rightarrow \bigvee_{\delta \in \Delta} Ab_{\delta} x)$ is an expanded exemption clause for Σ iff there are $\psi_1 \approx \psi_2 \approx \ldots \approx \psi_n \approx \varphi$ such that

$$\Sigma \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)} \cup \Sigma^{\psi_1(x)} \cup \ldots \cup \Sigma^{\psi_n(x)}} Ab_{\delta} x).$$

(ii) The expanded weak exemption extension Σ^{W} of Σ is given by

$$\Sigma^{W} = \Sigma \cup \{\varphi \mid \varphi \text{ is an expanded exemption clause for } \Sigma\}.$$

Of course, showing that $\Sigma^{\in}$ and Σ^{WI} are equally strong, will require some elaborations. Note that Lemmas A.4 and A.5 are already stated and proven in terms of general networks which may include cycles.

Proposition A.10. $\Sigma^{\in} \models \Sigma^{WI}$.

Proof. We first prove that $\Sigma^{\in} \models \Sigma^{W}$. Let $\theta \in \Sigma^{W}$. We may assume that $\Sigma \not\models \theta$ (otherwise $\Sigma^{\in} \models \theta$ follows immediately). Therefore, θ is of the form

$$\theta = \forall x(\varphi(x) \rightarrow \bigvee_{\delta \in \Delta} Ab_{\delta} x)$$

with

$$\Sigma \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)} \cup \Sigma^{\psi_1(x)} \cup \ldots \cup \Sigma^{\psi_n(x)}} Ab_{\delta} x)$$

for some $\psi_1 \approx \psi_2 \approx \ldots \approx \psi_n \approx \varphi$.

Since $\psi_1 \approx \varphi$, (repeated) use of the inheritance principle lets us conclude

$$\Sigma^{\in} \models \forall x(\psi_1(x) \rightarrow \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)} \cup \Sigma^{\psi_1(x)} \cup \ldots \cup \Sigma^{\psi_n(x)}} Ab_{\delta} x).$$

By taking $\Delta' = \Delta \cup \Sigma^{\varphi(x)}$, we may use the exemption principle to conclude

$$\Sigma^{\in} \models \forall x(\psi_1(x) \rightarrow \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)} \cup \Sigma^{\psi_2(x)} \cup \ldots \cup \Sigma^{\psi_n(x)}} Ab_{\delta} x).$$

Now, by (repeatedly) using the inheritance principle again we arrive at

$$\Sigma^{\in} \models \forall x(\varphi(x) \rightarrow \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)} \cup \Sigma^{\psi_2(x)} \cup \ldots \cup \Sigma^{\psi_n(x)}} Ab_{\delta} x).$$

The same process can be repeated for all ψ_i, leaving us with

$$\Sigma^{\in} \models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta \cup \Sigma^{\varphi(x)}} Ab_\delta x),$$

from which it follows through the exemption principle that

$$\Sigma^{\in} \models \forall x(\varphi(x) \to \bigvee_{\delta \in \Delta} Ab_\delta x).$$

This proves that $\Sigma^{\in} \models \Sigma^W$. Therefore, $\Sigma^{\in I} \models \Sigma^{WI}$. Since the exemption principle implies the inheritance principle, $\Sigma^{\in} \models \Sigma^{\in I}$, which means that $\Sigma^{\in} \models \Sigma^{WI}$. □

Lemma A.11. *Let $\Sigma^\uparrow$ be the lift of an inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_m\}$. Let $\Gamma = \Sigma^{\uparrow WI} - \Sigma^\uparrow$, and assume $\Sigma^\uparrow \cup \Gamma \cup \{\exists x P_i x\}$ is consistent. If $\Sigma^{\uparrow WI} \models \forall x(P_i x \to \bigvee_{\delta\in\Delta} Ab_\delta x)$, then there is some $\Delta' \subseteq \Delta$ such that $\forall x(P_i x \to \bigvee_{\delta\in\Delta'} Ab_\delta x) \in \Gamma$.*

Proof. Analogous to the proof of Lemma A.6.

Theorem A.12. *If $\Sigma^\uparrow$ is the lift of an inheritance network $\langle V, \Sigma\rangle$ and is coherent, then $(\Sigma^\uparrow)^{WI} \models (\Sigma^\uparrow)^{WI\in}$.*

Proof. Mostly analogous to the proof of Theorem A.7, but we need to look more closely at Case (ii). Analogous to what we concluded in the simple case, we have

$$\Sigma^\uparrow \models \forall x(P_j x \to \bigvee_{\delta \in \Delta \cup (\Sigma^\uparrow)^{P_i x} \cup (\Sigma^\uparrow)^{Q_1 x} \cup \ldots \cup (\Sigma^\uparrow)^{Q_n x} \cup (\Sigma^\uparrow)^{P_j x}} Ab_\delta x)$$

with $P_i \approx Q_1 \approx \ldots \approx Q_n$. Now if there is a positive path from v_j to v_i then $P_j \approx P_i$. Then by construction $\forall x(P_j x \to \bigvee_{\delta\in\Delta} Ab_\delta x) \in (\Sigma^\uparrow)^W$. Since there is a path from v_i to v_j, this in turn leads to $\theta \in (\Sigma^\uparrow)^{WI}$.

(If there is no positive path from v_j to v_i then this part is also analogous to the proof of Theorem A.7.) □

A.1.2. $(\Sigma^\uparrow)^{WI}$ and the construction of the D function

The preceding subsection establishes that $\Sigma^\in$ and Σ^{WI} have the same models. What is easier to see – but still important to prove – is that the alternative constraints leading to Σ^{WI} correctly model what happens in constructing the D function. The following lemma and proposition cover this part of the completeness proof.

Lemma A.13. *Let $\Sigma^\uparrow$ be the lift of some network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_n\}$. Then $X \subseteq \Sigma$ is a conflicting set relative to v_i if and only if*[27]

$$\Sigma^\uparrow \models \forall x \left(P_i x \to \bigvee_{\alpha\in X} Ab_\alpha x \right).$$

Proof. Suppose $X \subseteq \Sigma$ is a conflicting set relative to v_i. Suppose towards contradiction that there is a model $\mathfrak{A}$ of $\Sigma^\uparrow$ such that

$$\mathfrak{A} \models \exists x \left(P_i x \wedge \bigwedge_{\alpha\in X} \neg Ab_\alpha x \right).$$

Since X is a conflicting set relative to v_i, there is some v_j such that X contains both a positive and a negative path to v_j. Therefore by repeated modus ponens (as well as modus tollens, possibly) it follows that both $P_j x$ and $\neg P_j x$. Contradiction.

For the other direction, suppose X is not a conflicting set relative to v_i. Let $\mathfrak{A}$ be a model in which $\forall x P_i x$ is true, and $\forall x \neg Ab_\alpha x$ is true for all $\alpha \in X$, while the extension of the remaining predicates is determined by applying the rules in $\Sigma^\uparrow$. Since there are no logical relations between the predicates other than those provided by $\Sigma^\uparrow$, this can be done. The resulting model $\mathfrak{A}$ is a model of $\Sigma^\uparrow$, but $\forall x \left(P_i x \to \bigvee_{\alpha\in X} Ab_\alpha x\right)$ is false on $\mathfrak{A}$. □

Proposition A.14. *Let $\Sigma^\uparrow$ be the lift of an inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_n\}$. Let*

$$\phi = \forall x \left(P_i x \to \bigvee_{\alpha\in X} Ab_\alpha x \right).$$

If $(\Sigma^\uparrow)^{WI} \models \phi$, then $Y \in D(v_i)$ for some $Y \subseteq X$. Conversely, if $X \in D(v_i)$ then $(\Sigma^\uparrow)^{WI} \models \phi$.

[27] Note that this proposition states that the formula is true on every model of $\Sigma^\uparrow$, even those which are not models of $(\Sigma^\uparrow)^{WI}$.

Proof. Suppose $(\Sigma^{\uparrow})^{WI} \models \phi$. By the construction of $(\Sigma^{\uparrow})^{WI}$, there must be some k such that there is a positive path from v_i to v_k and

$$(\Sigma^{\uparrow})^{W} \models \forall x \left(P_k x \rightarrow \bigvee_{\alpha \in X} Ab_\alpha x \right).$$

By the construction of $(\Sigma^{\uparrow})^{W}$, it follows that

$$\Sigma^{\uparrow} \models \forall x \left(P_k x \rightarrow \bigvee_{\alpha \in X \cup Ess_\Sigma(v_k)} Ab_\alpha x \right).$$

By Lemma A.13, this means that $X \cup Ess_\Sigma(v_k)$ is a conflicting set relative to v_k. Therefore, $Y \in d(v_k)$ and hence, $Y \in D(v_i)$, where $Y = X - Ess_\Sigma(v_k) \subseteq X$.

For the converse, suppose $X \in D(v_i)$. Then there is some v_j such that there is a positive path from v_i to v_j and $X \in d(v_j)$. Therefore, $X \cup Ess_\Sigma(v_j)$ is a conflicting set relative to v_j. By Lemma A.13,

$$\Sigma^{\uparrow} \models \forall x \left(P_j x \rightarrow \bigvee_{\alpha \in (X \cup Min_\Sigma(v_j))} Ab_\alpha x \right).$$

By construction of $(\Sigma^{\uparrow})^{W}$,

$$(\Sigma^{\uparrow})^{W} \models \forall x \left(P_j x \rightarrow \bigvee_{\alpha \in X} Ab_\alpha x \right),$$

and therefore $(\Sigma^{\uparrow})^{WI} \models \phi$. □

A.2. Completeness

Knowing (via Σ^{WI}) how the D function and $\Sigma^{\in}$ are related is an important step on our way to completeness, but we are far from done. One thing we do not yet know is what on the inheritance network side corresponds to the models in the sets $\mathcal{F}$ of the states $\langle \mathcal{U}, \mathcal{F} \rangle$. The bulk of the completeness proof lies in showing that these models correspond to acceptable exception sets, with optimal models corresponding to minimal exception sets.

Proposition A.15. *Let $\Sigma^{\uparrow}$ be the lift of the inheritance network $\langle V, \Sigma \rangle$, with $V = \{v_1, \ldots, v_n\}$. Let $I = \langle \Sigma^{\uparrow}, \{P_i c\} \rangle$. Let $\langle \mathcal{U}, \mathcal{F} \rangle$ be the information state generated by I. Then for all $\mathfrak{A} \in \mathcal{F}$, the set $X = \{\alpha \in \Sigma \mid \mathfrak{A} \models Ab_\alpha c\}$ is an acceptable exception set for v_i.*

Proof. Consider $Y \in D(v_i)$. We must show that there is some $\delta \in Y$ such that $\delta \in X$.

By Proposition A.14,

$$(\Sigma^{\uparrow})^{WI} \models \forall x \left(P_i x \rightarrow \bigvee_{\alpha \in Y} Ab_{\alpha^{\uparrow}} x \right).$$

Therefore, $\mathfrak{A} \models \bigvee_{\alpha \in Y} Ab_\alpha c$. Hence, there is some $\alpha \in Y$ such that $\mathfrak{A} \models Ab_\alpha c$, which means that $\alpha \in X$. □

Below, when $X = \{\alpha \in \Sigma \mid \mathfrak{A} \models Ab_\alpha c\}$, we will often say that X is the exception set represented by $\mathfrak{A}$.

Next proposition says that every minimal exception set is represented by at least one model in $\mathfrak{F}$.

Proposition A.16. *Let $\Sigma^{\uparrow}$ be the lift of the inheritance network $\langle V, \Sigma \rangle$, with $V = \{v_1, \ldots, v_n\}$. Let $I = \langle \Sigma^{\uparrow}, \{P_i c\} \rangle$. Let $\langle \mathcal{U}, \mathcal{F} \rangle$ be the information state generated by I.*

For every minimal exception set X for v_i there is a model $\mathfrak{A} \in \mathcal{F}$ such that $\mathfrak{A} \models Ab_\alpha c$ iff $\alpha \in X$.

Proof. Let X be a minimal exception set for v_i. Construct $\mathfrak{A}$ as follows:

- For the domain $\mathcal{A}$, take the same domain as that of some other model in $\mathcal{F}$, and choose $\mathcal{I}(c)$ arbitrarily.
- Stipulate that $\mathcal{I}(c) \in \mathcal{I}(P_i)$ and that $\mathcal{I}(c) \in \mathcal{I}(Ab_\alpha)$ iff $\alpha \in X$.
- For all P_j, stipulate that $\mathcal{I}(c) \in \mathcal{I}(P_j)$ if and only if there is a positive path from v_i to v_j that does not contain an element of X.
- For all $d \neq \mathcal{I}(c)$ and for all P_j, stipulate that $d \notin \mathcal{I}(P)$.

Clearly, $\mathfrak{A} \models Ab_\alpha c$ iff $\alpha \in X$. To show that $\mathfrak{A} \in \mathcal{F}$, it is sufficient to show that $\mathfrak{A} \in \mathcal{U}$. For this it suffices to show that $\mathfrak{A} \models (\Sigma^\uparrow)^{WI}$.

For elements other than c, the predicate assignments are trivially consistent with all rules and exemption clauses in $(\Sigma^\uparrow)^{WI}$. For c, we first look at the rules in $\Sigma^\uparrow$.

Rules in $\Sigma^\uparrow$: Consider $\phi \in \Sigma^\uparrow$, where

$$\phi = \forall x((P_j x \wedge \neg Ab_{P_j P_k} x) \rightarrow P_k x).$$

We may assume that $\mathfrak{A} \models P_j c \wedge \neg Ab_{P_j P_k} c$. (Otherwise, c is trivially consistent with the rule.) Thus, there is a positive path from v_i to v_j that does not contain an element of X, and the arc from v_j to v_k is not in X. Therefore, there is also such a path from v_i to v_k, and thus $P_k c$.

For negative rules, again take $\phi \in \Sigma^\uparrow$ but now with

$$\phi = \forall x((P_j x \wedge \neg Ab_{P_j \neg P_k} x) \rightarrow \neg P_k x).$$

Again we may assume that $\mathfrak{A} \models P_j c \wedge \neg Ab_{P_j \neg P_k} c$. Thus there is a negative path from v_i to v_k containing no element of X. To prove that $\mathfrak{A} \models \neg P_k c$ we have to show that there is no positive path from v_i to v_k. Suppose there is such a positive path, and let Y be the union of these two paths. Then Y is a conflicting set relative to v_i. Since X is a minimal exception set for v_i, some $\alpha \in Y$ must be in X. Since the negative path had no such overlap, this α must be part of the positive path.

As we've shown that every such positive path contains an element of X, it follows by construction that $\mathfrak{A} \models \neg P_k c$. Therefore the valuation for c is consistent with this rule.

Exemption clauses in $(\Sigma^\uparrow)^{WI}$: Suppose $\theta \in (\Sigma^\uparrow)^{WI}$, where

$$\theta = \forall x(P_j x \rightarrow \bigvee_{\alpha \in \Delta} Ab_{\alpha^\uparrow} x).$$

By Proposition A.14, $Y \in D(v_j)$ for some $Y \subseteq \Delta$. We may assume that $P_j c$. Therefore there is a positive path from v_i to v_j, and thus $Y \in D(v_i)$. Since X is a minimal exception set for v_i, it follows that there is some $\alpha' \in Y$ for which $\alpha' \in X$. By construction, $\mathfrak{A} \models Ab_{\alpha'} c$, and therefore, θ holds for c. □

Finally, we show that minimal exception sets correspond to optimal models.

Proposition A.17. *Let $\Sigma^\uparrow$ be the lift of the inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_n\}$. Let $I = \langle \Sigma^\uparrow, \{P_i c\}\rangle$. Let $\langle \mathcal{U}, \mathcal{F}\rangle$ be the information state generated by I. Then $\mathfrak{A}$ is optimal in $\mathcal{F}$ iff there is a minimal exception set X for v_i that is represented by $\mathfrak{A}$.*

Proof. For the proof from left to right let $\mathfrak{A}$ be optimal in $\mathcal{F}$. By Proposition A.15, the set $X = \{\alpha \in \Sigma \mid \mathfrak{A} \models Ab_\alpha c\}$ is an acceptable exception set for v_i. Assume towards contradiction that X is not a minimal exception set for v_i, and that $X' \subset X$ is. By Proposition A.16, there is a $\mathfrak{A}' \in \mathcal{F}$ which represents X'.

Now construct model $\mathfrak{A}''$ to be exactly like $\mathfrak{A}$ except that when evaluating predicates (including abnormality predicates) applied to c, it uses the same evaluation as $\mathfrak{A}'$. Showing that $\mathfrak{A}'' \in \mathcal{F}$ is fairly trivial and left to the reader. The abnormality predicates made true by $\mathfrak{A}''$ are a strict subset of those made true by $\mathfrak{A}$. So $\mathfrak{A}''$ is strictly more normal than $\mathfrak{A}$, which is therefore not optimal.

For the other direction, let X be a minimal exception set for v_i. By Proposition A.16, there are models $\mathfrak{A} \in \mathcal{F}$ such that $\mathfrak{A} \models Ab_\alpha c$ iff $\alpha \in X$.

Now, suppose $\mathfrak{B} \in \mathcal{F}$ is at least as normal as $\mathfrak{A}$. By Proposition A.15, $\mathfrak{B}$ models some acceptable exception set Y for v_i. Since $\mathfrak{B}$ is at least as normal as $\mathfrak{A}$, we have $Y \subseteq X$. Since X is minimal, this means $Y = X$. So, $\mathfrak{A}$ is at least as normal as $\mathfrak{B}$, which means that $\mathfrak{A}$ is an optimal model. □

Having established the correspondence between optimal models and minimal exception sets, the last step in the completeness proof is to go from these models to the allowable inferences as defined in Definition 4.8.

Theorem A.18. *Let $\Sigma^\uparrow$ be the lift of the inheritance network $\langle V, \Sigma\rangle$, with $V = \{v_1, \ldots, v_n\}$. Let $I = \langle \Sigma^\uparrow, \{P_i c\}\rangle$. Let $\langle \mathcal{U}, \mathcal{F}\rangle$ be the information state generated by I.*

(i) *If X is a minimal exception set for v_i, and there is a positive path from v_i to v_j that does not contain any element of X, then $\mathfrak{A} \models P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing X in c.*
(ii) *If X is a minimal exception set for v_i, and there is a negative path from v_i to v_j which does not contain any element of X, then $\mathfrak{A} \models \neg P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing X in c.*
(iii) *If X is not an acceptable exception set for v_j, then $\mathfrak{A} \models \neg P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing X in c.*
(iv) *If $\mathfrak{A} \models P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing X in c, then there is a positive path from v_i to v_j which does not contain any element of X.*
(v) *If $\mathfrak{A} \models \neg P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing X in c, then either there is a negative path from v_i to v_j which does not contain any element of X or X is not an acceptable exception set for v_j.*

Proof. The proofs of (i) and (ii) are left to the reader.

The proof of (iii) is also straightforward: If X is not an acceptable exception set for v_j, then there is some $Y \in D(v_j)$ such that $X \cap Y = \emptyset$. Since $Y \in D(v_j)$, $(\Sigma^{\uparrow})^{WI} \models \forall x(P_j x \rightarrow \bigvee_{\alpha \in Y} Ab_\alpha x)$ (Proposition A.14).

Suppose $\mathfrak{A} \in \mathfrak{F}$ represents X in c. Since $X \cap Y = \emptyset$, $\mathfrak{A}$ does not make $\bigvee_{\alpha \in Y} Ab_\alpha c$ true. Therefore, $\mathfrak{A} \models \neg P_j c$.

To prove (iv), suppose every $\mathfrak{A} \in \mathcal{F}$ representing X in c makes $P_j c$ true. Construct $\mathfrak{B} = \langle \mathcal{B}, \mathcal{I} \rangle$ as follows:

- For the domain $\mathcal{B}$, take the same domain as that of some other model in $\mathcal{F}$, and choose $\mathcal{I}(c)$ arbitrarily.
- Stipulate that $\mathcal{I}(c) \in \mathcal{I}(P_i)$ and that $\mathcal{I}(c) \in \mathcal{I}(Ab_\alpha)$ iff $\alpha \in X$.
- For all P_j, stipulate that $\mathcal{I}(c) \in \mathcal{I}(P_j)$ if and only if there is a positive path from v_i to v_j that does not contain an element of X.
- For all $d \neq \mathcal{I}(c)$ and for all P_j, stipulate that $d \notin \mathcal{I}(P_j)$.

We have shown in the proof of Proposition A.16 that $\mathfrak{B} \in \mathcal{F}$. Thus, by construction there is a positive path from v_i to v_j that does not contain an element of X.

To prove (v), suppose every $\mathfrak{A} \in \mathcal{F}$ representing X in c makes $\neg P_j c$ true. Construct $\mathfrak{B}'$ like $\mathfrak{B}$ in (iv), but stipulate that $\mathcal{I}(c) \in \mathcal{I}(P_j)$. Then $\mathfrak{B}' \notin \mathcal{F}$, and more specifically $\mathfrak{B}' \not\models (\Sigma^{\uparrow})^{WI}$. Pick $\phi \in (\Sigma^{\uparrow})^{WI}$ such that $\mathfrak{B}' \models \neg\phi$. A number of cases arise, depending on ϕ.

(a) $\phi = \forall x(P_k x \wedge \neg Ab_\phi x \rightarrow \neg P_j x)$ for some k, with $\mathfrak{B}' \models P_k c \wedge \neg Ab_\phi c$. In this case, there is a negative path from v_i to v_j (via v_k) that does not contain an element of X.

(b) $\phi = \forall x(P_j x \wedge \neg Ab_\phi x \rightarrow \neg P_k x)$ for some k, with $\mathfrak{B}' \models P_k c \wedge \neg Ab_\phi c$. In this case too, there is a negative path from v_i to v_j (via v_k using modus tollens at the end) that does not contain an element of X.

(c) $\phi = \forall x(P_j x \rightarrow \bigvee_{\delta \in \Delta} Ab_\delta x)$ for some Δ, with $\mathfrak{B}' \models \neg \bigvee_{\delta \in \Delta} Ab_\delta c$. Then it follows that $X \cap \Delta = \emptyset$. By Proposition A.14, $Y \in D(v_j)$ for some $Y \subseteq \Delta$. Since X contains no element of Δ, it contains no element of this Y. Therefore, X is not an acceptable exception set for v_j.

(d) $\phi = \forall x(P_j x \wedge \neg Ab_\phi \rightarrow P_k x)$ for some k, with $\mathfrak{B}' \models \neg P_k c \wedge \neg Ab_\phi c$. In this case, change the model one step further, making $P_k c$ true. Since the new model $\mathfrak{B}'' \notin \mathcal{F}$, find a new $\phi \in (\Sigma^{\uparrow})^{WI}$ such that $\mathfrak{B}'' \models \neg\phi$. If this ϕ' is like in case (a) or (b), there is still a negative path from v_i to v_j, which is just one arc longer than the path we found in case (a) and (b). (Recall that a negative path can go through any amount of positive arcs 'in the wrong direction' at the end.) If ϕ' is like case (c), then the Y which is found is also part of v_j. If ϕ' is itself like case (d), then we continue to proceed in the same way. Since no amount of making predicates true will make the model part of $\mathcal{F}$, going on long enough will lead to a ϕ' of one of the first three forms. The only potential complication in this induction is the possibility that we are led to a formula like type a or b where P_k is true merely because of a change we made to the model. In this case there is a negative path from v_j to itself of which no element is in X. Since this path is a contradicting set relative to v_j, it follows that X is not an acceptable exception set for v_j. □

Corollary A.19 *(Soundness–completeness). Let $\langle V, \Sigma \rangle$ be an inheritance network, and suppose $\Sigma^{\uparrow}$ is coherent. Then $v_i \vdash_\Sigma v_j$ if and only if $\Sigma^{\uparrow}, \{P_i c\} \models_d P_j c$, and $v_i \vdash_\Sigma \neg v_j$ if and only if $\Sigma^{\uparrow}, \{P_i c\} \models_d \neg P_j c$.*

Proof. The proof of the second equivalence is left to the reader. As for the first: let $\langle \mathcal{U}, \mathcal{F} \rangle$ correspond to $\langle \Sigma^{\uparrow}, \{P_i c\} \rangle$.

By definition, $v_i \vdash_\Sigma v_j$ if and only if (a) for every minimal exception set X for v_i, there is a positive path Y from v_i to v_j with $X \cap Y = \emptyset$.

By Theorem A.18, (a) holds iff (b) $\mathfrak{A} \models P_j c$ for every $\mathfrak{A} \in \mathcal{F}$ representing some minimal exception set for v_i in c.

By Proposition A.17, (b) holds iff (c) $\mathfrak{A} \models P_j c$ for every optimal model $\mathfrak{A}$ in $\mathcal{F}$.

By definition, (c) is true iff $\Sigma^{\uparrow}, \{P_i c\} \models_d P_j c$ $(\neg P_j c)$. □

A.3. Coherence

Theorem A.20. *Let $\langle V, \Sigma \rangle$ be an inheritance network with $V = \{v_1, \ldots, v_n\}$. Then $\Sigma^{\uparrow}$ is incoherent if and only if $\emptyset \in d(v_i)$ for some i.*

Proof. $\Sigma^{\uparrow}$ is incoherent if and only if there is some P_i such that $\Sigma^{\uparrow WI} \cup \{\exists x P_j x\}$ is inconsistent. This is so if and only if $(\Sigma^{\uparrow})^{WI} \models \forall x \neg P_j x$ for some P_j. By the convention on empty disjunctions, $\forall x \neg P_j x$ is equivalent to $\forall x(P_j x \rightarrow \bigvee_{\alpha \in \emptyset} Ab_\alpha x)$. By Proposition A.14 this means that $\emptyset \in D(v_j)$. The latter implies that $\emptyset \in d(v_i)$ for some i. □

Proposition A.21. *Let $\langle V, \Sigma \rangle$ be an inheritance network without strict arcs. If $\emptyset \in d(x)$, then there are some z and some $y \rightleftharpoons x$, $y' \rightleftharpoons x$ such that Σ contains a positive arc from y to z and a negative arc from y' to z.*

Proof. Suppose $\emptyset \in d(x)$. Then there is some minimal conflicting set $X \subseteq Ess_\Sigma(x)$. We may assume without loss of generality that X is the union of a positive path $\{xy_1, y_1 y_2, \ldots, y_m z\}$ and a negative path $\{xy'_1, y'_1 y'_2, \ldots, y'_n z^-\}$.

Since $y_m z \in X$, it follows that $y_m z \in Ess_\Sigma(x)$. Therefore, $x \rightleftharpoons y_m$. Analogously, $x \rightleftharpoons y'_n$. □

Proposition A.22. *Let $\langle V, \Sigma\rangle$ be an inheritance network. If $\emptyset \in d(x)$, then there are some z and some $y \rightleftharpoons x$, $y' \rightleftharpoons x$ such that y semi-strictly implies z and y' semi-strictly refutes z.*

Proof. Suppose $\emptyset \in d(x)$. Then there is some minimal conflicting set $X \subseteq Ess_\Sigma(x)$. We may assume without loss of generality that X is the union of a positive path $\{xy_1, y_1y_2, \ldots, y_nz\}$ and a negative path $\{xy'_1, y'_1y'_2, \ldots, y'_nz^-\}$ (where some of these may actually be strict).

Pick the smallest i for which y_i strictly implies z.[28] Since $y_{i-1}y_i \in X$, it follows that $y_{i-1}y_i \in Ess_\Sigma(x)$. But by construction $y_{i-1}y_i$ is not strict. Therefore, $y_{i-1} \rightleftharpoons x$.

Analogously, $y'_{j-1} \rightleftharpoons x$ when we pick the smallest j for which y'_j strictly refutes z. (If no y'_j does so, pick $j = n+1$ instead.) Now let $y = y_{i-1}$, $y' = y'_{j-1}$. By construction, y semi-strictly implies z and y' semi-strictly refutes it. □

Corollary A.23. *Let $\langle V, \Sigma\rangle$ be an inheritance network. $\Sigma^\uparrow$ is incoherent if and only if there are $x, y, z \in V$ such that $x \rightleftharpoons y$, x semi-strictly implies z, and y semi-strictly refutes z.*

References

[1] J. McCarthy, Circumscription: a form of non-monotonic reasoning, in: M.L. Ginsberg (Ed.), Readings in Nonmonotonic Reasoning, Kaufmann, Los Altos, CA, 1987, pp. 145–151.
[2] J. McCarthy, Applications of circumscription to formalizing common sense knowledge, in: V. Lifschitz (Ed.), Formalizing Common Sense: Papers by John McCarthy, Ablex Publishing Corporation, Norwood, New Jersey, 1990, pp. 198–225.
[3] J.F. Horty, R.M. Thomason, D.S. Touretzky, A skeptical theory of inheritance in nonmonotonic semantic networks, Artif. Intell. 42 (1990) 311–348.
[4] P.A. Bonatti, C. Lutz, F. Wolter, Description logics with circumscription, in: P. Doherty, J. Mylopoulos, C. Welty (Eds.), Proceedings of the 10th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2006, AAAI Press, 2006, pp. 400–410.
[5] P.A. Bonatti, C. Lutz, F. Wolter, The complexity of circumscription in DLs, J. Artif. Intell. Res. (2009) 717–773.
[6] J. Delgrande, An approach to default reasoning based on a first-order conditional logic: revised report, Artif. Intell. 36 (1988) 63–90.
[7] N. Asher, M. Morreau, Commonsense entailment: a modal theory of nonmonotonic reasoning, in: J. van Eijck (Ed.), Logics in AI: Proc. of the European Workshop JELIA'90, Springer, Berlin, Heidelberg, 1991, pp. 1–30.
[8] C. Boutilier, The complexity of circumscription in DLs, Artif. Intell. 68 (1994) 87–154.
[9] F. Veltman, Defaults in update semantics, J. Philos. Log. (1996) 221–261.
[10] H. Geffner, Default Reasoning: Causal and Conditional Theories, vol. 4, MIT Press, Cambridge, MA, 1992.
[11] D. Makinson, K. Schlechta, Floating conclusions and zombie paths: two deep difficulties in the "directly skeptical" approach to defeasible inheritance nets, Artif. Intell. 48 (1991) 199–209.
[12] M. Ginsberg, Essentials of Artificial Intelligence, Morgan Kaufmann, 1993.
[13] J.F. Horty, Skepticism and floating conclusions, Artif. Intell. 135 (2002) 55–72.
[14] H. Prakken, Intuitions and the modelling of defeasible reasoning: some case studies, arXiv preprint, cs/0207031.
[15] R.H. Thomason, J.F. Horty, D.S. Touretzky, A calculus for inheritance in monotonic semantic nets, in: Z. Ras, M. Zemankova (Eds.), Proceedings of the Second International Symposium on Methodologies for Intelligent Systems, North Holland, 1987, pp. 280–287.
[16] A. Bochman, Default theory of defeasible entailment, in: Proceedings of the 11th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2008, AAAI Press, 2008, pp. 466–475.
[17] R. Reiter, A logic for default reasoning, Artif. Intell. 13 (1) (1980) 81–132.
[18] J.F. Horty, Some direct theories of nonmonotonic inheritance, in: D. Gabbay, C. Hogger, J. Robinson (Eds.), Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 3, Oxford University Press, 1994, pp. 111–187.
[19] Z. Qian, K.B. Irani, Circumscribing defaults, in: Proceedings of the 12th International Joint Conference on Artificial Intelligence, IJCAI, Morgan Kaufmann, 1991, pp. 438–445.

[28] For y_{i-1} to exist we must assume x does not semi-strictly imply z, but this is safe because if it does then we can pick $y = x$ and skip the next couple of steps in the proof.