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.