Logical Foundations of Well-Founded Semantics*

Pedro Cabalar
Department of Computer Science,
Corunna University, Spain.
cabalar@dc.fi.udc.es

Sergei Odintsov
Sobolev Institute of Mathematics,
Novosibirsk, Russia.
odintsov@math.nsc.ru

David Pearce
Universidad Rey Juan Carlos,
Madrid, Spain.
davidandrew.pearce@urjc.es

Abstract

We propose a solution to a long-standing problem in the foundations of well-founded semantics (WFS) for logic programs. The problem addressed is this: which (non-modal) logic can be considered adequate for well-founded semantics in the sense that its minimal models (appropriately defined) coincide with the partial stable models of a logic program? We approach this problem by identifying the HT^2 frames previously proposed by Cabalar to capture WFS as structures of a kind used by Došen to characterise a family of logics weaker than intuitionistic and minimal logic. We define a notion of minimal, total HT^2 model which we call *partial equilibrium model*. Since for normal logic programs these models coincide with partial stable models, we propose the resulting partial equilibrium logic as a logical foundation for well-founded semantics. In addition we axiomatise the logic of HT^2-models and prove that it captures the strong equivalence of theories in partial equilibrium logic.

Keywords: well-founded semantics, partial stable models, equilibrium logic

Introduction

Of the various proposals for dealing with default negation in logic programming that go beyond the methods of ordinary Prolog, the *well-founded semantics* (WFS) of Van Gelder, Ross and Schlipf (van Gelder, Ross, & Schlipf 1991) has proved to be one of the most attractive and resilient. Particularly its favourable computational properties have made it popular among system developers and the well-known implementation XSB-Prolog[1] is now extensively used in AI problem solving and applications in knowledge representation and reasoning. The present paper studies the logical foundations of WFS and proposes a solution to a long-standing open problem: how to characterise partial stable models as minimal models in a suitable nonclassical, nonmodal logic. We thereby obtain a deductive base logic (in the sense of (Dietrich 1994)) for well-founded inference as well as a means to extend WFS to disjunctive programs and arbitrary propositional theories.[2] A major challenge of the paper is to axiomatise the base logic.

Well-founded semantics defines a nonmonotonic inference relation whose properties have been keenly studied. But compared with other logical formalisms popular in knowledge representation and reasoning, the logical *foundations* of WFS remain largely uncharted territory. One cause seems to be that an adequate underlying logic for WFS has simply not been identified and studied. While well-founded models are easy to describe using a three-valued semantics, it is much harder to say under which logic or logics is well-founded inference closed. Also WFS is still closely tied to a very restricted syntax, that of normal logic programs. Various attempts have been made to extend this syntax, mainly in two directions, by adding a second, explicit negation operator and, secondly, by permitting disjunctive rules, ie rules whose heads may comprise disjunctions of atoms. While the first avenue has seen some success in terms of implemented systems and practical applications (Pereira & Alferes 1992), progress on the second path has been limited. The main problem is that there have been several different proposals for how to extend WFS with disjunction and virtually no agreement on which is more adequate or even on the general criteria by which adequacy can be assessed.

A natural way to identify a logical foundation for a logic programming semantics or inference relation is to represent the intended models of the semantics as minimal models in a suitable logic. In the case of stable models and answer sets a solution of this sort was found several years ago. The logic of here-and-there, HT, (also known as Gödel's 3-valued logic) was used in (Pearce 1997) to represent stable models as minimal models and was identified as a maximal logic with the property that equivalent theories have the same (stable model) semantics. The weakest extension of the logic containing strong negation and its axioms plays the same role for answer set programming with two negations. Later, in (Lifschitz, Pearce, & Valverde 2001), it was shown that HT characterises the strong equivalence of programs (see below) under stable semantics, and subsequently (de Jongh & Hendriks 2002) identified the family of all su-

*Partially supported by CICyT project TIC-2003-9001-C02 and WASP (IST-2001-37004)

[1]See http://www.cs.sunysb.edu/~sbprolog/xsb-page.html

[2]Without going into technicalities, very roughly if $\mid\!\sim$ is some nonmonotonic inference relation then a monotonic logic $\mathcal{L}$ forms a deductive base for $\mid\!\sim$ if $\mid\!\sim$ extends $\mathcal{L}$-inference and $\mathcal{L}$-equivalent theories are also equivalent under $\mid\!\sim$.

perintuitionistic logics with this property. This work has led to a flourishing research programme in the area of logic programming and nonmonotonic reasoning, influencing research topics such as how to: (i) extend the syntax of answer set programs eg. to allow rules with boolean formulas in heads and bodies (Lifschitz, Tang, & Turner 1999), or even to arbitrary propositional formulas (Pearce 1997; Ferraris 2005); (ii) provide a semantics for answer set programs with additional constructs such as cardinality constraints, weight constraints (Ferraris & Lifschitz 2005) or aggregates (Ferraris 2005); (iii) characterise strong equivalence for other nonmonotonic inference relations and LP semantics (Turner 2001; 2004; Odintsov & Pearce 2005); (iv) define variants of strong equivalence such as partial equivalence (Woltran 2004), equivalence wrt particular classes of formulas (Eiter & Fink 2003; Pearce & Valverde 2004b) or equivalence across different vocabularies (Pearce & Valverde 2004a); (v) use monotonic base logics to verify valid program transformations (Osorio, Navarro, & Arrazola 2001; Pearce 2004).

The approach and main results of the paper

Our aim here is to initiate a similar foundational study for WFS following the same approach adopted for stable models and answer sets: find a minimal model characterisation of the intended logic programming structures and axiomatise the resulting base logic. In the case of stable models the corresponding minimal models in the logic HT are called *equilibrium* models. For reasons that will become clear, in the case of partial stable semantics and WFS it is natural to call the resulting minimal models *partial equilibrium* models.

There are several possible approaches to understanding and extending WFS. For example the technique of (Brass & Dix 1994) to capture WFS via a set of program transformations has been much discussed in the literature. Our approach by contrast proceeds via partial stable (p-stable) models. There are several reasons for this. First, p-stable models, though defined via program reducts, are not too far removed conceptually from ordinary model theoretic semantics. Thus we can hope to analyse them via logical and model-theoretic methods. Second, p-stable models are a natural generalisation of ordinary (2-valued) stable models to a multi-valued setting; and as we have seen stable models do admit a natural logical foundation. Third, the well-founded model of a normal logic program coincides with the unique minimal p-stable model. So if we can capture p-stable semantics for normal programs in terms of minimal models for some logic, then a further minimisation process will yield the well-founded model.

Before proceeding to the technical details of the paper let us see briefly why, unlike in the case of stable model reasoning, we cannot use superintuitionistic logics as a basis for well-founded semantics. When dealing with deductive bases for logic programming we are assuming the direct or 'trivial' translation of program rules into logical formulas. In this paper we use the negation symbol '$\neg$' corresponding to the operator not of logic programs, and ordinary implication'$\rightarrow$' replaces the inverted arrow '$\leftarrow$'. So a program rule of the form $q \leftarrow p, not\ r$ is treated simply as a logical formula

$$p \wedge \neg r \rightarrow q.$$

Now consider the simple rule $p \leftarrow not\ p$, or $\neg p \rightarrow p$ written as a logical formula, and let us suppose that the non-logical constants are p and q. The well-founded model of this program makes the atom p undecided and q false. Now suppose we enlarge our program by adding a new rule $q \leftarrow not\ p$. Viewed as a formula $\neg p \rightarrow q$, from the standpoint of intuitionistic logic we have added no new 'content', since intuitionistically $\neg p \rightarrow q$ follows logically from $\neg p \rightarrow p$, so the second program is equivalent to the first. But consider the well-founded model of the enlarged program: now the atom q has changed truth value and become undecided; the two programs are not equivalent under WFS. So intuitionistically equivalent formulas need not have the same well-founded semantics.

The paper is organised as follows. We first review Došen semantics (Došen 1986) for a logic called N, a general framework for dealing with weak negation, and then proceed to study a particular case N^* that results from combining this approach with Routley semantics (Routley & Routley 1972) (also used in (Odintsov & Pearce 2005) for paraconsistent answer sets). We then consider the semantics of HT^2 frames introduced by (Cabalar 2001) to model partial stable models and we show that they are a particular case of N^* frames. We also present in this section a 6-valued characterisation of HT^2 that sheds some new light on the comparison with Przymusinski's 3-valued definition of partial stable models. Next we define a concept of minimal and total HT^2 model called *partial equilibrium* model and we show that on logic programs these coincide with partial stable models, so that partial equilibrium logic provides a foundation for and extension of well-founded inference. Two of the main contributions of the paper follow. First we axiomatise the HT^2 logic and prove a completeness theorem, then we establish in the following section a strong equivalence theorem to show how the logic HT^2 captures strongly equivalent theories in partial equilibrium logic. The paper concludes with a brief discussion of related work and with some topics for future investigation.

Our study is still at a preliminary stage and many issues are left open for future work. For instance, it is evident that partial equilibrium logic provides a means to extend well-founded and p-stable semantics beyond the syntax of normal programs, even to arbitrary propositional theories. However, a more detailed analysis of the behaviour of this extension and a comparison with other extensions of WFS in the literature are topics we are currently exploring and hope to present in future work.

Došen and Routley semantics

The logic we are going to investigate is an extension of a logic introduced by Došen in (Došen 1986) (see also (Došen 1999)) which he denotes by N. Došen's aim was to study logics weaker than Johansson's minimal logic. We recall here the main definitions and facts regarding N. Formulas of N are built-up in the usual way using atoms from a given propositional signature At and the standard logical constants: $\wedge$, $\vee$, $\rightarrow$, $\neg$, respectively standing for conjunction, disjunction, implication and negation. We write For to stand for the set of all well-formed formulae of this language. The rules of inference for N are *modus ponens* and the contraposition rule

$$\frac{\alpha \rightarrow \beta}{\neg\beta \rightarrow \neg\alpha} \tag{RC}$$

The set of axioms contains the axiom schemata of positive logic plus:

$$\neg\alpha \wedge \neg\beta \rightarrow \neg(\alpha \vee \beta) \tag{1}$$

Definition 1 (N model) *A model for N is a quadruple $\mathcal{M} = \langle W, \leq, R, V\rangle$ such that: (i) $\langle W, \leq\rangle$ is a partial ordering (of worlds), (ii) $R \subseteq W^2$ is an accessibility relation among worlds verifying $(\leq R) \subseteq R$, (iii) and finally, V is a valuation function from $At \times W \longrightarrow \{0, 1\}$ satisfying:*

$$V(p, u) = 1 \;\;\&\;\; u \leq w \;\;\Rightarrow\;\; V(p, w) = 1 \tag{2}$$

V is extended to a valuation on all formulas via the following conditions.

- $V(\varphi \wedge \psi, w) = 1$ iff $V(\varphi, w) = V(\psi, w) = 1$
- $V(\varphi \vee \psi, w) = 1$ iff $V(\varphi, w) = 1$ or $V(\psi, w) = 1$
- $V(\varphi \rightarrow \psi, w) = 1$ iff for every w' such that $w \leq w'$, $V(\varphi, w') = 1 \Rightarrow V(\psi, w') = 1$
- $V(\neg\varphi, w) = 1$ iff for every w' such that wRw' $V(\varphi, w') = 0$

As the reader may have already observed, the main difference with respect to intuitionistic frames is the presence of a new accessibility relation R used for interpreting negation, while $\leq$ remains for implication. Extending valuation V to all formulas, we use for positive connectives the same conditions as in the case of intuitionistic logic, but for negation we use instead the condition involving the relation R.

A proposition φ is said to be *true* in an N model $\mathcal{M} = \langle W, \leq, R, V\rangle$, if $V(\varphi, v) = 1$, for all $v \in W$. A formula φ is *valid*, in symbols $\models \varphi$, if it is true in every N model. It is easy to prove by induction that condition (2) above holds for any formula φ, ie

$$V(\varphi, u) = 1 \;\;\&\;\; u \leq w \;\;\Rightarrow\;\; V(\varphi, w) = 1 \tag{3}$$

Moreover, N is complete, that is, a formula is valid iff it is a theorem of N.

Let us consider now the logic N^* obtained by adding to N the following axioms

$$\neg(\alpha \rightarrow \alpha) \;\;\rightarrow\;\; \beta \tag{4}$$

$$\neg(\alpha \wedge \beta) \;\;\rightarrow\;\; \neg\alpha \vee \neg\beta \tag{5}$$

Thus, both De Morgan laws are provable in N^*

$$N^* \vdash \neg(\alpha \wedge \beta) \leftrightarrow \neg\alpha \vee \neg\beta,\; \neg(\alpha \vee \beta) \leftrightarrow \neg\alpha \wedge \neg\beta$$

since axioms (1) and (5) explicitly provide one direction of each law, whereas the opposite implications can be inferred via the rule (RC).

An N model $\mathcal{M} = \langle W, \leq, R, V\rangle$ is called an N^* *model* if it satisfies

$$\forall x \exists x^* (xRx^* \wedge \forall y(xRy \Rightarrow y \leq x^*)) \tag{6}$$

ie, for any world x of $\mathcal{M}$ there exists a $\leq$-greatest world x^* among those accessible from x by R. It can be easily checked that the new axioms of N^* are valid in all N^* models. Moreover, N^* is complete wrt to that class of models.

The persistence of validity of formulas (3) plus the existence of a greatest R-accessible world (6) implies that the validity of negated formulas at x is equivalent to

$$V(\neg\varphi, x) = 1 \;\Leftrightarrow\; V(\varphi, x^*) = 1.$$

This observation allows us to define a Routley style semantics (Routley & Routley 1972) for extensions of N^*.

Definition 2 *A Routley frame is a triple $\langle W, \leq, *\rangle$, where W is a set, $\leq$ a partial order on W and $* : W \rightarrow W$ is such that $x \leq y$ iff $y^* \leq x^*$. A Routley model is a Routley frame together with a valuation $V : At \times W \longrightarrow \{0, 1\}$ as for N models.*

As usual, a formula φ is valid in a Routley model if it is valid at every world of this model.

Completeness proofs for N and N^* can be obtained via the method of canonical models. We now sketch this for N^*. Let S be some logic extending N^*, ie some set of formulas containing N^* and closed under substitution, (RC), and *modus ponens*. First we say that a set of formulas Γ is a *theory* wrt S (*S theory*) if it contains S and is closed under *modus ponens* and a *prime S theory* if it additionally satisfies the disjunction property:

$$\alpha \vee \beta \in \Gamma \Rightarrow \alpha \in \Gamma \;\text{ or }\; \beta \in \Gamma$$

Next we note a standard extension lemma. Let Σ and Δ be sets of formulas. A relation $\Sigma \vdash_S \Delta$ means that for some $\varphi_0, \ldots, \varphi_n \in \Delta$ the disjunction $\varphi_0 \vee \ldots \vee \varphi_n$ can be obtained from elements of S and Σ using the rule of *modus ponens*.

Lemma 1 (Extension lemma) *For any extension S of N^*, any sets of formulas Σ and Δ, if $\Sigma \not\vdash_S \Delta$, then there is a prime S theory $\Gamma \supseteq \Sigma$ such that $\Gamma \not\vdash_S \Delta$.*

On this basis one defines canonical models as follows.

Definition 3 (Canonical model) *Let S be any extension of N^*. The* canonical S frame *is the triple $\langle W^c, \leq^c, *^c\rangle$ where (i) W^c is the set of prime theories wrt S, (ii) $\Gamma \leq^c \Delta := \Gamma \subseteq \Delta$, (iii) $\Gamma^{*^c} := \{\alpha | \neg\alpha \notin \Gamma\}$. The* canonical S model *is the canonical S frame together with the valuation function V^c such that $V^c(p, \Gamma) = 1$ iff $p \in \Gamma$.*

It is not hard to check that the canonical model S is indeed a Routley model. The only non-trivial item is to prove that the $*$-function is well defined, ie that Γ^{*^c} is a prime theory.

Lemma 2 *For any prime S theory Γ, the set Γ^{*^c} is also a prime S theory.*

Proof. For $\varphi \in S$, we have $\neg\varphi \notin \Gamma$, otherwise Γ is trivial by axiom (4). Thus, $S \subseteq \Gamma^*$. Let φ and $\varphi \rightarrow \psi$ be in Γ^*, ie $\neg\varphi, \neg(\varphi \rightarrow \psi) \notin \Gamma$. By the disjunction property of Γ

and De Morgan laws $\neg(\varphi \wedge (\varphi \rightarrow \psi)) \notin \Gamma$, equivalently, $\neg(\varphi \wedge \psi) \notin \Gamma$. Since $\neg\psi \rightarrow \neg(\varphi \wedge \psi) \in N^*$, $\neg\psi \notin \Gamma$, i.e. $\psi \in \Gamma^*$. We have proved that Γ^* is closed under *modus ponens*. The disjunction property of Γ^* follows by De Morgan laws. □

Now the completeness for N^* follows from the lemma for the canonical model.

Lemma 3 *In the canonical S model, for every $\Gamma \in W_c$ and every φ,*

$$V^c(\varphi, \Gamma) = 1 \Leftrightarrow \varphi \in \Gamma.$$

Proof. By induction on the complexity of φ. □

The completeness property follows by noting that if $\not\vdash_{N^*} \varphi$ then by the extension lemma there is a prime N^* theory Γ such that $\varphi \notin \Gamma$. It follows from Lemma 3 that φ does not hold in the canonical N^* model and therefore is not N^*-valid.

Theorem 1 *For any formula φ, $\vdash_{N^*} \varphi$ iff φ is valid in every Routley model.*

Finally, we note that an intuitionistic negation can be defined in N^*. Fix some propositional variable p_0 and put

$$\bot := \neg(p_0 \rightarrow p_0) \text{ and } -\alpha := \alpha \rightarrow \bot.$$

From axiom (4) it follows that for any Routley model $\mathcal{M} = \langle W, \leq, *, V\rangle$, the constant $\bot$ is not satisfied at any world $w \in W$. Therefore, the satisfaction of the derived expression $-\alpha$ coincides exactly with the interpretation of negation in intuitionistic logic:

$$V(-\alpha, w) = 1 \text{ iff } \forall w' \text{ such that } w \leq w', V(\alpha, w') = 0.$$

Since satisfaction of positive connectives is also defined as in intuitionistic logic, we actually have

Proposition 1 *The $\langle\vee, \wedge, \rightarrow, -\rangle$-fragment of N^* coincides with intuitionistic logic.*

HT^2-models

As mentioned in the introduction, it was shown in (Pearce 1997) that the so-called logic of *here-and-there*, HT, can be used as a foundation for the stable model semantics for logic programs. In the semantics for intermediate or super-intuitionistic logics, HT can be captured by rooted Kripke frames with two elements, commonly denoted by h and t and called 'here' and 'there', with $h \leq t$. In (Cabalar 2001) a notion of HT^2 model was introduced and studied in order to capture partial stable models for logic programs. The motivation for the notation is that HT^2 models are based on frames that include for each world w in an HT-model an additional world w' accessible from w via the $\leq$ relation. In addition, just as we have $h \leq t$ in an HT-model, we have also $h' \leq t'$ in an HT^2-model. More precisely we define HT^2 in terms of N models as follows.

Definition 4 (HT^2 model) *An HT^2* model *is an N model $\mathcal{M} = \langle W, \leq, R, V\rangle$ such that (i) W comprises 4 worlds denoted by h, h', t, t', (ii) $\leq$ is a partial ordering on W satisfying $h \leq t$, $h \leq h'$, $h' \leq t'$ and $t \leq t'$, (iii) $R \subseteq W^2$ is given by hRh', $h'Rh$, tRt', $t'Rt$, hRt', $h'Rt$. (iv) V is an N-valuation.*

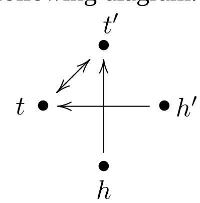
An interesting observation is that when we force $h = h'$ and $t = t'$ we actually obtain that $\leq$ and R collapse into the same relation and, in fact, the whole structure becomes an HT frame. Thus, it is easy to see that:

Proposition 2 *Any valid formula in HT^2 is also a valid formula in HT.*

Proposition 3 *In HT^2 models the following formulas are valid: $\alpha \rightarrow \neg\neg\alpha$, $\neg\alpha \leftrightarrow \neg\neg\neg\alpha$, $\neg(\alpha \rightarrow \alpha) \rightarrow \beta$, $\neg(\alpha \wedge \beta) \rightarrow \neg\alpha \vee \neg\beta$, $\alpha \wedge \neg\alpha \rightarrow \neg\beta \vee \neg\neg\beta$.*

Proof. Since HT^2 frames are finite, the validity of formulas can be verified directly. □

According to Proposition 3, the HT^2 frame defines an extension of N^* and we can replace the above defined models by models based on the Routley frame $\mathcal{W}^{HT^2} = \langle W^{HT^2}, \leq, *\rangle$, where $W^{HT^2} = \{h, h', t, t'\}$ and the ordering $\leq$ and the action of $*$ are represented in the following diagram.



That is, $h^* = t^* = t'$, $(h')^* = (t')^* = t$ and $u < v$ iff v is strictly higher than u in the diagram.

Now, fix some HT^2 model $\mathcal{M} = \langle \mathcal{W}^{HT^2}, V\rangle$. For $w \in W^{HT^2}$ let us set $\Delta_w^{\mathcal{M}} := \{\varphi : V(\varphi, w) = 1\}$.

Lemma 4 *For an arbitrary HT^2 model $\mathcal{M} = \langle \mathcal{W}^{HT^2}, V\rangle$ the following hold. (i) Δ_w is a prime HT^2 theory for any $w \in W^{HT^2}$. (ii) $\Delta_u \subseteq \Delta_v$ iff $u \leq v$. (iii) $\Delta_{t'} = \Delta_h^*$ and $\Delta_t = \Delta_{h'}^*$. (iv) $\varphi \rightarrow \psi \in \Delta_w$ iff for all $v \geq w$ either $\varphi \notin \Delta_v$ or $\psi \in \Delta_v$.*

Proof. All these properties follow straightforwardly from the definition of validity of formulas in Routley models and the structure of an HT^2 frame. □

Lemma 5 *Let $\mathsf{Q} = \langle \Delta_h, \Delta_{h'}, \Delta_t, \Delta_{t'}\rangle$ be a quadruple of prime HT^2 theories satisfying all conditions of Lemma 4. Define an HT^2 model $\mathcal{M}_\mathsf{Q} = \langle \mathcal{W}^{HT^2}, V_\mathsf{Q}\rangle$ as follows:*

$$V_\mathsf{Q}(p, w) = 1 \textit{ iff } p \in \Delta_w, \ w \in W^{HT^2}.$$

Then for all $w \in W^{HT^2}$ we have

$$\Delta_w^{\mathcal{M}_\mathsf{Q}} = \Delta_w.$$

Proof. By induction on the structure of formulas. □

Due to the last two lemmas, the Routley HT^2 model $\mathcal{M}$ can be identified with the quadruple of prime HT^2 theories $\langle \Delta_h^{\mathcal{M}}, \Delta_{h'}^{\mathcal{M}}, \Delta_t^{\mathcal{M}}, \Delta_{t'}^{\mathcal{M}}\rangle$. Now, let us say that a prime theory Δ is: *consistent* if $\varphi \wedge \neg\varphi \notin \Delta$ for any φ; *inconsistent* if it is not consistent; *complete* if $\varphi \notin \Delta$ implies $\neg\varphi \in \Delta$; and *weakly complete* if $\neg\varphi \notin \Delta$ implies $\neg\neg\varphi \in \Delta$.

Lemma 6 *Let Δ be a prime HT^2 theory. (i) $\Delta^{**} = \{\varphi : \neg\neg\varphi \in \Delta\}$; (ii) $\Delta \subseteq \Delta^{**}$ and $\Delta^* = \Delta^{***}$; (iii) $\Delta = \Gamma^*$ iff Δ is closed under the rule $\neg\neg\varphi/\varphi$; (iv) if Δ is inconsistent, then Δ is weakly complete; (v) Δ^* is consistent iff Δ is weakly complete; (vi) $\Delta \subseteq \Delta^*$ iff Δ is consistent; (vii) $\Delta^* \subseteq \Delta$ iff Δ is complete; (viii) $\Delta = \Delta^*$ iff Δ is consistent and complete.*

Proof. (i). By definition $\varphi \in \Delta^{**}$ iff $\neg\varphi \notin \Delta^*$ iff $\neg\neg\varphi \in \Delta$.

(ii) The inclusion $\Delta \subseteq \Delta^{**}$ follows from the previous item and the formula $\varphi \to \neg\neg\varphi \in HT^2$. By definition $\varphi \in \Delta^{***}$ iff $\neg\neg\neg\varphi \notin \Delta$. The latter is equivalent to $\neg\varphi \notin \Delta$ due to $\neg\varphi \leftrightarrow \neg\neg\neg\varphi \in HT^2$, ie $\varphi \in \Delta^*$.

(iii) If $\Delta = \Gamma^*$, then $\Delta = \Delta^{**}$ by the previous item. Let $\neg\neg\varphi \in \Delta$, then $\varphi \in \Delta^{**} = \Delta$ by item (i). Conversely, if Δ is closed under the rule $\neg\neg\varphi/\varphi$, then $\Delta^{**} = \Delta$ by item (i).

(iv) This follows from the formula $\varphi \wedge \neg\varphi \to \neg\psi \vee \neg\neg\psi \in HT^2$ and the disjunction property of Δ.

(v) The consistency of Δ^* means that for every φ, either $\varphi \notin \Delta^*$ or $\neg\varphi \notin \Delta^*$. By definition of Δ^* this is equivalent to $\neg\varphi \in \Delta$ or $\neg\neg\varphi \in \Delta$ for every φ, ie to the weak completeness of Δ.

(vi) Let $\Delta \subseteq \Delta^*$. If $\varphi \in \Delta$, then $\varphi \in \Delta^*$, ie $\neg\varphi \notin \Delta$.

Assume Δ is consistent, then $\varphi \in \Delta$ implies $\neg\varphi \notin \Delta$, ie $\varphi \in \Delta^*$.

(vii) Let $\Delta^* \subseteq \Delta$. If $\neg\varphi \notin \Delta$, then $\varphi \in \Delta^*$, and so $\varphi \in \Delta$.

Let Δ be complete. If $\varphi \in \Delta^*$, then $\neg\varphi \notin \Delta$ and $\varphi \in \Delta$ by completeness.

(viii) Follows from items (vi) and (vii). □

Proposition 4 *If $\varphi \notin HT^2$, then*

$$\not\vdash_{HT^2} \{\varphi\} \cup \{\psi \wedge \neg\psi : \psi \in For\}.$$

Proof. Let $\varphi \notin HT^2$ and $\varphi \vee (\psi_0 \wedge \neg\psi_0) \vee \ldots \vee (\psi_n \wedge \neg\psi_n) \in HT^2$. The latter means by the extension lemma that for any prime theory Δ either $\varphi \in \Delta$ or Δ is inconsistent. However, φ is refutable at some HT^2 model $\langle \Delta_h, \Delta_{h'}, \Delta_t, \Delta_{t'} \rangle$, which means $\varphi \notin \Delta_h$. Since $\Delta_h \subseteq \Delta_{t'} = \Delta_h^*$, Δ_h is consistent by item (vi) of the previous lemma. □

By contraposition we obtain the following consequence.

Corollary 1 *HT^2 is closed under the rule*

$$\frac{\alpha \vee (\beta \wedge \neg\beta)}{\alpha}.$$

Minimal HT^2 models and partial stable models

The correspondence between HT^2 and WFS is established by the fact that, as we prove in this section, some minimal HT^2 models coincide with Przymusinki's partial stable models (Przymusinski 1994), when we restrict the syntax to that of normal logic programs. We recall next some basic definitions from Przymusinski's 3-valued setting, and proceed later to introduce a related multi-valued characterisation of HT^2 that will be very useful for comparison purposes.

A *3-valued interpretation* $\mathbf{T}$ is a mapping from the propositional signature At to the set of truth values[3] $\{0, 1, 2\}$ respectively standing for *false*, *undefined* and *true*. We can also represent the interpretation $\mathbf{T}$ as a pair of sets of atoms (T, T') satisfying $T \subseteq T'$ where $\mathbf{T}(p) = 0$ iff $p \notin T'$, $\mathbf{T}(p) = 2$ iff $p \in T$ and $\mathbf{T}(p) = 1$ otherwise (ie, $p \in T' \setminus T$). Notice that in the literature, it is perhaps more usual to find the alternative forms:

- a pair of sets (T^+, T^-), respectively denoting true and false atoms, with $T^+ \cap T^- = \emptyset$,
- a set L of literals which is *consistent* (it contains no pair $p, \neg p$).

but it is clear that we may equivalently use any of the three representations.

Two ordering relations among 3-valued interpretations are defined such that, if $\mathbf{T}_1 = (T_1, T_1')$ and $\mathbf{T}_2 = (T_2, T_2')$, then:

i) $\mathbf{T}_1 \leq \mathbf{T}_2$ iff $T_1 \subseteq T_2$ and $T_1' \subseteq T_2'$,

ii) $\mathbf{T}_1 \preceq \mathbf{T}_2$ iff $T_1 \subseteq T_2$ and $T_2' \subseteq T_1'$.

In (Przymusinski 1994), these relations receive the names of *standard* and *Fitting's ordering* respectively. The $\leq$ relation intuitively represents that one interpretation contains "less truth" than the other. It is equivalent to the condition: $\forall p \in V,\ \mathbf{T}_1(p) \leq \mathbf{T}_2(p)$, where $\leq$ denotes now the integer ordering for values. The other relation, $\preceq$, measures the degree of knowledge in terms of undefined atoms. Interpretations with shape (T, T) are called *complete* (they have no undefined atoms).

Given a 3-valued interpretation $\mathbf{T}$, Przymusinski's valuation of formulas is defined so that conjunction is the minimum, disjunction the maximum, and negation and implication are defined as:

- $\mathbf{T}(\neg\varphi) = 2 - \mathbf{T}(\varphi)$
- $\mathbf{T}(\varphi \to \psi) = \begin{cases} 2 & \text{if } \mathbf{T}(\varphi) \leq \mathbf{T}(\psi) \\ 0 & \text{otherwise} \end{cases}$

Additionally, valuation of truth constants is fixed as $\mathbf{T}(\top) = 2$, $\mathbf{T}(\bot) = 0$ and $\mathbf{T}(\mathbf{u}) = 1$ (the latter is a new constant representing undefinedness).

The definition of partial stable model relies on a generalisation of the program reduct (Gelfond & Lifschitz 1988) to the 3-valued case. Given a 3-valued interpretation $\mathbf{T}$, the *reduct* $\Pi^{\mathbf{T}}$ is formed by replacing each negated literal $\neg p$ in program Π by $\top, \mathbf{u}$ or $\bot$ depending on whether $\mathbf{T}(p)$ is $0, 1$ or 2 respectively.

Definition 5 (Partial stable model) *A 3-valued interpretation $\mathbf{T}$ is a* partial stable model *if it is the $\leq$-minimal model of $\Pi^{\mathbf{T}}$.*

In (Przymusinski 1994) it is shown that a positive program (like $\Pi^{\mathbf{T}}$ when Π is normal) has a unique $\leq$-minimal model. It was also shown that the *well founded model* $\mathbf{T}$ of a normal program Π is the $\preceq$-minimal partial stable model. Again, for the case of normal programs, it has also been proved that there exists a $\preceq$-minimum partial stable model, ie, a unique well founded model.

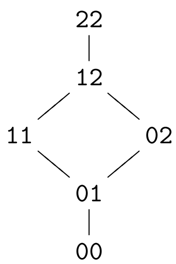
Now, let us return to HT^2 and consider a model $\mathcal{M} = \langle W, \leq, ^*, V \rangle$ denoting by H, H', T, T' the four sets of atoms respectively verified at each corresponding point or world h, h', t, t'. Since, by construction, $H \subseteq H'$ and $T \subseteq T'$, it is clear that we can represent $\mathcal{M}$ as a pair $\langle \mathbf{H}, \mathbf{T} \rangle$ of 3-valued interpretations $\mathbf{H} = (H, H')$ and $\mathbf{T} = (T, T')$. In this way, we could define the possible "situations" of a formula in HT^2 by using a pair of values xy with $x, y \in \{0, 1, 2\}$. Condition (3) restricts the number of these situations to the following six $\texttt{00} := \emptyset,\ \texttt{01} := \{t'\},\ \texttt{11} := \{h', t'\},\ \texttt{02} :=$

[3] In (Przymusinski 1994), 1 and 2 are respectively represented as 1/2 and 1 instead.

$\{t, t'\}$, $12 := \{h', t, t'\}$, $22 := \{h, h', t, t'\}$, where each set shows the worlds at which the formula is satisfied. Thus, an alternative way of describing HT^2 is by providing its logical matrix in terms of a 6-valued logic. As a result, the above setting becomes an algebra of 6 cones: $\mathcal{A}^{HT^2} := \langle\{00, 01, 11, 02, 12, 22\}, \vee, \wedge, \rightarrow, \neg\rangle$ where $\vee$ and $\wedge$ are set theoretical join and meet, whereas $\rightarrow$ and $\neg$ are defined as follows:

$$x \rightarrow y \;\; := \;\; \{w : w \leq w' \Rightarrow (w' \in x \Rightarrow w' \in y)\},$$
$$\neg\, x \;\; := \;\; \{w : w^* \notin x\}.$$

The only distinguished element is 22. The lattice structure of this algebra can be described by the condition $xy \leq zt \Leftrightarrow x \leq z \;\&\; y \leq t$ and is shown in Figure 1, together with the resulting truth-tables.



φ	$\neg\varphi$
00	22
01	11
11	11
02	00
12	00
22	00

$\rightarrow$	00	01	11	02	12	22
00	22	22	22	22	22	22
01	00	22	22	22	22	22
11	00	02	22	02	22	22
02	00	11	11	22	22	22
12	00	01	11	02	22	22
22	00	01	11	02	12	22

Given $V(\phi) = xy$ and $V(\psi) = zt$:

$$V(\phi \wedge \psi) = uv \Leftrightarrow u = min(x, z) \;\&\; v = min(y, t)$$
$$V(\phi \vee \psi) = uv \Leftrightarrow u = max(x, z) \;\&\; v = max(y, t)$$

Figure 1: Lattice structure and truth tables for the 6-valued HT^2 description.

An interesting observation is that, by the semantics, if $\langle\mathbf{H}, \mathbf{T}\rangle$ is a model then necessarily $\mathbf{H} \leq \mathbf{T}$, since it is easy to check that this condition is equivalent to $H \subseteq T$ and $H' \subseteq T'$. Moreover, for any theory Π note that if $\langle\mathbf{H}, \mathbf{T}\rangle \models \Pi$ then also $\langle\mathbf{T}, \mathbf{T}\rangle \models \Pi$.

The ordering $\leq$ can be extended to a partial ordering $\trianglelefteq$ among models as follows. We set $\langle\mathbf{H}_1, \mathbf{T}_1\rangle \trianglelefteq \langle\mathbf{H}_2, \mathbf{T}_2\rangle$ if (i) $\mathbf{T}_1 = \mathbf{T}_2$; (ii) $\mathbf{H}_1 \leq \mathbf{H}_2$. A model $\langle\mathbf{H}, \mathbf{T}\rangle$ in which $\mathbf{H} = \mathbf{T}$ is said to be *total*. Note that the term *total* model does not refer to the absence of undefined atoms. To represent this, we further say that a total partial equilibrium model is *complete* if $\mathbf{T}$ has the form (T, T).

We are interested here in a special kind of minimal model that we call a partial equilibrium model.

Definition 6 (Partial equilibrium model) *A model $\mathcal{M}$ of a theory Π is said to be a* partial equilibrium *model of Π if (i) $\mathcal{M}$ is total; (ii) $\mathcal{M}$ is minimal among models of Π under the ordering $\trianglelefteq$.*

In other words a partial equilibrium model of Π has the form $\langle\mathbf{T}, \mathbf{T}\rangle$ and is such that if $\langle\mathbf{H}, \mathbf{T}\rangle$ is any model of Π with $\mathbf{H} \leq \mathbf{T}$, then $\mathbf{H} = \mathbf{T}$. *Partial equilibrium logic* is the logic determined by truth in all partial equilibrium models of a theory. Formally we can define a nonmonotonic inference relation by:

Definition 7 (entailment) *Let Π be a theory, φ a formula and $\mathcal{PEM}(\Pi)$ the collection of all partial equilibrium models of Π. We say that Π entails φ, in symbols $\Pi \mid\!\sim \varphi$, if either (i) or (ii) holds: (i) $\mathcal{PEM}(\Pi) \neq \emptyset$ and $\mathcal{M} \models \varphi$ for every $\mathcal{M} \in \mathcal{PEM}(\Pi)$; (ii) $\mathcal{PEM}(\Pi) = \emptyset$ and φ is true in all HT^2-models of Π.*

In this definition we consider the skeptical or cautious entailment relation; a credulous variant is easily given if needed. Clause (ii) is needed because not all consistent theories have partial equilibrium models. Again (ii) represents one possible route to understanding entailment in the absence of intended models; other possibilities may be considered depending on context.

Finally, we proceed now to use the representation based on pairs of 3-valued interpretations to establish a straightforward correspondence to partial stable models. We begin by noting a property we will use below: examining the table for implication in Figure 1, it is easy to see that $M(\varphi \rightarrow \psi) = 22$ iff, given $M(\varphi) = xy$ and $M(\psi) = uv$, we have both $x \leq u$ and $y \leq v$. We also fix the HT^2 valuation of constants as $M(\top) = 22$, $M(\mathbf{u}) = 11$ and $M(\bot) = 00$.

Lemma 7 *For any $M = \langle\mathbf{H}, \mathbf{T}\rangle$ and any atom p: $M(\neg p) = M((\neg p)^{\mathbf{T}})$.*

Proof. Assume $\mathbf{T}$ has the form (T, T'). We have three cases, depending on $\mathbf{T}(p)$.

(i) For $\mathbf{T}(p) = 2$ the reduct is $(\neg p)^{\mathbf{T}} = \bot$, but we also have $p \in T$, ie, $M(p) \in \{02, 12, 22\}$ and so $M(\neg p) = 00 = M(\bot)$.

(ii) If $\mathbf{T}(p) = 1$ the reduct is $(\neg p)^{\mathbf{T}} = \mathbf{u}$ and we also have $p \in T' \setminus T$, ie, $M(p) \in \{01, 11\}$, which means $M(\neg p) = 11 = M(\mathbf{u})$.

(iii) If $\mathbf{T}(p) = 0$ the reduct is $(\neg p)^{\mathbf{T}} = \top$ and we also get $p \notin T'$, ie, $M(p) = 00$, which means $M(\neg p) = 22 = M(\top)$. □

Corollary 2 *For any HT^2 interpretation $M = \langle\mathbf{H}, \mathbf{T}\rangle$ and any normal logic program Π: $M \models \Pi$ iff $M \models \Pi^{\mathbf{T}}$.* □

Lemma 8 *Let Π be a positive logic program (possibly containing constants in the body) and let $\mathbf{T}$ be a 3-valued model of Π. Then, for any $M = \langle\mathbf{H}, \mathbf{T}\rangle$ and any rule $r \in \Pi$: $M(r) = 22$ iff $\mathbf{H}(r) = 2$.*

Proof. First, we note that for any atom or constant φ, $M(\varphi) = xy$ iff $\mathbf{H}(\varphi) = x$ and $\mathbf{T}(\varphi) = y$. Now, let r have the form $(A_i \wedge \cdots \wedge A_n \rightarrow B)$ and let $M(A_i) = x_i y_i$ and $M(B) = uv$. Condition $M(r) = 22$ means that both $min\{x_i\} \leq u$ and $min\{y_i\} \leq v$. However, the former is equivalent to $\mathbf{H}(r) = 2$, whereas the latter means $\mathbf{T}(r) = 2$ that in our case is always true, as $\mathbf{T}$ is a 3-valued model of Π. □

Theorem 2 *Let Π be a normal logic program. Then $\langle \mathbf{T}, \mathbf{T}\rangle$ is a partial equilibrium model of Π if and only if $\mathbf{T}$ is a partial stable model of Π.*

Proof. From Corollary 2, we can safely replace program Π by $\Pi^{\mathbf{T}}$ in the claim, provided that for determining if $\langle \mathbf{T}, \mathbf{T}\rangle$ is in partial equilibrium, we fix the second component of HT^2 models to $\mathbf{T}$. But now, as Π^T is positive, we can apply Lemma 8. In particular, we get first that $\langle \mathbf{T}, \mathbf{T}\rangle$ is an HT^2 model iff $\mathbf{T}$ is a 3-valued model. And similarly, we also get that for any $\mathbf{H} < \mathbf{T}$, $\langle \mathbf{H}, \mathbf{T}\rangle$ is an HT^2 model if $\mathbf{H}$ is a 3-valued model. □

Following (Przymusinski 1994), once partial stable models are captured, we can further minimise among them wrt the amount of information (ie, defined atoms) to obtain a well-founded model. Thus, an HT^2 *well-founded model* would just be a partial equilibrium model, $\preceq$-minimal among the partial equilibrium models of Γ.

Axiomatisation of HT^2

Although we have described HT^2 via a class of frames, it can be considered a logic in the sense that it defines a set of formulas: those valid on these frames. A more constructive (and perhaps more standard) definition of HT^2 is also possible using a calculus, that is, a set of axioms and inference rules.

Let HT^* be an N^* extension obtained by adding the following axioms:

A1. $-\alpha \vee - - \alpha$
A2. $-\alpha \vee (\alpha \rightarrow (\beta \vee (\beta \rightarrow (\gamma \vee -\gamma))))$
A3. $\bigwedge_{i=0}^{2}((\alpha_i \rightarrow \bigvee_{j\neq i} \alpha_j) \rightarrow \bigvee_{j\neq i} \alpha_j) \rightarrow \bigvee_{i=0}^{2} \alpha_i$
A4. $\alpha \rightarrow \neg\neg\alpha$
A5. $\alpha \wedge \neg\alpha \rightarrow \neg\beta \vee \neg\neg\beta$
A6. $\neg\alpha \wedge \neg(\alpha \rightarrow \beta) \rightarrow \neg\neg\alpha$
A7. $\neg\neg\alpha \vee \neg\neg\beta \vee \neg(\alpha \rightarrow \beta) \vee \neg\neg(\alpha \rightarrow \beta)$
A8. $\neg\neg\alpha \wedge \neg\neg\beta \rightarrow (\alpha \rightarrow \beta) \vee (\beta \rightarrow \alpha)$,

and the elimination of contradictions rule

$$\frac{\alpha \vee (\beta \wedge \neg\beta)}{\alpha} \qquad \text{(EC)}$$

Proposition 5 *The canonical frame $\mathcal{W}^{HT^*}$ satisfies the following properties: (i) $\mathcal{W}^{HT^*}$ is strongly directed; (ii) $\mathcal{W}^{HT^*}$ is of depth 3; (iii) each element of $\mathcal{W}^{HT^*}$ has at most two immediate successors; (iv) elements of $\mathcal{W}^{HT^*}$ satisfy all properties listed in Lemma 6.*

Proof. Items (i)-(iii) can be inferred from axioms A1, A2 and A3 respectively in the same way as for superintuitionistic logics determined by these axioms. Item (iv) holds since only axioms of HT^* were used in the proof of Lemma 6. □

Theorem 3 $HT^* = HT^2$.

Proof. The inclusion $HT^* \subseteq HT^2$ follows from the definition of HT^*. All its axioms are HT^2 tautologies, which can be verified directly. Moreover, HT^2 is closed under the rule (EC) by Corollary 1.

We prove the nontrivial inclusion $HT^* \subseteq HT^2$. Take some φ_0 non-provable in HT^* and construct an HT^2 model refuting φ_0.

Let Δ be a prime HT^* theory such that $\varphi_0 \notin \Delta$. By the rule (EC) if $HT^* \not\vdash \varphi_0$, then $HT^* \not\vdash \{\varphi_0\} \cup \{\beta \wedge \neg\beta | \beta \in For\}$. Thus we may assume that Δ is consistent.

1. Assume that Δ is consistent and complete. We prove that for any φ and ψ,

$$\varphi \rightarrow \psi \in \Delta \;\; \Leftrightarrow \;\; \varphi \notin \Delta \text{ or } \psi \in \Delta. \qquad (7)$$

The direct implication is obvious. If $\psi \in \Delta$, then $\varphi \rightarrow \psi \in \Delta$ by the positive axiom $\psi \rightarrow (\varphi \rightarrow \psi)$. Let $\varphi \notin \Delta$ and $\varphi \rightarrow \psi \notin \Delta$. By completeness $\neg\varphi, \neg(\varphi \rightarrow \psi) \in \Delta$, whence $\neg\neg\varphi \in \Delta$ by A6. Consistency and completeness of Δ imply in this case $\varphi \in \Delta$. This contradiction proves the desired equivalence.

By Lemma 6 we have $\Delta = \Delta^*$. Thus, the quadruple $\langle \Delta, \Delta, \Delta, \Delta\rangle$ is an HT^2 model refuting φ_0. Note that we have established also the following fact

Lemma 9 *If Δ is a complete prime HT^* theory closed under the rule $\neg\neg\varphi/\varphi$, then Δ is a maximal element of $\mathcal{W}^{HT^*}$.*

Proof. In the above reasoning the consistency of Δ was used to establish that Δ is closed under the rule $\neg\neg\varphi/\varphi$. Therefore, equivalence (7) holds for Δ. Assume Δ is not maximal. Let $\Delta' \in \mathcal{W}^{HT^*}$ and ψ_0 be such that $\Delta \subset \Delta'$ and $\psi_0 \in \Delta' \setminus \Delta$. On one hand, $-\psi_0 \in \Delta$ by (7). On the other hand, $-\psi_0 \notin \Delta$ by canonical model lemma. □

2. Assume that Δ is not complete, but is consistent and closed under $\neg\neg\varphi/\varphi$. Item (vi) of Lemma 6 implies $\Delta \subseteq \Delta^*$, where as item (iii) implies that $\Delta = \Gamma^*$. Thus $\Delta = \Delta^{**}$ by item (ii) of the same lemma. Note that Δ^* is complete since $\Delta^{**} \subseteq \Delta^*$. By the last lemma Δ^* is a maximal prime HT^* theory.

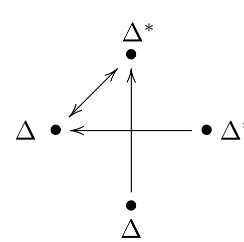
We claim that

$$\varphi \rightarrow \psi \in \Delta \;\Leftrightarrow\; (\varphi \notin \Delta \vee \psi \in \Delta) \wedge (\varphi \notin \Delta^* \vee \psi \in \Delta^*).$$

The direct implication is obvious. We prove the converse implication. Since Δ^* is maximal, the second conjunctive term means exactly that $\varphi \rightarrow \psi \in \Delta^*$. Therefore, $\neg(\varphi \rightarrow \psi) \notin \Delta$.

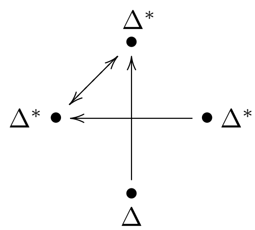
If $\psi \in \Delta$, then $\varphi \rightarrow \psi \in \Delta$. Assume $\varphi \notin \Delta$ and $\psi \notin \Delta$. By the rule $\neg\neg\varphi/\varphi$ we have $\neg\neg\varphi \notin \Delta$ and $\neg\neg\psi \notin \Delta$. By A7 at least one of formulas $\neg\neg\varphi$,$\neg\neg\psi$,$\neg(\varphi \rightarrow \psi)$ or $\neg\neg(\varphi \rightarrow \psi)$ belongs to Δ. Therefore, $\neg\neg(\varphi \rightarrow \psi) \in \Delta$ and so $\varphi \rightarrow \psi \in \Delta$.

One can see that all conditions of Lemma 4 hold for the quadruple $\langle \Delta, \Delta^*, \Delta, \Delta^*\rangle$. We have thus constructed the following countermodel for φ_0.



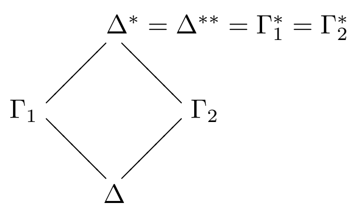
3. Now suppose $\varphi_0 \notin \Delta$, where Δ is a consistent weakly complete prime theory not closed under $\neg\neg\varphi/\varphi$. By item (vi) of Lemma 6 we have $\Delta \subseteq \Delta^*$, whence Δ^* is also weakly complete. By item (iii) of the same lemma Δ^* is

closed under the rule $\neg\neg\varphi/\varphi$. Therefore, Δ^* is complete and it is maximal in $\mathcal{W}^{HT^*}$ by Lemma 9. Item (v) of Lemma 6 implies that Δ^* is consistent. Consistency and completeness imply the equality $\Delta^{**} = \Delta^*$ by item (viii). The inclusion $\Delta \subseteq \Delta^*$ is proper since Δ is not closed under $\neg\neg\varphi/\varphi$. If there is no other proper extension of Δ, we obtain a countermodel $\langle \Delta, \Delta^*, \Delta^*, \Delta^* \rangle$.



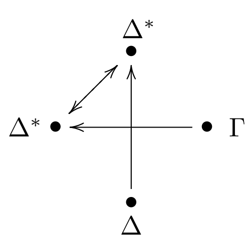
Assume that there is $\Gamma \in \mathcal{W}^{HT^*}$ such that $\Gamma \notin \{\Delta, \Delta^*\}$ and $\Delta \subseteq \Gamma$. Since Δ^* is maximal and $\mathcal{W}^{HT^*}$ is strongly directed, we have $\Gamma \subseteq \Delta^*$. By the antimonotonicity of the $*$-operation we obtain $\Gamma^* = \Delta^*$.

By item (iii) of Proposition 5 there exists at most two prime theories Γ_1 and Γ_2 between Δ and Δ^*. Assume $\Gamma_1 \neq \Gamma_2$. Then Γ_1 and Γ_2 are mutually incomparable by item (ii) of Proposition 5.



Let $\varphi \in \Gamma_1 \setminus \Gamma_2$ and $\psi \in \Gamma_2 \setminus \Gamma_1$. In this case $\varphi, \psi \in \Delta^{**}$ and $\neg\neg\varphi, \neg\neg\psi \in \Delta$ by item (i) of Lemma 6. By axiom A8 we obtain $(\varphi \to \psi) \vee (\psi \to \varphi) \in \Delta$. Since Δ possesses the disjunction property we have $\varphi \to \psi \in \Delta$ or $\psi \to \varphi \in \Delta$. Both cases contradict the choice of φ and ψ. In the first case, there is an extension Γ_1 of Δ such that $\varphi \in \Gamma_1$ and $\psi \notin \Gamma_1$. In the second case, we have $\psi \in \Gamma_2$ and $\varphi \notin \Gamma_2$.

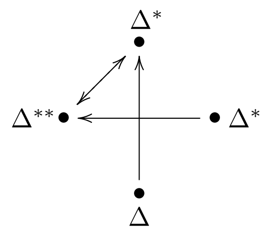
Thus, if there is a proper extension Γ of Δ different from Δ^*, it is unique and we obtain for φ_0 the countermodel $\langle \Delta, \Gamma, \Delta^*, \Delta^* \rangle$.



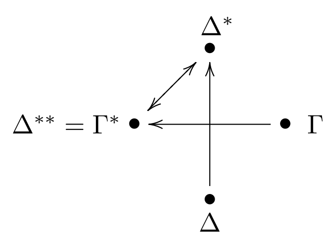
4. Consider the last case. Let $\varphi_0 \notin \Delta$, where Δ is consistent but is neither weakly complete nor closed under $\neg\neg\varphi/\varphi$. Again by Lemma 6 we have $\Delta \subseteq \Delta^*$ and Δ^* is closed under $\neg\neg\varphi/\varphi$. Since Δ is not weakly complete, Δ^* is not consistent by item (v) of the same lemma.

We show that Δ^* is complete. For any formula φ we have either $\neg\varphi \notin \Delta$ or $\neg\neg\varphi \notin \Delta$, consequently, $\varphi \in \Delta$ or $\neg\varphi \in \Delta$ by the definition of the $*$-operation.

Completeness of Δ^* implies its maximality and the inclusion $\Delta^{**} \subseteq \Delta^*$. Since $\Delta^* = \Delta^{***}$, we conclude that Δ^{**} is consistent. By this fact and inconsistency of Δ^* the inclusion $\Delta^{**} \subseteq \Delta^*$ is proper. By item (ii) of Lemma 6. $\Delta \subseteq \Delta^{**}$. This inclusion is also proper since Δ is not closed under $\neg\neg\varphi/\varphi$. If there is no other extension of Δ we obtain the countermodel $\langle \Delta, \Delta^*, \Delta^{**}, \Delta^* \rangle$.



If there is one more Γ such that $\Delta \subseteq \Gamma$, it is unique by item (iii) of Proposition 5. The strong directedness of $\mathcal{W}^{HT^*}$ and the maximality of Δ^* imply $\Gamma \subseteq \Delta^*$. Since $\mathcal{W}^{HT^*}$ is of depth 3, theories Δ^{**} and Γ are incomparable. By antimonotonicity of $*$ we have $\Gamma^* \subseteq \Delta^*$ and $\Delta^{**} \subseteq \Gamma^*$. Thus, $\Gamma^* \in \{\Delta^*, \Delta^{**}\}$. Taking into account $\Gamma \subseteq \Gamma^{**}$ we obtain $\Gamma^* = \Delta^{**}$. We arrive at the countermodel $\langle \Delta, \Gamma, \Delta^{**}(= \Gamma^*), \Delta^* \rangle$, and we are done.



□

Strong equivalence wrt partial equilibrium logic

We now establish a strong equivalence theorem for partial equilibrium logic. The notion of *strong equivalence* is important both conceptually and as a potential tool for simplifying nonmonotonic programs and theories and optimising their computation. For stable semantics strong equivalence can be completely captured in the logic HT (Lifschitz, Pearce, & Valverde 2001) and in ASP this fact has given rise to a lively programme of research into defining and computing different equivalence concepts, see eg (Eiter, Fink & Woltran 2006; Woltran 2004). In the case of WFS and p-stable semantics, however, to our knowledge until now, with the exception of (Nomikos, Rondogiannis & Wadge 2005), there have been no studies of strong equivalence and related notions.

We begin by noting that, when considering logic programs, equivalence under Przymusinki's 3-valued logic is not adequate for testing strong equivalence, much in the same way as classical logic is not suitable for strong equivalence under stable models. In fact, as happens in that case, it is not even suitable for checking regular equivalence. As an example, the programs $\{p \to q, \neg p \to q\}$ and $\{p \to q, \neg q \to p\}$ are equivalent under Przymusinski's 3-valued logic although they clearly have different well-founded models – the first one makes p false and q true, while the second leaves both atoms undefined.

Returning to arbitrary theories, in the present context we say that two propositional theories Γ_1 and Γ_2 are *strongly equivalent* if for any theory Γ, theories $\Gamma_1 \cup \Gamma$ and $\Gamma_2 \cup \Gamma$ have the same partial equilibrium models.

Proposition 6 *Theories Γ_1 and Γ_2 are strongly equivalent iff Γ_1 and Γ_2 are equivalent in HT^2.*

Proof. We consider the non-trivial direction. Let us assume that Γ_1 and Γ_2 have different models and construct a set of formulas Γ such that $\Gamma_1 \cup \Gamma$ and $\Gamma_2 \cup \Gamma$ have different partial equilibrium models.

Let $\langle \mathbf{H}, \mathbf{T} \rangle$ be an HT^2-model such that $\langle \mathbf{H}, \mathbf{T} \rangle \models \Gamma_1$ and $\langle \mathbf{H}, \mathbf{T} \rangle \not\models \Gamma_2$. Note that in this case we have $\langle \mathbf{T}, \mathbf{T} \rangle \models \Gamma_1$.

Case 1. Assume $\langle \mathbf{T}, \mathbf{T} \rangle \not\models \Gamma_2$. If $T = T'$, put $\Gamma := T$. In this case it can be easily seen that $\langle \mathbf{T}, \mathbf{T} \rangle$ is a partial equilibrium model of $\Gamma_1 \cup \Gamma$, but it is not a model of $\Gamma_2 \cup \Gamma$.

Assume $T \neq T'$ and $p_0 \in T' \setminus T$. Now we put

$$\Gamma := T \cup \{\neg p_0 \to q \mid q \in T'\}.$$

Clearly, $\langle \mathbf{T}, \mathbf{T} \rangle \not\models \Gamma_2 \cup \Gamma$. Let $\langle \mathbf{J}, \mathbf{T} \rangle \models \Gamma_1 \cup \Gamma$. Since $T \subseteq \Gamma$, we have $J = T$. The atom p_0 is refuted at t (by choice $p_0 \notin T$), therefore, $\neg p_0 \in \Delta_{h'}$ by $h'^* = t$. From the definition of Γ we have that $T' \subseteq \Delta_{h'}$, whence $J' = T'$. We have thus proved that $\langle \mathbf{T}, \mathbf{T} \rangle$ is a partial equilibrium model of $\Gamma_1 \cup \Gamma$.

Case 2. Let $\langle \mathbf{T}, \mathbf{T} \rangle \models \Gamma_2$. Fix some atom p_0 such that $p_0 \notin T'$ and p_0 does not occur in formulas of Γ_1 and Γ_2. We put

$$\Gamma := H \cup \{p_0 \leftrightarrow \neg p_0\} \cup \Pi_0 \cup \Pi_1 \cup \Pi_2 \cup \Pi_3,$$

where

$$\begin{array}{lcl} \Pi_0 & = & \{p_0 \to p \mid p \in H'\} \\ \Pi_1 & = & \{p \to q \vee p_0 \mid p, q \in T \setminus H\} \\ \Pi_2 & = & \{(p_0 \to p) \to q \mid p \in T' \setminus H',\ q \in T \setminus H\} \\ \Pi_3 & = & \{p_0 \wedge p \to q \mid p \in T' \setminus H,\ q \in T' \setminus H'\} \end{array}$$

Note that for any model $\langle \mathbf{H}, \mathbf{T} \rangle$, validity of $p_0 \leftrightarrow \neg p_0$ at this model means that $p_0 \in H' \setminus T$, ie p_0 is true exactly at h' and t'.[4]

Consider the new models $\langle \mathbf{H}, \mathbf{T}_1 \rangle$ and $\langle \mathbf{T}_1, \mathbf{T}_1 \rangle$, where $\mathbf{T}_1 := (T, T' \cup \{p_0\})$. Since p_0 is not involved in the computation of validity of formulas from Γ_1 and Γ_2 we still have

$$\langle \mathbf{H}, \mathbf{T}_1 \rangle \models \Gamma_1,\ \ \langle \mathbf{H}, \mathbf{T}_1 \rangle \not\models \Gamma_2,$$

$$\langle \mathbf{T}_1, \mathbf{T}_1 \rangle \models \Gamma_1,\ \ \langle \mathbf{T}_1, \mathbf{T}_1 \rangle \models \Gamma_2.$$

It is routine to check that both $\langle \mathbf{H}, \mathbf{T}_1 \rangle$ and $\langle \mathbf{T}_1, \mathbf{T}_1 \rangle$ are models of Γ, which proves that $\langle \mathbf{T}_1, \mathbf{T}_1 \rangle$ is not a partial equilibrium model of $\Gamma_1 \cup \Gamma$.

Let us prove that $\langle \mathbf{T}_1, \mathbf{T}_1 \rangle$ is a partial equilibrium model of $\Gamma_2 \cup \Gamma$. Assume that $\langle \mathbf{J}, \mathbf{T}_1 \rangle \models \Gamma_2 \cup \Gamma$. Since $H \subseteq \Gamma$ we have $H \subseteq J$, on the other hand the inclusion $\Pi_0 \subseteq \Gamma$ guarantees that $H' \subseteq J'$. One of these inclusions must be proper, because $\langle \mathbf{H}, \mathbf{T}_1 \rangle$ is not a model of Γ_2.

If $H \neq J$, the satisfiability of Π_1 implies the equality $J = T$ since p_0 is false at h. At the same time, formulas of Π_3 imply $J' = T' \cup \{p_0\}$. Indeed, let $p_1 \in J \setminus H$. Since both p_0 and p_1 are true at h', the validity of $p_0 \wedge p_1 \to q$ means that $q \in J$ for all $q \in T' \setminus H'$.

Assume now $J' \setminus H' \neq \emptyset$ and $p_2 \in J' \setminus H'$. All implications $p_0 \wedge p_2 \to q$, $q \in T' \setminus H'$, are in Π_2, whence $J' = T'$. At the same time $\langle \mathbf{J}, \mathbf{T}_1 \rangle \models p_0 \to p_2$. Now the equality $H = T$ follows from the fact that $(p_0 \to p_2) \to q \in \Pi_3$ for all $q \in T \setminus H$. $\square$

The above result can be extended to show that HT^2 also captures strong equivalence wrt well-founded models (ie, $\preceq$-minimal partial equilibrium models)[5].

Note that unlike in the case of strong equivalence under stable model semantics, we cannot assume in the general case that the formulas in Γ have the syntax of logic program rules. So when Γ_1 and Γ_2 have the form of logic programs, it is clear that HT^2 equivalence is a sufficient condition for strong equivalence, but it is an open question whether Γ can be taken to be a logic program (of whatever kind) in the case of non-equivalence.

Related work

There has been a number of attempts to provide a foundation for well-founded semantics; some are more or less logical in nature, others employ alternative mathematical methods. Of the former kind, we should mention:

- The approach of (Bochman 1998a; 1998b) which analyses several logic programming semantics, including WFS, in a generalised framework of Gentzen-style deduction. A strong point of Bochman's method of *bi-consequence relations* is its ability to capture different semantics within the same framework. The method is somewhat removed from ordinary logic and model theory and does not provide Hilbert-style axiomatisations. It remains to be seen whether it might complement the methods described here.
- Another type of technique can be found in (Rondogiannis & Wadge 2002) which proposes an infinite-valued logic to capture WFS. However it is unclear how this logic relates to other known multi-valued logics and how it can be used to extend the semantics beyond the format of normal programs.
- Another recent approach is that of (Alcântara, Damásio, & Pereira 2004) which studies WFS and variants using semantical frames. These are closely related to the HT^2-frames described here and in (Cabalar 2001). However no logical axiomatisation of the semantics is presented.

[4]If in the model under consideration $T \neq T'$, instead of adding to Γ the formula $p_0 \leftrightarrow \neg p_0$ for a new atom p_0 we can take $p_1 \in T' \setminus T$ and replace p_0 by $\neg p_1$ in $\Pi_1, \ldots, \Pi_4$. The negation $\neg p_1$ is true exactly at h' and t'.

Alternatively, we could pass to the conservative extension of HT^2 obtained by adding a new constant u together with axiom $u \leftrightarrow \neg u$. In this case, we replace p_0 by u.

[5]The details will be included in a sequel to the present paper.

- The method of representing WFS via embeddings into nonmonotonic modal logics. This has been explored notably in (Przymusinski 1995; Bonatti 1995). Though this is quite different from our aim to study WFS in a language close to the original logical syntax, we also hope in the future to examine modal counterparts of partial equilibrium logic and thereby make comparisons with frameworks such as those of (Przymusinski 1995; Bonatti 1995).

Among efforts to capture well-founded reasoning using other mathematical methods, we should mention:

- Argumentation theory as applied in (Bondarenko, Dung, Kowalski, & Toni 1997). This method has proved flexible enough to model several kinds of semantics for logic programming and nonmonotonic reasoning and implementations are now being developed (Dung, Kowalski, & Toni 2006). This approach can provide ways to enlarge the syntactical scope of well-founded reasoning. It remains to be seen how they relate to logical systems such as partial equilibrium logic.
- The infinite-game semantics recently proposed by (Rondogiannis & Wadge 2005). This appears at present to be restricted to the syntax of normal programs.

Conclusions and future work

We have proposed partial equilibrium logic as a general system of nonmonotonic logic to act as a foundation for the semantics of partial stable models and thereby for well-founded inference. Our approach has been to identify an underlying monotonic logical framework to be used as a basis. The natural choice is a logic in which partial stability can be expressed as a simple minimality condition with well-foundedness as a special case. The condition of equilibrium that captures stable models in the logic of here-and-there can be readily generalised to a minimality condition that captures partial stability in a logic HT^2 which corresponds in a natural way to HT. In this paper we have shown how the resulting logic has a six-valued truth matrix and can be axiomatised as an extension of Došen's logic N. Although the negation of HT^2, corresponding to the well-founded negation, is rather weak, intuitionistic negation is actually definable in HT^2. We have seen also that HT^2 captures the strong equivalence of theories in partial equilibrium logic.

The present paper reports on ongoing work that will continue to investigate many more issues in the foundations of WFS and p-stable semantics. Work currently in progress is examining a series of further topics including:

- the complexity of reasoning with partial equilibrium logic, for the general case as well as for specific classes of extended logic programs;
- the behaviour of partial equilibrium logic on disjunctive and nested logic programs and its comparison with other semantics;
- further study of the relation of HT^2 to HT and of partial equilibrium logic to equilibrium logic;
- general properties of partial equilibrium entailment;
- strong equivalence results for special classes of models such as the well-founded models defined above;
- how to add strong or explicit negation to partial equilibrium logic and compare this with the well-known system WFSX with explicit negation (Pereira & Alferes 1992).
- proof theory and implementation methods for partial equilibrium logic.

We hope to present some of this ongoing work in a sequel to the present paper.

References

Alcântara, J.; Damá-sio, C.; and Pereira, L. M. 2004. A frame-based characterisation of the paraconsistent well-founded semantics with explicit negation.

Bochman, A. 1998a. A logical foundation for logic programming I: Biconsequence relations and nonmonotonic completion. *Journal of Logic Programming* 35:151–170.

Bochman, A. 1998b. A logical foundation for logic programming II: Semantics for logic programs. *Journal of Logic Programming* 35:171–194.

Bonatti, P. 1995. Autoepistemic logics as a unifying framework for the semantic of logic programs. *Journal of Logic Programming* 22(2):91–149.

Bondarenko, A, Dung, P.M., Kowalski, R.A., Toni, F. 1997. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence 93(1-2)* , 63-101.

Brass, S., and Dix, J. 1994. A disjunctive semantics based on unfolding and bottom-up evaluation. In *IFIP'94 Congress, Workshop FG2: Disjunctive Logic Programming and Disjunctive Databases*, 83–91.

Cabalar, P. 2001. Well-founded semantics as two-dimensional here-and-there. In *Proceedings of ASP 01, 2001 AAAI Spring Symposium Series*.

de Jongh, D., and Hendriks, A. 2002. Characterization of strongly equivalent logic programs in intermediate logics. *Theory and Practice of Logic Programming* 3(3):259–270.

Dietrich, J. 1994. Deductive bases of nonmonotonic inference operations. Technical report, NTZ Report, Universität Leipzig.

Došen, K. 1986. Negation as a modal operator. *Rep. Math. Logic* 20:15–28.

Došen, K. 1999. Negation in the light of modal logic. In Gabbay, D & Wansing, H. (eds), *What is negation? Appl. Log. Ser. 13*. Dordrecht: Kluwer Academic Publishers. 77–86.

Dung, P.M., Kowalski, R.A., Toni, F. 2006. Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence 107(2)* , 14-159.

Eiter, T., and Fink, M. 2003. Uniform equivalence of logic programs under the stable model semantics. In *Proceedings of the Int. Conf. of Logic Programming, ICLP'03*. Mumbay, India: Springer.

Eiter, T., Fink, M, and Woltran, S. 2006. Semantical Characterizations and Complexity of Equivalences in Answer

Set Programming. *ACM Transactions on Computational Logic*. to appear.

Ferraris, P. 2005. Answer sets for propositional theories. In *Logic Programming and Nonmonotonic Reasoning. Proceedings LPNMR 05, LNAI 3662*, 119–131. Springer.

Ferraris, P., and Lifschitz, V. 2005. Weight constraints as nested expressions. *Theory and Practice of Logic Programming* 5:45–74.

Gelfond, M., and Lifschitz, V. 1988. The stable models semantics for logic programming. In *Proc. of the 5th Intl. Conf. on Logic Programming*, 1070–1080.

Lifschitz, V.; Pearce, D.; and Valverde, A. 2001. Strongly equivalent logic programs. *ACM Transactions on Computational Logic* 2:526–541.

Lifschitz, V.; Tang, L. R.; and Turner, H. 1999. Nested expressions in logic programs. *Annals of Mathematics and Artificial Intelligence* 25:369–389.

C. Nomikos, Rondogiannis, P., and Wadge, W. W. 2005. A Sufficient Condition for Strong Equivalence under the Well-Founded Semantics In *21st International Conference on Logic Programming (ICLP 2005), LNCS 3668*, 414-415. Springer.

Odintsov, S., and Pearce, D. 2005. Routley semantics for answer sets. In *Proceedings LPNMR05*. Springer LNAI.

Osorio, M.; Navarro, J.; and Arrazola, J. 2001. Equivalence in answer set programming. In *Proc. LOPSTR 2001, LNCS 2372*, 57–75. Springer.

Pearce, D. 1997. A new logical characterisation of stable models and answer sets. In *Non monotonic extensions of logic programming. Proc. NMELP'96. (LNCS 1216)*. Springer-Verlag. 57–70.

Pearce, D. 2004. Simplifying logic programs under answer set semantics. In Lifschitz, V., and Demoen, B., eds., *Proc. of ICLP04, LNCS 3132*. Springer.

Pearce, D., and Valverde, A. 2004a. Synonymous theories in answer set programming and equilibrium logic. In de Mántaras, R. L., and Saitta, L., eds., *Proc. of ECAI 2004*, 388–392. IOS Press.

Pearce, D., and Valverde, A. 2004b. Uniform equivalence for equilibrium logic and logic programs. In *Proc. of LPNMR'04, LNAI 2923*, 194–206. Springer.

Pereira, L. M., and Alferes, J. J. 1992. Well founded semantics for logic programs with explicit negation. In *Proceedings of the European Conference on Artificial Intelligence (ECAI'92)*, 102–106. Montreal, Canada: John Wiley & Sons.

Przymusinski, T. 1994. Well-founded and stationary models of logic programs. *Annals of Mathematics and Artificial Intelligence* 12:141–187.

Przymusinski, T. 1995. Static semantics for normal and disjunctive logic programs. *Annals of Mathematics and Artificial Intelligence* 14:323–357.

Rondogiannis, P., and Wadge, W. W. 2002. An infinite-valued semantics for logic programs with negation. In *Proceedings of the 8th European Conference of Logics in Artificial Intelligence (JELIA 2002)*, 456–467. Springer LNAI.

Rondogiannis, P., and Wadge, W. W. 2005. An infinite-game semantics for well-founded negation in logic programming. In *Proceedings of Games for Logic and Programming Languages (GaLoP)*, 77-91. ETAPS 2005, Edinburgh.

Routley, R., and Routley, V. 1972. The semantics of first degree entailment. *Noûs* 6:335–359.

Turner, H. 2001. Strong equivalence for logic programs and default theories (made easy). In *Proc. of the Sixth Int'l Conf. on Logic Programming and Nonmonotonic Reasoning (LPNMR'01)*, 81–92.

Turner, H. 2004. Strong equivalence for causal theories. In Lifschitz, V., and Niemelä, I. (eds.), *Proceedings of Logic Programming and Nonmonotonic Reasoning, LPNMR 2004*. Springer LNAI 2923.

van Gelder, A.; Ross, K. A.; and Schlipf, J. S. 1991. The well-founded semantics for general logic programs. *Journal of the ACM* 38(3):620–650.

Woltran, S. 2004. Characterizations for relativized notions of equivalence in answer set programming. In Alferes, J. J., and Leite, J., eds., *Proceedings of the 9th European Conference on Logics in Artificial Intelligence (JELIA'04)*.