The Well-Founded Semantics Is the Principle of Inductive Definition, Revisited

Marc Denecker
Department of Computer Science
K.U. Leuven
3001 Heverlee, Belgium
marc.denecker@cs.kuleuven.be

Joost Vennekens
Campus De Nayer | K.U. Leuven
Department of Computer Science
2860 Sint-Katelijne-Waver, Belgium
joost.vennekens@cs.kuleuven.be

Abstract

In the past, there have been several attempts to explain logic programming under the well-founded semantics as a logic of inductive definitions. A weakness in all is the absence of an obvious connection between how we understand various types of informal inductive definitions in mathematical text and the complex mathematics of the well-founded semantics. In this paper, we close this gap. We formalize the induction process in the most common principles and prove that the well-founded model construction generalizes them all.

Introduction

Ever since the well-founded semantics was first defined (Van Gelder, Ross, and Schlipf 1991), researchers have referred to the concept of *an inductive definition*, as we know it from mathematics, to explain the intuitions behind this formal semantics. This link was made explicit in several publications (Denecker 1998; Denecker, Bruynooghe, and Marek 2001; Denecker and Ternovska 2008), where we gave various arguments that, under the well-founded semantics, each rule in a set of rules can be seen as an (inductive or base) case of an inductive definition. A weakness in all is the absence of an obvious connection between how we understand various types of informal inductive definitions in mathematical text and the complex mathematics of the well-founded semantics. This paper aims to close this gap.

To open the discussion, let us consider two prototypical examples illustrating the two most common forms of inductive definitions: the monotone inductive Definition 1 of the transitive closure of a graph and the Definition 2 of the satisfaction relation of propositional logic by induction over the sub-formula order.

Definition 1. *The reachability graph R of a directed graph G is defined inductively:*

- $(x, y) \in R$ *if* $(x, y) \in G$*;*
- $(x, y) \in R$ *if there exists a vertex* z *such that* $(x, z), (z, y) \in R$.

Definition 2. *Given a propositional vocabulary Σ, the satisfaction relation $\models$ between Σ-structures and Σ-formulas of propositional logic is defined by induction over the structure of formulas:*

- $I \models P$ *if* P *is a propositional symbol and* $P \in I$.
- $I \models \alpha \wedge \beta$ *if* $I \models \alpha$ *and* $I \models \beta$.
- $I \models \alpha \vee \beta$ *if* $I \models \alpha$ *or* $I \models \beta$ *(or both).*
- $I \models \neg\alpha$ *if* $I \not\models \alpha$.

(Inductive) definitions serve to define formal objects, but they are not formal objects themselves. As such, we will refer to them as *informal definitions*. They are commonly phrased in natural language as collections of informal rules, possibly with an induction order. They define a set (or more than one, in case of simultaneous induction) *in terms of* other sets, which we will call *parameters*. E.g., Definition 1 has graph G as a parameter and Definition 2 the vocabulary Σ. Informal inductive definitions are broadly used in mathematics, broadly understood and, despite their informal nature, they are of mathematical precision. The set defined by it can be characterized in two quite different ways: "non-constructively", as the least set closed under rule application, and "constructively", as the set obtained by iterated rule application. By Tarski's cherished result on the least fixpoint of monotone operators, both principles coincide.

It comes as a surprise to many but the last statement is only half true. Def. 2 is non-monotone due to its 4th rule, and Tarski's result does not apply for it. There are infinitely many minimal sets that are closed under these rules and some are just weird: e.g., in some of them $\{P\} \models \neg P$ holds (See Example 1). What this shows is that the constructive principle is the more fundamental of the two. An inductive definition defines a set by describing how to construct it through an *induction process*. This process starts from the empty set, proceeds by applying rules till the set is closed under the rules. In case of induction over a well-founded order, rules must be applied "along" the specified order. This is the intuition that we guess all of us share and that will be formalized below.

The precision of informal inductive definitions makes them an ideal target for a formal empirical study. This is what we do in the next section, We define a simple formal, rule-based syntax with atomic heads and first order logic (FO) bodies. This leads to the following representation of the above informal definitions.

$$\left\{ \begin{array}{l} \forall x \forall y (R(x, y) \leftarrow G(x, y)) \\ \forall x \forall y (R(x, y) \leftarrow \exists z (R(x, z) \wedge R(z, y)) \end{array} \right\}$$

$$\left\{ \begin{array}{l} \forall i \forall p (Sat(i,p) \leftarrow Atom(p) \land In(p,i)) \\ \forall i \forall f \forall g (Sat(i, And(f,g)) \leftarrow Sat(i,f) \land Sat(i,g)) \\ \forall i \forall f \forall g (Sat(i, Or(f,g)) \leftarrow Sat(i,f) \lor Sat(i,g)) \\ \forall i \forall f (Sat(i, Not(f)) \leftarrow \neg Sat(i,f)) \end{array} \right\}$$

These two sets of rules, denoted respectively Δ_{TC} and $\Delta_{\models}$, will serve as examples throughout the paper. We then formalize several semantic aspects of informal inductive definitions: the induction process, the induction order, two sorts of inductive definitions and their generalization, etc. This study reveals several unexpected but fundamental aspects of inductive definitions. After that, we are on mathematical ground and we will prove that the set defined by a definition is its well-founded model.

The main contributions of the paper are then as follows. First, we provide a number of *definitions from first principles* that formalize intuitions and implicit conventions regarding informal inductive definitions. We do this for monotone inductive definitions, definitions by induction over a well-founded order as well as for their generalization, iterated inductive definitions. Second, we prove the equivalence of formal rule sets under the formalized semantics and the well-founded semantics.

Due to space restrictions, proofs are omitted.

Formal definitions and natural inductions

A vocabulary Σ consists of a set of non-logical predicate, function, constant and variable symbols. As usual, terms are built from the function, constant and variable symbols of Σ, and formulas are built from atomic formulas (a predicate symbol applied to a tuple of terms), the logical symbols $\mathbf{t}$ (true), $\mathbf{f}$ (false), $=$ (identity), connectives $\land, \lor, \neg$ and quantifiers $\exists, \forall$.

An occurrence of a variable x in a formula φ is free if it does not occur in a sub-formula $\exists x \psi$ or $\forall x \psi$ of φ. The set $free(\varphi)$ is the set of all variables with a free occurrence in φ. A sentence is a formula φ with $free(\varphi) = \emptyset$. An occurrence of a sub-formula φ in ψ is called *positive* in ψ if it occurs in the scope of an even number of negations, otherwise it is called *negative*. A formula φ is positive with respect to a set σ of predicate symbols if there are no atoms $P(\bar{t})$ with $P \in \sigma$ that have a negative occurrence in φ.

Σ-structures $\mathfrak{A}$ consist of a domain $D^{\mathfrak{A}}$ and an assignment of appropriate values $\tau^{\mathfrak{A}}$ to symbols $\tau \in \Sigma$. The value $P^{\mathfrak{A}}$ of a predicate symbol P/n is a function from $(D^{\mathfrak{A}})^n$ to $\{\mathbf{t}, \mathbf{f}\}$ (hence the characteristic function of a set of n-tuples). Let D be a domain. A *domain atom* of D is a pair $(P, \bar{a})$ with P/n a predicate symbol and $\bar{a} \in D^n$. Abusing notation, we write domain atoms as atoms $P(\bar{a}), Q(\bar{b})$. We use A, B, C as mathematical variables for domain atoms. We define $P(\bar{a})^{\mathfrak{A}} = P^{\mathfrak{A}}(\bar{a})$. For a given set σ of predicate symbols, we denote the set of domain atoms with predicates in σ by At^{σ}_{D}. There is a one-to-one correspondence between subsets of At^{σ}_{D} and σ-structures with domain D. We often exploit this correspondence in our notations by treating σ-structures as such sets. E.g., $\emptyset$ stands for the structure $\mathfrak{A}$ with $P^{\mathfrak{A}} = \emptyset$, for every $P \in \sigma$. We write $P(\bar{a}) \in \mathfrak{A}$ to denote that $P^{\mathfrak{A}}(\bar{a}) = \mathbf{t}$, and $\mathfrak{A} \subseteq \mathfrak{B}$ to denote that $P(\bar{a})^{\mathfrak{A}} \leq P(\bar{a})^{\mathfrak{B}}$, for every $P(\bar{a}) \in At^{\sigma}_{D}$. This order based on $\mathbf{f} < \mathbf{t}$ is sometimes called the *truth order*.

Definition 3. *A (formal)* definition *over Σ is a set of definition rules of the form*

$$\forall \bar{x}\ (P(\bar{t}) \leftarrow \phi)$$

where ϕ is a FO formula and $P(\bar{t})$ is an atomic formula over Σ such that P is not $=$.

We call $P(\bar{t})$ the head of the rule, and ϕ the body. The connective $\leftarrow$ is called *definitional implication*. Rules of this kind are similar in nature to and generalize *productions* in (Martin-Löf 1971). A predicate appearing in the head of a rule of a definition Δ is called a *defined predicate* of Δ; all other non-logical symbols in Δ are called *parameters* of Δ. The sets of defined predicates and parameters of Δ are denoted by $def(\Delta)$ and $pars(\Delta)$, respectively. We assume without loss of generality that every rule is of the form $\forall \bar{x}\ (P(\bar{x}) \leftarrow \phi)$ where $\bar{x}$ is a tuple of distinct variables. Indeed, every rule $\forall \bar{x}\ (P(\bar{t}) \leftarrow \phi)$ can be put into that form as $\forall \bar{y}\ (P(\bar{y}) \leftarrow \exists \bar{x} (\bar{y} = \bar{t} \land \phi))$. Below, a domain atom $P(\bar{a})$ of a defined predicate of Δ is called a *defined domain atom*. Note that this formal notion of definition does not (yet) include an induction order.

A definition is always evaluated in a *context* providing values for its parameters. We formalize this.

Definition 4. *We call a $pars(\Delta)$-structure $\mathcal{O}$ a* context *of Δ.*

Example 1. *The definition $\Delta_{\models}$ of Sat formalizes the informal Definition 2 of satisfaction. Its parameter symbols are $Atom, In, And, Or$ and Not. We view $\Delta_{\models}$ as a many-sorted definition, with two sorts for structures and formulas. Any set S of propositional symbols induces a context $\mathcal{O}_S$ which is the $pars(\Delta)$-structure defined as follows:*

- *$D^{\mathcal{O}_S} = PropF(S) \cup Struct(S)$, where $PropF(S)$ is the set of propositional formulas over S and $Struct(S)$ is the set of propositional S-structures.*
- *$And^{\mathcal{O}_S}$ is the function that maps pairs of formulas $(\psi, \phi) \in PropF^2$ to $\psi \land \phi$. The functions $Or^{\mathcal{O}_S}$ and $Not^{\mathcal{O}_S}$ are defined in a similar vein.*
- *Finally, $In^{\mathcal{O}_S}$ is $\{(I,p) | I \in Struct(S), p \in I\}$.*

As an example, defined domain atoms for $S = \{P\}$ are of the form $Sat(\{P\}, P), Sat(\{\}, P \land \neg P), \ldots$. The value $t^{\mathcal{O}_s}$ of the term $t = And(P, Not(P))$ is the formula $P \land \neg P$.

From here till the end of this section, we assume the presence of a definition Δ, and a context $\mathcal{O}$ for Δ with domain D. The induction process associated with Δ in context $\mathcal{O}$ will be a sequence of $def(\Delta)$-structures $\mathfrak{A}$ with domain D. By $\mathcal{O} \circ \mathfrak{A}$, we denote the structure $\mathfrak{B}$ with domain D and such that $P^{\mathfrak{B}} = P^{\mathfrak{A}}$ if $P \in def(\Delta)$ and $\tau^{\mathfrak{B}} = \tau^{\mathcal{O}}$ if $\tau \in pars(\Delta)$. In the sequel, we frequently evaluate formulas with respect to structures $\mathcal{O} \circ \mathfrak{A}$. Because $\mathcal{O}$ is given and fixed, we take the liberty to write only the "variable" part and write e.g., $\mathfrak{A} \models \varphi$ instead of $\mathcal{O} \circ \mathfrak{A} \models \varphi$, or $A^{\mathfrak{A}}$ instead of $A^{\mathcal{O} \circ \mathfrak{A}}$, etc.

The next definitions are formalizations of the concept of an element being *derivable* from a definition, and a set being *closed* or *saturated* under a definition.

Definition 5. *Given a context $\mathcal{O}$, we say that $P(\bar{a})$ is derivable from rule $\forall \bar{x}(P(\bar{x}) \leftarrow \varphi)$ in $(\mathcal{O}\circ)\mathfrak{A}$ if $\varphi^{\mathfrak{A}[\bar{x}:\bar{a}]} = \mathbf{t}$. We say that $P(\bar{a})$ is Δ-derivable in $\mathfrak{A}$ if it is derivable from a rule of Δ in $\mathfrak{A}$. Below, we denote this by $\mathfrak{A} \vdash_\Delta P(\bar{a})$.*

Definition 6. *Given a context $\mathcal{O}$, we say that $\mathfrak{A}$ is closed (or saturated) on a set S of defined domain atoms under Δ in $\mathcal{O}$ if for every $A \in S$, $\mathcal{O} \circ \mathfrak{A} \vdash_\Delta A$ implies $A \in \mathfrak{A}$. In general, $\mathfrak{A}$ is* closed *(or saturated) under Δ (in $\mathcal{O}$) if it is closed on $At_D^{def(\Delta)}$ under Δ in $\mathcal{O}$.*

Example 2. *Let Δ_{ev} be the following non-monotone definition:*

$$\left\{ \begin{array}{l} Even(0) \leftarrow \\ \forall x(Even(x+1) \leftarrow \neg Even(x)) \end{array} \right\} \qquad (1)$$

It can be viewed as the formalization of an informal definition of the even numbers by induction over the standard order of numbers: 0 is even; if n is not even then n+1 is even. This definition is structurally equivalent to the first and last rule of Definition 2. For succinctness, we abbreviate $Even$ often to Ev.

The context $\mathcal{O}$ is the structure of the natural numbers, with the standard interpretation of $0, 1$ and $+$. Consider the following sets, for every $N \in \mathbb{N}$:

$$\{Ev(0), Ev(2), Ev(4), \dots\}$$
$$\{Ev(0), Ev(2), \dots, Ev(2N), Ev(2N+1), Ev(2N+3), \dots\}$$

Each of these sets is closed under Δ_{ev}. None has a subset that is closed. Thus, the defined set is not the least set closed under the rules. A similar phenomenon arises for Definition 2.

The main concept of this paper is that of the induction process. It is formalized as follows.

Definition 7. *A* natural induction *$\mathcal{N}$ of Δ in $\mathcal{O}$ (with domain D) is an increasing sequence $(\mathfrak{A}_\alpha)_{0 \le \alpha \le \beta}$ of $def(\Delta)$-structures with domain D such that:*

- *$\mathfrak{A}_0$ is the empty structure $\emptyset$.*
- *For each successor ordinal $i+1 \le \beta$, for each domain atom $A \in \mathfrak{A}_{i+1} \setminus \mathfrak{A}_i$, A is derivable from Δ in $\mathfrak{A}_i$ ($\mathfrak{A}_i \vdash_\Delta A$). We say that A is derived at i and define $\|A\|_{\mathcal{N}} := i$, the rank of A in $\mathcal{N}$.*
- *For each limit ordinal $\lambda \le \beta$, $\mathfrak{A}_\lambda = \bigcup_{\alpha<\lambda} \mathfrak{A}_\alpha$.*

A natural induction is called terminal *if $\mathfrak{A}_\beta$ is closed under Δ (in $\mathcal{O}$).*

Natural inductions will be denoted compactly as a sequence of the (disjoint) sets of atoms that are derived at each step. E.g.,

$$\rightarrow \{A_1, \dots, A_n\} \rightarrow \{B_1, \dots, B_m\} \rightarrow \dots$$

derives the A_i's in step 0 and the B_j's in step 1. If such a set is a singleton we drop the brackets.

Example 3. *Consider the formal definition Δ_{TC} formalizing the transitive closure Definition 1. Take context $\mathcal{O}$ such that $D = \{a, b, c\}$, $G^{\mathcal{O}} = \{(a,a), (b,c), (c,b)\}$. All terminal natural inductions converge to $\{(a,a), (b,c), (c,b), (b,b), (c,c)\}$. E.g.,*

$$\rightarrow T(a,a) \rightarrow T(b,c) \rightarrow T(c,b) \rightarrow T(b,b) \rightarrow T(c,c)$$
$$\rightarrow \{T(c,b), T(b,c)\} \rightarrow T(c,c) \rightarrow T(b,b) \rightarrow T(a,a)$$

The example illustrates some basic points about informal inductive definitions. They indeed define the concept in terms of the parameters, by describing how to construct it through iterated rule application. The description of the construction process is however highly *non-deterministic*: rules can be applied in many orders. From a practical point of view, it is all-important that such induction sequences converge to the same fixpoint, otherwise the definition would be ambiguous. For a monotonic informal definition, such as Definition 1, the order of rule application is not important, because—given the graph G—all sequences converge to the intended set, which is the least relation that is closed under the rules. In the above framework of natural induction sequences, this can be proven formally. First we propose a formalization of the notion of monotone inductive definition.

Definition 8. *We call Δ monotone in $\mathcal{O}$ if for all pairs of $def(\Delta)$-structures $\mathfrak{A} \subseteq \mathfrak{B}$, for all defined domain atoms A, if $\mathfrak{A} \vdash_\Delta A$ then $\mathfrak{B} \vdash_\Delta A$.*

Proposition 1. *Each terminal natural induction of a monotone Δ in $\mathcal{O}$ converges to the least $def(\Delta)$-structure $\mathfrak{A}$ that is closed under Δ in $\mathcal{O}$.*

This proposition is not difficult to prove but follows from the general Theorem 3 below.

The convergence property does not hold for non-monotone definitions. The problem is that the body of a non-monotone rule may eventually become false, after it has already been true. Natural inductions that apply a rule during the "window" where its body holds will derive its head, whereas natural inductions that miss this window may not.

Example 4 (Continuation of Example 1). *Consider definition $\Delta_\models$ in the context of the structure $\mathcal{O}_S$ for the singleton vocabulary $S = \{P\}$. There are only two structures for the vocabulary S, namely, $\emptyset$ and $\{P\}$. Below is an initial segment of a stepwise natural induction that derives a wrong fact.*

$$\rightarrow Sat(\{P\}, \neg P) \rightarrow Sat(\{P\}, P) \rightarrow \dots$$

In the first step, with $\mathfrak{A}_0 = \emptyset$, all instances of the rule for negation are applicable. Here, we use it to derive $Sat(\{P\}, \neg P)$. However, the next step applies the base rule to derive $Sat(\{P\}, P)$, thus falsifying the condition of the rule that was applied in the first step.

We realize that the role of the induction order in informal definitions is to delay the application of rules until it is safe to do so, that is, until later rule applications cannot longer falsify the premise of a rule that has been applied before. We now formally define the notion of definition by induction over a well-founded order. For brevity, we call it an *ordered definition*.

Definition 9. *Given is a domain D. An ordered definition is a pair $(\Delta, \prec)$ with Δ a definition and $\prec$ a strict well-founded order on $At_D^{def(\Delta)}$.*

Recall that a strict order is irreflexive, transitive and asymmetric. A strict order $\prec$ is well-founded if it has no infinite descending chains $x_0 \succ x_1 \succ x_2 \succ \dots$.

Example 5. *The induction order of informal Definition 2 is the sub-formula order. In the context of definition $\Delta_{\models}$ and a structure $\mathcal{O}_S$, it corresponds to a strict well-founded order $\prec$ on domain atoms, where $Sat(I, \psi) \prec Sat(J, \phi)$ if $I = J$ and ψ is a strict sub-formula of ϕ.*

The induction order provided with an informal definition serves to constrain the order of rule application in natural inductions. How does this work? Intuition says that no rule should be applied to derive a fact as long as there are derivable but not yet derived facts that are strictly smaller. E.g., assume that at some point in the induction process $I \models \varphi$ is derivable. We are allowed to do so only if there is no strict subformula ϕ of φ for which $I \models \varphi$ is derivable but was not yet derived. In general, we can derive $P(\bar{a})$ if the current set $\mathfrak{A}_i$ is saturated on atoms preceding $P(\bar{a})$ in the induction order. This is formalized in the following definition.

Recall that the rank $\|A\|_{\mathcal{N}}$ of A in a natural induction $\mathcal{N}$ is the ordinal i such that $A \in \mathfrak{A}_{i+1} \setminus \mathfrak{A}_i$.

Definition 10. *A natural induction $\mathcal{N}$* respects *$\prec$ (w.r.t. Δ and $\mathcal{O}$) if for any domain atom A with $\|A\|_{\mathcal{N}} = i$, $\mathfrak{A}_i$ is saturated on $\{B \mid B \prec A\}$ (under Δ in $\mathcal{O}$).*

We say that $\mathcal{N}$ follows *$\prec$ if for every A and B derived by $\mathcal{N}$, $A \prec B$ implies $\|A\|_{\mathcal{N}} < \|B\|_{\mathcal{N}}$.*

Example 6. *The natural induction of Example 4:*

$$\rightarrow Sat(\{P\}, \neg P) \rightarrow Sat(\{P\}, P) \rightarrow \ldots$$

does not respect the subformula order. The atom $Sat(\{P\}, \neg P)$ is derived in the first step, when the empty set is not saturated in $\{A \mid A \prec Sat(\{P\}, \neg P)\}$ since $Sat(\{P\}, P)$ is derivable.

In general the induction process is highly underspecified, even if an induction order is given.

Example 7. *(Example 5 continued). Natural inductions of the informal Definition 2 will derive $I \models \varphi$ only after the satisfaction of all sub-formulas has been derived. This constrains the order of rule application, but much freedom is left. There are infinitely many such natural inductions. A few non-terminal ones are:*

$$\rightarrow Sat(\{P\}, P) \rightarrow Sat(\{P\}, P \wedge P) \rightarrow Sat(\{P\}, \neg\neg P)$$

$$\rightarrow Sat(\{P\}, P) \rightarrow Sat(\{P\}, \neg\neg P) \rightarrow Sat(\{P\}, P \vee P)$$

Note that both natural inductions respect the sub-formula order and follow it. Intuition suggests that these sequences can be extended to converging terminal natural inductions, and this will be proven below.

Given our experience with informal definitions, we expect some "good" properties of natural inductions that respect the induction order $\prec$: (1) that they all converge, (2) that such a natural induction *follows* the induction order, (3) that once an element is derived, it remains derivable, and (4) that in the limit, the defined set is the intended one. However, none of these properties hold right now.

The major question is related to (1). It is essential that all natural inductions that respect $\prec$ converge. However, it is straightforward to see that this is not the case. Take the empty induction order $\emptyset$ for the definition $\Delta_{\models}$ in context $\mathcal{O}_S$. This order is a strict well-founded order and all natural inductions respect it in a trivial way. As we saw in Example 4, not all of these natural inductions converge.

As for (2), a counterexample is below.

Example 8. *Consider the order $P \prec Q$ and definition:*

$$\left\{ \begin{array}{l} Q \leftarrow \mathbf{t} \\ P \leftarrow Q \end{array} \right\}$$

Here is a terminal natural induction:

$$\rightarrow Q \rightarrow P$$

It obviously does not follow $\prec$ since $P \prec Q$. However, it respects $\prec$. In the first step, when Q is derived, the structure $\mathfrak{A}_0 = \emptyset$ is saturated on $\{A \mid A \prec Q\} = \{P\}$, since P is not derivable. In the second step, $\mathfrak{A}_1 = \{Q\}$ is trivially saturated on $\{A \mid A \prec P\} = \{\}$.

A counterexample for (3) and (4) is given below.

Example 9. *We reconsider Δ_{ev} and $\mathcal{O}$ from Example 2.*

$$\left\{ \begin{array}{l} Even(0) \leftarrow \\ \forall x (Even(x+1) \leftarrow \neg Even(x)) \end{array} \right\} \qquad (2)$$

Let $\prec$ be the order induced by the standard order on the natural numbers. That is, $Ev(n) \prec Ev(m)$ if $n < m$. This is a total order, and the unique terminal natural induction that respects and follows it constructs the set of even numbers:

$$\rightarrow Ev(0) \rightarrow Ev(2) \rightarrow Ev(4) \rightarrow \ldots \rightarrow Ev(2n) \rightarrow \ldots$$

Now take the following non-standard induction order:

$$Ev(1) \prec Ev(0) \prec Ev(2) \prec Ev(3) \prec \ldots$$

Also this is a total strict well-founded order. The unique terminal natural induction that respects $\prec$ is:

$$\rightarrow Ev(1) \rightarrow Ev(0) \rightarrow Ev(3) \rightarrow Ev(5) \rightarrow \ldots$$

Note that $Ev(1)$ is not longer derivable after step 2. Also, it clearly does not define the intended set.

In non-monotone informal definitions, we impose a well-founded induction order to obtain convergence of the induction process. However, it is clear from the above examples that in selecting the induction order, great care is required. In general, imposing an unsuitable induction order w.r.t. Δ and $\mathcal{O}$ may have a number of undesired effects as just shown.

Something clearly wrong with the second induction order in the above example is that it does not *match* the structure of the given rules. In particular, $Ev(1)$ is defined in terms of $Ev(0)$, even though $Ev(0)$ is strictly larger than $Ev(1)$ in the proposed induction order. This would be unacceptable in an informal definition.

The above examples expose one of the implicit conventions of informal inductive definitions. This is that the induction order should "match" the structure of the rules of a definition. Intuitively, this means that defined facts may only "depend" on facts that are strictly smaller in the induction order. We now formalize this.

First we formalize the notion of dependency relation. Below, for binary relation $\propto$, $\mathfrak{A}|_{\propto A}$ denotes $\mathfrak{A} \cap \{B \mid B \propto A\}$.

Definition 11. *A binary relation $\propto$ on $At_D^{def(\Delta)}$ is a* dependency relation *of Δ in $\mathcal{O}$ if $\propto$ is transitive and for all A and all $\mathfrak{A}, \mathfrak{B}$, if $\mathfrak{A}|_{\propto A} = \mathfrak{B}|_{\propto A}$ then $\mathfrak{A} \vdash_\Delta A$ iff $\mathfrak{B} \vdash_\Delta A$.*

If $\propto$ is a dependency relation, then for any defined atom A, the set $\{B \mid B \propto A\}$ is (a superset of) the set of atoms on which A depends. Indeed, in any pair of structures that coincide on this set, A is derivable in both or in none.

That an induction order "matches" the rules of a definition simply means that $\prec$ is a dependency relation.

Definition 12. *We say that $\prec$* strictly orders Δ *in $\mathcal{O}$ if $\prec$ is a strict well-founded order and a dependency relation of Δ in $\mathcal{O}$.*

Example 10. *In case of definition Δ_{ev} of even numbers and the structure $\mathcal{O}$ of Example 2, we see that the first order*

$$Ev(0) \prec Ev(1) \prec Ev(2) \prec Ev(3) \prec \ldots$$

strictly orders Δ_{ev}, while the second order

$$Ev(1) \prec Ev(0) \prec Ev(2) \prec Ev(3) \prec \ldots$$

does not. E.g., $\emptyset$ and $\{Ev(0)\}$ are identical on $\{B \mid B \prec Ev(1)\} = \emptyset$, but $\emptyset \vdash_\Delta Ev(1)$ while $\{Ev(0)\} \not\vdash_\Delta Ev(1)$.

Natural inductions that respect an order $\prec$ that strictly orders Δ in $\mathcal{O}$ satisfy all the good properties (1-4) above, as shown by the following two propositions.

Proposition 2. *If $\prec$ strictly orders Δ then any natural induction $\mathcal{N}$ that respects $\prec$ also follows $\prec$.*

Proposition 3. *If $(\Delta, \prec)$ is an ordered definition in context $\mathcal{O}$ and $\prec$ strictly orders Δ in $\mathcal{O}$, then all terminal natural inductions that respect $\prec$ converge. Moreover the limit is independent of $\prec$. (It is the ultimate well-founded fixpoint of Δ in $\mathcal{O}$ - see Definition 19).*

Also this proposition follows from the stronger Theorem 3.

This theorem shows that an ordered definition in which $\prec$ strictly orders Δ unambiguously defines a set. The definition of an ordered definition can now be refined as follows.

Definition 13. *A definition Δ is a definition by well-founded induction over $\prec$ in $\mathcal{O}$ (or briefly, an ordered definition) if $\prec$ strictly orders Δ in $\mathcal{O}$. The structure defined by it is the limit of any terminal natural induction that respects $\prec$.*

Interestingly, the convergence property states that the limit is independent of the selected order. Sometimes this phenomenon can be seen in mathematical text.

Example 11. *We defined the satisfaction relation $\models$ over the subformula order but it is not uncommon to define it over alternative induction orders. For example, we could define $\models$ by induction on the* size *of formulas. Formally, we define $Sat(I, \psi) \prec Sat(J, \phi))$ if $I = J$ and the size of ψ (the number of nodes in its parse tree) is strictly less than the size of ϕ. Alternatively, we may define $\models$ by induction on the* depth *of formulas, i.e., the length of the longest branch in the parse tree of ϕ. The three orders lead to three variants of Definition 2. Intuition suggests that they are equivalent.*

It is indeed easy to verify that each of them strictly orders $\Delta_\models$ in $\mathcal{O}_S$: for each rule instance, the body refers to formulas that are strict subformula's, and have smaller size and depth than the formula in the head. Hence, it follows from the proposition that these definitions are indeed equivalent.

This does not mean that they have the same natural inductions. E.g., reconsider the natural induction of Example 7:

$$\to Sat(\{P\}, P) \to Sat(\{P\}, \neg\neg P) \to Sat(\{P\}, P \vee P)$$

This one respects and follows the subformula order and the size order. However, it does not respect the depth order, since $\mathfrak{A}_1$ is not saturated on $\{B \mid B \prec Sat(\{I\}, \neg\neg P)\}$. For instance, $Sat(\{P\}, P \vee P)$ is derivable but not derived and $P \vee P$ has strictly smaller depth than $\neg\neg P$.

Generalizing monotone and ordered definitions. There is an obvious similarity between Propositions 1 and 3. However, the former is not a generalization of the latter, because not all monotone definitions are ordered. For instance, there is no $\prec$ that strictly orders definition Δ_{TC} of transitive closure. Due to the transitivity rule, all defined domain atoms depend on each other; the only dependency relation is the total one and this is not a strict order. In this section, we define the more general class of *iterated inductive definitions*, which encompasses all ordered definitions as well as all monotone definitions. We will then prove a theorem for this more general class, which generalizes both of the earlier results.

The general idea of iterated inductive definitions is that they admit a dependency $\propto$ that is not a strict order. However, if atom A, B depend on each other (that is, $A \propto B \propto A$), then they depend *monotonically* on each other: deriving B may make A derivable and vice versa.

We define $A \prec_\propto B$ if $A \propto B$ and $B \not\propto A$. Since $\propto$ is transitive, $\prec_\propto$ is a strict order. $\prec_\propto$ divides the set of domain atoms into a set of strictly ordered "layers" such that, for all A, B, if $A \prec_\propto B$, then A is in a strictly lower layer than B, and if $A \propto B \propto A$, they are in the same layer.

Natural inductions of an iterated inductive definition proceed along the order $\prec_\propto$. Just as for an ordered definition, an atom A may be derived at step i only if $\mathfrak{A}_i$ is saturated on $\{B \mid B \prec_\propto A\}$. In this way, a natural induction closes layer by layer, and starts a new monotone induction in the next layer as soon as a layer is saturated. The following definition will serve to ensure that the "sub-inductions" that take place inside a single layer are monotone.

Definition 14. *A relation $\propto$ monotonically orders Δ in $\mathcal{O}$ if $\prec_\propto$ is a strict well-founded order and for all defined A, for all $\mathfrak{A}, \mathfrak{B}$ such that $\mathfrak{A}|_{\prec_\propto A} = \mathfrak{B}|_{\prec_\propto A}$ and $\mathfrak{A}|_{\propto A} \subseteq \mathfrak{B}|_{\propto A}$, if $\mathfrak{A} \vdash_\Delta A$ then $\mathfrak{B} \vdash_\Delta A$.*

Proposition 4. *If $\propto$ monotonically orders Δ in $\mathcal{O}$ then $\propto$ is a dependency relation of Δ in $\mathcal{O}$.*

Definition 15. *We say that a natural induction $\mathcal{N}$ respects (follows) a dependency relation $\propto$ if it respects (follows) $\prec_\propto$ according to Definition 10.*

Proposition 5. *If a natural induction $\mathcal{N}$ respects a relation $\propto$ that monotonically orders Δ in $\mathcal{O}$, then $\mathcal{N}$ follows $\propto$.*

Proposition 6. *Assume that $\propto$ monotonically orders Δ in $\mathcal{O}$. Then all terminal natural inductions that respect $\propto$ converge. Moreover, the limit is independent of $\propto$. (It is the ultimate well-founded fixpoint of Δ in $\mathcal{O}$.)*

Again, this proposition follows from Theorem 3.

We have already defined the concept of a monotone and ordered definition in context $\mathcal{O}$. Now, we also define the concept of an iterated inductive definition (in $\mathcal{O}$).

Definition 16. *A definition Δ is a definition by iterated induction over $\propto$ in $\mathcal{O}$ if $\propto$ monotonically orders Δ in $\mathcal{O}$. Its defined structure is the limit of any terminal natural induction.*

Now we can show that iterated inductive definitions generalize monotone and ordered definition. For a monotone definition, the entire set of all domain atoms can serve as a single layer. Let $\propto_t$ denote the total binary relation on $At_{\Delta}^{def(\Delta)}$. Note that $<_{\propto_t}= \emptyset$.

Proposition 7. *A definition Δ is monotone in $\mathcal{O}$ iff Δ is a definition by iterated induction over $\propto_t$ in $\mathcal{O}$. A natural induction of Δ in $\mathcal{O}$ (trivially) respects $\propto_t$.*

Proposition 8. *For a binary relation $\propto$, a definition Δ is a definition by well-founded induction over $\propto$ in $\mathcal{O}$ iff Δ is by iterated induction over $\propto$ in $\mathcal{O}$ and $\propto$ is irreflexive and asymmetric (hence, a strict order).*

As a consequence, Prop. 6 is a generalization of both Prop. 1 and Prop. 3.

Informal iterated inductive definitions are quite common in mathematical text although they are only very rarely formulated as sets of informal rules. To phrase them, formal scientists use other tools from their toolbox, for example fixpoints of operators. A well-known iterated inductive definition is the alternating fixpoint definition of the well-founded model in (Van Gelder 1993). In this theory, a (stable) operator $\mathcal{A}$ is defined on structures by defining $\mathcal{A}(\mathfrak{A})$ as the least fixpoint of a monotone operator $\lambda x T(x, \mathfrak{A})$. This stable operator is anti-monotone. The well-founded fixpoint is then characterised as the limit of an alternating fixpoint construction using $\mathcal{A}$. This is iterated induction in the sense that each of the steps involves itself a monotone inductive construction.

A rare case where iterated induction is explicitly available in rule form is in the definition of *stable theory* (Marek 1989) which is the set of modal propositional formulas closed under the standard inference rules and two additional ones:

$$\frac{\vdash \psi}{\vdash K\psi} \qquad \frac{\not\vdash \psi}{\vdash \neg K\psi}$$

The second is a non-monotone rule. The set is computed by iterated induction for increasing modal nesting depth of modal formulas.

In knowledge representation, there are many applications of iterated inductive definitions that have a natural representation as rule sets. This is for instance the case in representations of dynamic systems with ramifications (Denecker and Ternovska 2007).

Summary of informal definitions. The above theory exposes several important issues of informal definitions. First, that the "non-constructive" characterization is incorrect in case of non-monotone (ordered) definitions. Second, that the induction process is highly non-deterministic, and therefore that convergence is all-important. In mathematical practice, we typically take this property for granted. In fact, it is not trivial at all. It is a fundamentally important property of inductive definitions.

Third, in mathematical texts, we have a certain degree of freedom when it comes to choosing the induction order for an inductive definition. Nevertheless, the order is far from arbitrary and needs to match the structure of the rules. Our exposition clarifies the role and nature of the induction order, and how it constrains the order of rule application.

Finally, as shown by Proposition 6, the choice of the induction order is irrelevant as long as it matches the rules. The order does not affect the semantics of the definition.

In view of this, one may wonder why an induction order is specified at all in mathematical text? One possible explanation is that it serves to help the reader better understand the definition. Moreover, the specified order may help him/her as a kind of *parity check* of the soundness of the definition. Indeed, not all sets of informal rules form sensible definitions (far from). The induction order helps in verifying that the (informal) rules indeed form a sensible definition (see also the "parity check" discussion following Corollary 1).

Developing a logic of definitions. In the mathematical logic of ordered and iterated induction (IID) presented in (Buchholz et al. 1981), an iterated inductive definition is expressed via SO formula's that express a definition Δ and, independently, an induction order $\prec$. They use this logic system to study proof-theoretic strength and expressivity of iterated definitions. From a representational point of view however, we see two problems with an approach in which the induction order is explicitly expressed. First, expressing an induction order in logic might be as complex as expressing the definition itself, if not more. It is a needless complication of the knowledge representation process. Second, it also makes the knowledge representation process more error-prone. Even though the logic of (Buchholz et al. 1981) imposes strong additional constraints on the induction process so that convergence can be guaranteed, it is still possible to express combinations of a definition and an induction order that will converge to the wrong limit. For example, one can encode the definition Δ_{ev} with the non-matching order $Ev(1) \prec Ev(0) \prec Ev(2) \prec \ldots$, in which case the unintended set $\{Ev(1), Ev(0), Ev(3), Ev(5), \ldots\}$ will be constructed.

It is preferable to design a logic of definitions in which only the rules need to be represented and the order is left implicit. Indeed, Proposition 6 gives us license to do this, because it shows that all induction orders that fit the structure of the rules of Δ produce the same unique limit of their terminal natural inductions.

Designing a logic in which all and only iterated inductive definitions can be expressed is impossible. A logic's syntax should be decidable, while it is easy to see that it is undecidable whether a rule set is a monotone, ordered or iterated definition in a structure $\mathcal{O}$. One option out is by imposing syntactical constraints. For monotone definitions, we could impose the simple syntactic criterion that only rules with positive bodies are allowed. Beyond monotone definitions, however, this seems unfeasible. One issue is that whether a rule sets admits a $\propto$ that strictly or monotonically orders Δ in $\mathcal{O}$ may depend on the context $\mathcal{O}$.

Example 12. *The definition Δ_{ev} is an ordered inductive definition over the standard order in the natural numbers, but does not admit an induction order in the context $\mathcal{O}$ with domain $\{0,1\}$, $0^{\mathcal{O}} = 0, 1^{\mathcal{O}} = 1, +^{\mathcal{O}} = \{(0,0,0),(0,1,1),(1,0,1),(1,1,1)\}$. Indeed, in this structure, the two instances of the inductive rule are: $even(1) \leftarrow \neg even(0)$ and $even(1) \leftarrow \neg even(1)$. The least dependency relation is $even(0) \propto even(1) \propto even(1)$, but $\propto$ does not strictly or monotonically order Δ_{ev} in $\mathcal{O}$. Similarly, definition $\Delta_{\models}$ is not a legal definition in contexts where $Not^{\mathcal{O}}$ contains cycles.*

Requiring syntactic stratification is a possibility but it would eliminate interesting cases such as definition $\Delta_{\models}$. A more refined condition like local stratification (Przymusinski 1988) offers more liberty but it applies only to conjunctive rule bodies, in Herbrand contexts, is also undecidable and moreover is very brittle. Simple equivalence preserving transformations may transform a locally stratified rule set into one that is not. For instance, transforming rules $\forall \bar{x}(P(\bar{t}) \leftarrow \varphi)$ to $\forall \bar{y}(P(\bar{y}) \leftarrow \exists \bar{x}(\bar{t} = \bar{y} \land \varphi))$ breaks local stratification. This is not a good idea.

The alternative option is to waive any syntactic restriction but to design a "partial" semantics that assigns the "right" values to the defined predicates if Δ is a sensible inductive definition in $\mathcal{O}$ and otherwise does not assign a value. It is here that the well-founded semantics comes in. It will help us to cope with the absence of an explicit induction order in two ways. First, if this semantics assigns a value, then we can be certain that the definition is correct and that the assigned value is indeed the intended one. Second, as we will see, this semantics also provides mathematical tools that we can use to understand the workings of an inductive definition, without making reference to a specific induction order.

Well-founded inductions

The goal of this section is to show that the well-founded semantics allows us to perform the natural induction along some induction order, without actually knowing this order up front. There exist many formalizations of the well-founded semantics, but they all have in common that they construct a sequence of three-valued structures of increasing precision. This is quite different from a natural induction, which is an increasing sequence of two-valued structures. The following observation gives a clue as to how such sequences might be related.

Consider a (stepwise) natural induction of Δ in context $\mathcal{O}$ that respects $\prec_\propto$.

$$\to A_1 \to A_2 \to \ldots \to A_\beta$$

At each i, the structure $\mathfrak{A}_i = \{A_1, \ldots, A_i\}$ already provides the following partial information about the limit $\mathfrak{A}_\beta$:

- $A \in \mathfrak{A}_\beta$ if $A \in \mathfrak{A}_i$;
- $A \notin \mathfrak{A}_\beta$ if $\mathfrak{A}_i \not\vdash_\Delta A$ and $\mathfrak{A}_i$ is saturated in $\{B | B \propto A\}$. Indeed, since $\mathfrak{A}_i$ is saturated on $\{B | B \propto A\}$, it is saturated on every $\{B \mid B \propto C\} \subseteq \{B \mid B \propto A\}$. As a consequence, no atoms $C \propto A$ will ever be derived in the future. Since A is not derivable now, it will never become derivable.
- It is unknown whether $A \in \mathfrak{A}_\beta$ otherwise.

This tells us how to construct from each $\mathfrak{A}_i$ a three-valued structure $\mathcal{I}_i$ that "approximates" the defined set $\mathfrak{A}_\beta$. As such, a natural induction implicitly specifies a sequence of three-valued structures of increasing precision, that converges to $\mathfrak{A}_\beta$. Let us explore this in a few of the examples.

Example 13. *The transitive closure definition Δ_{TC} in context $\mathcal{O}$ of Example 3 is an iterated inductive definition over the total dependency relation $\propto_t$ ($\prec_{\propto_t} = \emptyset$). This induces a unique layer consisting of all domain atoms. Consider the following terminal natural induction:*

$$\to R(a,a) \to R(b,c) \to R(c,b) \to R(b,b) \to R(c,c)$$

The intermediate structures $\mathfrak{A}_i$ ($i < 5$) in this sequence are not saturated and no negative information is available. Hence, for each $i < 5$, all atoms not in $\mathfrak{A}_i$ are unknown in $\mathcal{I}_i$. However, the limit $\mathfrak{A}_5 = R$ is saturated, and only then it can be derived that all atoms not in $\mathfrak{A}_5$ are false.

Example 14. *The definition Δ_{ev} in context $\mathcal{O}$ of Example 2 is by induction over the standard order. The unique natural induction is:*

$$\to Ev(0) \to Ev(2) \to \ldots \to Ev(2n) \to \ldots$$

At each step i, one can verify that $\mathfrak{A}_i$ is saturated on $\{Ev(j) \mid j \leq 2i-1\}$. It follows that, for $0 \leq j \leq i$, $\mathcal{I}_i \models \neg Ev(2j-1)$. The following sequence describes the derivation of positive and negative information during this natural induction:

$$\to \{\} \to Ev(0) \to \neg Ev(1) \to Ev(2) \to \neg Ev(3) \to \ldots$$

Thus, natural inductions correspond to sequences of three-valued structures $\langle \mathcal{I}_i \rangle_{0 \leq i \leq \beta}$. As we have seen, however, building a correct natural induction requires an induction order $\prec_\propto$. We will now show that we can exploit the additional information that is present in three-valued structures to construct $\mathcal{I}_{i+1}$ from $\mathcal{I}_i$, without using the induction order at all.

In the discussion below, we assume without loss of generality that a (finite) Δ contains exactly one rule $\forall \bar{x}(P(\bar{x}) \leftarrow \varphi)$ per defined predicate P (we can bundle finitely many multiple rules in one using a disjunction). We denote the body φ of this unique rule as φ_P. Below, when $A = P(\bar{a})$, we write $\varphi_A^{\mathfrak{A}}$ as a shorthand for $\varphi_P^{\mathfrak{A}[\bar{x}:\bar{a}]}$.

We first recall the formalization of the parametrized well-founded semantics of (Denecker and Vennekens 2007). We consider three-valued structures on $def(\Delta)$ on domain D. For each $P/n \in def(\Delta)$, $P^{\mathcal{I}}$ is a function $D^n \to \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$. The precision order on the three-valued structures is the point-wise extension of the partial order $\mathbf{u} \leq_{\mathrm{p}} \mathbf{t}$, $\mathbf{u} \leq_{\mathrm{p}} \mathbf{f}$. If U is a set of domain atoms, then $\mathcal{I}[U : \mathbf{f}]$ is identical to $\mathcal{I}$ except that every atom $A \in U$ is false; likewise for $\mathcal{I}[U : \mathbf{t}]$.

Rules are evaluated using a three-valued truth function $\varphi^{\mathcal{I}}$, defined for formulas φ and three-valued structures $\mathcal{I}$ that interpret all free symbols of φ. At this moment, this truth function is generic. All we require here is that it satisfies two properties: (1) if $\mathcal{I}_1 \leq_{\mathrm{p}} \mathcal{I}_2$ then $\varphi^{\mathcal{I}_1} \leq_{\mathrm{p}} \varphi^{\mathcal{I}_2}$ ($\leq_{\mathrm{p}}$-monotonicity) and (2) if $\mathfrak{A}$ is a (two-valued) structure, $\varphi^{\mathfrak{A}}$ is the standard truth value of φ in $\mathfrak{A}$.

Definition 17. *Given is a definition Δ and context $\mathcal{O}$. We say that $\mathcal{I}'$ is a Δ-refinement of $\mathcal{I}$ (in $\mathcal{O}$) if there is a set U of unknown domain atoms in $\mathcal{I}$ such that one of the following conditions is satisfied:*

- *for each $A \in U$, ${\varphi_A}^{\mathcal{I}} = \mathbf{t}$ and $\mathcal{I}' = \mathcal{I}[U : \mathbf{t}]$; or*
- *for each $A \in U$, ${\varphi_A}^{\mathcal{I}'} = \mathbf{f}$ and $\mathcal{I}' = \mathcal{I}[U : \mathbf{f}]$.*

As before, we used $\varphi^{\mathcal{I}}$ as a shorthand for $\varphi^{\mathcal{O} \circ \mathcal{I}}$.

Note the asymmetry: to derive A, its body must be true in $\mathcal{I}$ itself; but to derive $\neg A$, it suffices that its body is false in $\mathcal{I}[U : \mathbf{f}]$, i.e., we are free to assume that atoms are $\mathbf{f}$, as long as this prophecy turns out to fulfill itself. One recognizes the familiar concept of unfounded set in the second sort of refinement (Van Gelder, Ross, and Schlipf 1991).

Definition 18. *A* well-founded induction *of Δ in (context) $\mathcal{O}$ is a sequence $\langle \mathcal{I}_i \rangle_{0 \leq i \leq \beta}$ of three-valued $def(\Delta)$-structures with domain D such that $\mathcal{I}_0$ is $\mathbf{u}$ (the mapping of all defined domain atoms to $\mathbf{u}$), for each ordinal i, $\mathcal{I}_{i+1}$ is a Δ-refinement of $\mathcal{I}_i$ in $\mathcal{O}$, and for limit ordinals λ, $\mathcal{I}_\lambda$ is the $\leq_{\mathrm{p}}$-limit of $\langle \mathcal{I}_i \rangle_{i<\lambda}$. We call a well-founded induction terminal if its limit $\mathcal{I}_\beta$ has no strictly more precise Δ-refinement.*

A well-founded induction increases in precision, and hence, it has a $\leq_{\mathrm{p}}$-limit.

Theorem 1 ((Denecker and Vennekens 2007)). *All well-founded inductions in $\mathcal{O}$ can be extended to a terminal well-founded induction in $\mathcal{O}$. All terminal well-founded inductions in $\mathcal{O}$ converge.*

Definition 19. *The* well-founded model of Δ relative to $\mathcal{O}$ *(and the selected truth function) is the limit of any terminal well-founded induction of Δ in $\mathcal{O}$.*

We explain our intuition underlying well-founded inductions. It not only constructs elements of the defined set but also non-elements. Given such a partial set $\mathcal{I}_i$ at intermediate stage i, a new element A can be derived if what is known of the defined set, suffices to ascertain that A is derivable, that is ${\varphi_A}^{\mathcal{I}_i} = \mathbf{t}$. This explains the first refinement $\mathcal{I}_i[U : \mathbf{t}]$. A non-element A may be derived likewise, if it is certain that its condition does not hold, that is if ${\varphi_A}^{\mathcal{I}_i} = \mathbf{f}$. However, this is not enough. For example, in a positive definition such as Δ_{TC}, if we wait till ${\varphi_A}^{\mathcal{I}_i} = \mathbf{f}$ to derive that A is a non-element, we are waiting forever. A principle is needed that "looks ahead"; a principle that finds out that there are no more options to derive one or more domain atoms through natural induction from the current state. If U is such that $\mathcal{I}_i[U : \mathbf{f}]$ falsifies the premise φ_A of every $A \in U$, then no ways are left to construct such A's. This explains the second refinement $\mathcal{I}_i[U : \mathbf{f}]$.

The well-founded model is usually defined for normal logic programs, which correspond to the fragment of our rule based formalism that satisfies the following conditions: Δ is a set of normal rules (i.e., conjunctions of literals in the body), $pars(\Delta)$ contains only constant and function symbols (i.e., all predicates are defined), and $\mathcal{O}$ is the unique Herbrand interpretation of $pars(\Delta)$. If, in addition, we choose the standard Kleene three-valued truth function as our function $\varphi^{\mathcal{I}}$, then we obtain precisely the standard well-founded model of Δ.

An alternative three-valued truth function is the *supervaluation*, defined by $\varphi^{\mathcal{I}} = glb_{\leq_{\mathrm{p}}}\{\varphi^{\mathfrak{A}} \mid \mathcal{I} \leq_{\mathrm{p}} \mathfrak{A}$ and $\mathfrak{A}$ is two-valued $\}$ (van Fraassen 1966). It is more precise than Kleene's: e.g, for $P^{\mathcal{I}} = \mathbf{u}$ the supervaluation of $P \vee \neg P$ is $\mathbf{t}$ while Kleene's valuation is $\mathbf{u}$. We can also construct the well-founded model w.r.t. this truth function. This is the *ultimate* well-founded model of Δ relative to $\mathcal{O}$ (Denecker, Marek, and Truszczyński 2004). The standard well-founded model is less or equally precise than the ultimate one. This implies that if it is two-valued (i.e., maximally precise), it is also the ultimate well-founded model.

We are now ready to link well-founded and natural inductions. Let Δ be a definition by iterated induction over $\propto$ in context $\mathcal{O}$ with domain D. Let $\mathcal{N}$ be a natural induction that respects $\propto$:

$$\mathfrak{A}_0 \to \mathfrak{A}_1 \to \ldots \to \mathfrak{A}_\beta$$

With $\mathcal{N}$, we now associate the following sequence $WFI(\mathcal{N})$ of three-valued structures:

$$\mathcal{I}_0 \to \mathcal{I}_{0,1} \to \mathcal{I}_1 \to \mathcal{I}_{1,2} \to \mathcal{I}_2 \to \ldots \mathcal{I}_\beta \to \mathcal{I}_{\beta,\beta+1}$$

which we define by induction:

- $\mathcal{I}_0 = \mathbf{u}$ (for all domain atoms A, $A^{\mathcal{I}_0} = \mathbf{u}$).
- $\mathcal{I}_{i,i+1} = \mathcal{I}_i[U : \mathbf{f}]$ where $U = \{A \mid A^{\mathcal{I}_i} = \mathbf{u}, \mathfrak{A}_i \not\vdash_\Delta A$ and $\mathfrak{A}_i$ is saturated on $\{B \mid B \propto A\}\}$.
- $\mathcal{I}_{i+1} = \mathcal{I}_{i,i+1}[U : \mathbf{t}]$ where $U = \mathfrak{A}_{i+1} \setminus \mathfrak{A}_i$.
- $\mathcal{I}_\lambda$ is the $\leq_{\mathrm{p}}$-limit of its predecessors, for limit ordinals λ.

For the following theorem, we select supervaluation as three-valued truth function.

Theorem 2. *Let Δ be a definition by iterated induction over $\propto$ in $\mathcal{O}$. If $\mathcal{N}$ is a terminal natural induction of Δ in $\mathcal{O}$ that respects $\propto$ then the sequence $WFI(\mathcal{N})$ is a terminal well-founded induction and $\mathcal{I}_{\beta,\beta+1} = \mathfrak{A}_\beta$.*

It is on this theorem that the thesis of this paper rests. Indeed, it has the following theorem as a corollary, from which we directly obtain Proposition 6.

Theorem 3. *Let Δ be a definition by iterated induction over $\propto$ in $\mathcal{O}$. For any terminal natural induction $\mathcal{N}$ of Δ in $\mathcal{O}$ that respects $\propto$, its limit is the ultimate well-founded model of Δ in $\mathcal{O}$.*

Arguably, the construction process of a well-conceived formal or informal definition should result in a well-defined set. In our setting, this mean that the well-founded model in $\mathcal{O}$ should be two-valued. In this case we call Δ *total* in $\mathcal{O}$.

Corollary 1. *If Δ is a definition by iterated induction over $\propto$ in $\mathcal{O}$, then Δ is total in $\mathcal{O}$.*

Thus an iterated definition Δ has the desirable property of being total in $\mathcal{O}$. While the induction order has no semantical role in a definition, it gives insight in the structure of the definition and it gives a "parity check" of the correctness/totality of the definition.

From a technical point of view, Corollary 1 shows that one way to prove that a rule set has a 2-valued well-founded model is to prove that it has a dependency relation that monotonically orders it. This generalizes a condition in (Denecker and Ternovska 2008) and, to the best of our knowledge, it is the most general such condition currently known.

A comparison of natural and well-founded inductions. Below we write a well-founded induction as:

$$\rightarrow S_1 \rightarrow S_2 \rightarrow \ldots$$

where S_i is either the set of atoms U such that $\mathfrak{A}_i = \mathfrak{A}_{i-1}[U : \mathbf{t}]$ or the set of negations $\neg A$ of atoms A in the set U such that $\mathfrak{A}_i = \mathfrak{A}_{i-1}[U : \mathbf{f}]$.

Example 15. *For the natural induction $\mathcal{N}$ of Example 3:*

$$\rightarrow R(a,a) \rightarrow R(b,c) \rightarrow R(c,b) \rightarrow R(b,b) \rightarrow R(c,c)$$

the corresponding well-founded induction $WFI(\mathcal{N})$ is:

$$\rightarrow \emptyset \rightarrow R(a,a) \rightarrow \emptyset \rightarrow R(b,c) \rightarrow \emptyset \rightarrow R(c,b) \rightarrow$$
$$\emptyset \rightarrow R(b,b) \rightarrow \emptyset \rightarrow R(c,c) \rightarrow \{\neg R(x,y) \mid (x,y) \notin R\}.$$

Example 16. *The natural induction of Example 2:*

$$\rightarrow Ev(0) \rightarrow Ev(2) \rightarrow \ldots$$

has the corresponding well-founded induction $WFI(\mathcal{N})$:

$$\rightarrow \emptyset \rightarrow Ev(0) \rightarrow \neg Ev(1) \rightarrow Ev(2) \rightarrow \neg Ev(3) \rightarrow \ldots$$

Many definitions admit well-founded inductions that are not natural inductions. Well-founded inductions can go dramatically "faster" than natural inductions.

Example 17. *A well-founded induction of the satisfaction definition $\Delta_{\models}$ that does not respect the induction order, but is nevertheless a correct well-founded induction is:*

$$\rightarrow Sat(\{P\},P) \rightarrow \{Sat(\{P\},P\lor\varphi) \mid \varphi \textit{ a formula over } S\}$$

Indeed, after the first step $Sat(\{P\},P)$ is true in the three-valued structure $\mathcal{I}_1$. Hence, we have $[Sat(\{P\},P) \lor Sat(\{P\},\varphi)]^{\mathcal{I}_1} = \mathbf{t}$ for every φ. It follows that $\mathfrak{A}_1[U : \mathbf{t}]$ is a refinement of $\mathfrak{A}_1$ for $U = \{Sat(\{P\}, P \lor \varphi) | \varphi$ a formula over $S\}$.

The well-founded induction in the example matches how humans derive $\mathfrak{A} \models \varphi \lor \psi$: if a disjunct φ is derived to be satisfied, we jump to the conclusion that $\varphi \lor \psi$ is satisfied, even if the value of ψ is unknown. Strictly speaking, here we are violating the induction order! It is nevertheless safe: any fact derived during a well-founded induction is correct.

The next example shows that there are sensible definitions that are not iterated definitions.

Example 18. *Let $\mathcal{O}$ be the natural number context of Example 2 and Δ_{ev1} the following variant definition of even numbers and an additional predicate $Next$:*

$$\left\{\begin{array}{l} \forall x\forall y(Next(x,y) \leftarrow x = y+1) \\ \forall x(Even(x) \ \leftarrow x = 0\lor \\ \qquad\qquad\qquad \exists y(Next(x,y) \land \neg Even(y))) \end{array}\right\}$$

The definition defines and uses $Next$ as the successor predicate. Its well-founded model $\mathfrak{A}$ interprets $Next$ as the successor relation and $Even$ as the set of even numbers, as expected. However, this definition is not an iterated induction. Indeed, in any dependency relation of Δ_{ev1}, it holds that $Ev(n) \propto Ev(m)$ for all $n,m \in \mathbb{N}$. This follows from the fact that $Ev(m)$ is derivable in the structure $\{Next(m,n)\}$ but not in $\{Next(m,n), Ev(n)\}$. However, as the same two structures show, no such $\propto$ monotonically orders Δ_{ev1} in $\mathcal{O}$.

The above example shows a disturbing brittleness of the concept of a definition by iterated induction as defined in Definition 16, that fortunately is not shared with the rule formalism under the well-founded semantics.

To summarize, the concept of well-founded induction seems to provide a superior formalization of the induction process, one that is independent of the order, faster, more like humans reason on informal definitions and more robust under innocent syntactical variance.

Standard versus ultimate well-founded semantics. The result of the previous section (the convergence of natural and well-founded inductions) are formulated for the ultimate well-founded semantics, the one derived using the supervaluation. Here, we investigate to what extend the result holds for the standard well-founded semantics.

The first observation is that if the standard well-founded model is two-valued, it is the ultimate well-founded model. It follows that if the standard well-founded model of an iterated inductive definition is two-valued, it is the structure defined by this definition. In general however, an iterated inductive definition as defined here, may not have a two-valued standard well-founded model. An example is $\Delta = \{P \leftarrow P \lor \neg P\}$: it is monotone according to Definition 8 and its defined structure and ultimate well-founded model is $\{P\}$. On the other hand, P is unknown in Δ's standard well-founded model. Thus, the standard well-founded semantics is too weak for this definition.

How serious is the gap? To address this, we make two observations. First, standard and ultimate well-founded semantics break apart only in quite specific circumstances: when there is a *case splitting* involving one or multiple rule bodies leading to a loop over negation (see above example). This is a pattern that we *never* observe in our applications. In practice, standard and ultimate semantics coincide.

The second observation might explain the first one. Although our formal definitions of monotone, ordered and iterated inductive definitions admit rule sets showing such a pattern, the conventions of informal definitions seem to forbid it. Consider, for instance, a variant of the informal definition of satisfaction (Definition 2), that is obtained by replacing its first rule by the following one:

- $I \models P$ if either $I \models P$ and $P \in I$, or if $I \not\models P$ and $P \in I$;

Or after splitting this rule:

- $I \models P$ if $I \models P$ and $P \in I$;
- $I \models P$ if $I \not\models P$ and $P \in I$;

One could argue that these new rules "obviously" are equivalent to the original one by appealing to the fact that "$I \models P$

or $I \not\models P$" is tautologically true. But the modified rules mismatch the induction order and we think such a definition would not be accepted in mathematical text (we would not). This suggests that an implicit convention of informal definitions is not yet explicit in our theory: our definitions are too liberal. We think that some conventions regarding informal definitions remain to be discovered, and that their formalization will preclude rule sets of the above form and lead to a notion of iterated inductive definition that will have a two-valued standard well-founded model. In other words, we think that if one follows the conventions of informal inductive definitions while expressing them as a formal rule set, the standard well-founded model will be two-valued.

In summary, the computationally cheaper standard well-founded semantics, in practice, almost always coincides with the more expensive ultimate semantics. Moreover, in cases where they differ, it is far from obvious that the rule sets correspond to legal inductive definitions, and that the answer provided by the ultimate semantics is better than that of the standard well-founded model.

Conclusion. The paper is relevant from two angles. Informal (inductive) definitions are an important form of human expert knowledge and have many applications in scientific texts as well as in knowledge representation. In this paper we studied the general knowledge representation problem of representing informal definitions. From a very different angle, this paper studies the old but unresolved problem of the *informal semantics* of logic programming.

The first part of the paper was concerned with formalizing familiar but implicit intuitions and conventions of informal definitions. This was done by a series of definitions "from first principles" (Definitions 3-16). The central concept is our formalization of the induction process, the "natural induction". The analysis exposed non-trivial properties of informal definitions: the all-importance of convergence of the induction process, the role of the induction order, its tight link with the rules and ultimately, its irrelevance. To the best of our knowledge, our analysis provides the most detailed semantical account of the studied forms of informal definitions to date. The second part of the paper showed that the well-founded inductions underlying the well-founded semantics provide a natural, efficient and robust generalization of natural inductions. Taken together, this provides, for the first time, a direct argument why rule formalisms under the well-founded semantics correctly formalize informal inductive definitions.

We finish this paper with a speculation. Given that our conscious understanding of inductive definitions is quite partial, where does our proficiency to reason with them come from? Indeed, it is not because someone consciously (and erroneously) believes that the satisfaction relation is the least set closed under the familiar rules, that he or she is not capable to correctly reason about it. Indeed, daily practice provides ample evidence to the contrary. Arguably, this shows a gap between our conscious and subconscious understanding of inductive definitions. However, in contrast to similar gaps, e.g., in the context of reasoning on statistical information (Kahneman 2011), it is here the (slow) conscious understanding that is erroneous, and the (fast) subconscious reasoning that is correct (people do derive that $\{P\} \not\models \neg P$)! The explanation that we see for this phenomenon is that the principle(s) of inductive definition is not a primitive of human cognition, but is just a manifestation of a deeper common sense principle that is hard wired in the cognitive machinery of the human mind. As we have argued in several papers in the past, we believe that this base cognitive principle is that of causal reasoning: the induction process as a causal process that creates the defined object.

References

Buchholz, W.; Feferman, S.; Pohlers, W.; and Sieg, W. 1981. *Iterated Inductive Definitions and Subsystems of Analysis : Recent Proof-Theoretical Studies*, volume 897 of *Lecture Notes in Mathematics*. Springer.

Denecker, M., and Ternovska, E. 2007. Inductive situation calculus. *Artificial Intelligence* 171(5-6):332–360.

Denecker, M., and Ternovska, E. 2008. A logic of nonmonotone inductive definitions. *ACM Transactions on Computational Logic (TOCL)* 9(2):14:1–14:52.

Denecker, M., and Vennekens, J. 2007. Well-founded semantics and the algebraic theory of non-monotone inductive definitions. In Baral, C.; Brewka, G.; and Schlipf, J. S., eds., *LPNMR*, volume 4483 of *LNCS*, 84–96. Springer.

Denecker, M.; Bruynooghe, M.; and Marek, V. W. 2001. Logic programming revisited: Logic programs as inductive definitions. *ACM Transactions on Computational Logic (TOCL)* 2(4):623–654.

Denecker, M.; Marek, V. W.; and Truszczyński, M. 2004. Ultimate approximation and its application in nonmonotonic knowledge representation systems. *Information and Computation* 192(1):84–121.

Denecker, M. 1998. The well-founded semantics is the principle of inductive definition. In Dix, J.; del Cerro, L. F.; and Furbach, U., eds., *JELIA*, volume 1489 of *LNCS*, 1–16. Springer.

Kahneman, D. 2011. *Thinking, fast and slow*. Farrar, Strays and Giroux.

Marek, W. 1989. Stable theories in autoepistemic logic. *Fundamenta Informaticae* 12(2):243–254.

Martin-Löf, P. 1971. Hauptsatz for the intuitionistic theory of iterated inductive definitions. In Fenstad, J., ed., *Second Scandinavian Logic Symposium*, 179–216.

Przymusinski, T. C. 1988. On the declarative semantics of deductive databases and logic programs. In *Foundations of Deductive Databases and Logic Programming.* Morgan Kaufmann. 193–216.

van Fraassen, B. 1966. Singular terms, truth-value gaps and free logic. *Journal of Philosophy* 63(17):481–495.

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.

Van Gelder, A. 1993. The alternating fixpoint of logic programs with negation. *Journal of Computer and System Sciences* 47(1):185–221.