Self-Correcting Unsound Reasoning Agents (DARe 2017)

Yannick Chevalier

Self-Correcting Unsound Reasoning Agents

Yannick Chevalier[1]

Université Paul Sabatier*, IRIT, Toulouse, France yannick.chevalier@irit.fr

Abstract. This paper introduces a formal framework for relating learning and deduction in reasoning agents. Our goal is to capture imperfect reasoning as well as the progress, through *introspection*, towards a better reasoning ability. We capture the interleaving between these by a reasoning/deduction connection and we show how this—and related—definition apply to a setting in which agents are modeled by first-order logic theories. In this setting, we give a sufficient condition on the connection ensuring that under fairness assumptions the limit of introspection steps is a sound and complete deduction system. Under the same assumption we prove every falsehood is eventually refuted, hence the self-correction property.

Keywords: Reasoning Agents, Unsound Reasoning, First-Order Logic, Ordered Resolution, Saturation

1 Introduction

We consider in this paper a framework in which an agent is described by statements in a logic. These statements reflect the agent's view of the world, and can either be true—when the statement is known to be true in all possible cases—or just satisfiable—in which case we say the statement is merely believed. While the knowledge must be a consistent set of statements, we allow the set of beliefs to be inconsistent, as sometimes belief just reflects uncertainty, lack of knowledge, or simply queries related to the knowledge. Reasoning is accordingly split into two part:

- an *introspective* part, that aims at producing new known statements from existing ones;
- a *deductive* part, that aims at evaluating believed statements to eliminate those in contradiction with what is known to be true.

Specifically, we aim at capturing the dynamicity of reasoning, that is the interplay between these two aspects of reasoning. This agency framework is actually very close to what is routine in automated deduction. A given (usually first-order) theory T is saturated by adding consequences of already known clauses to obtain a new, hopefully finite, equivalent theory T' . The rationale for the development of a saturation procedure is to obtain a theory T' that can be used with a specific deduction strategy to answer queries.

* CIMI Project FMS

This work is built from the strong connection in first-order logic between a saturation procedure on the one hand, and the deduction strategy that can be proven complete for saturated sets of clauses on the other hand. For instance, linear ordered resolution is complete [2] for a set of clauses saturated up to redundancy under ordered resolution: given a saturated theory T and a ground clause C such that $T \cup \neg C$ is inconsistent, there exists a proof by ordered resolution of the empty clause in which each inference contains at most one clause from T. Similar results are obtained for superposition [2], equational theories [6, 5], with the possible addition of a selection function [8, 11]. In some works [1, 9], the output of the saturation process is directly a decision procedure for a set of queries (in the mentioned cases, a decision procedure for ground queries). In other cases it is proven that if a saturation procedure terminates then a given proof strategy is efficient, and in many cases this proof strategy can be turned into a decision procedures for ground query problems [2, 10, 3, 9].

In contrast with these more recent approaches, the dynamicity of reasoning that we consider is closer to the unfailing completion procedure [5] that applies to equational theories and does not have a subsumption rule. In [5] it is proved that a transformation of the theory can be formulated as a transformation of proofs that reduce their complexity for a well-founded ordering. Thus the absence of a subsumption rule, after which proofs may become more complex, is essential in that approach. Technically we aim at obtaining a similar main theorem, *i.e.* that the presentations computed step by step are eventually complete for refutation of any contradicted formula.

In the context of this paper, this means we are interested in the interleaving of saturation steps (called introspection above) and query evaluation steps (called deduction above). We introduce in this paper pairs $(\vdash^i, \vdash^d)$ of inference rules that reflect this situation. The relation $\vdash^i$ represents the inference rules employed during the saturation process, while the relation $\vdash^d$ represents the inference rules involved in the deduction.

Beside that technical justification for this paper, we note that we model a situation in which the lack of omniscience of agents is not the result of a partial knowledge, but rather stems from an incomplete form of deduction. This approach was proposed by Konolige in [7], but does not seem to have been pursued. In [7] what we call an agent state is called a deduction structure. While Konolige focuses on the properties of a deduction structure, for example by introducing a possible worlds semantics, this paper focuses on the analysis on the possible evolutions of deduction structures and on the general criteria on pairs $(\vdash^i, \vdash^d)$ needed to ensure that a contradictory formula will always be eventually (for relation $\vdash^i$) be refuted (using $\vdash^d$).

Outline. In this paper we first give basic definitions relating to first-order logic, and saturation in Sect. 2. We then present agent models and their evolution in Sect. 3, and we prove that under assumptions made on the introspection and deduction process, an agent may eventually remove all beliefs contradicting her knowledge. We give an example in Sect. 4 of how the processes of saturation and

deduction in first-order logic fit in that model. We conclude in Sect. 5 with the possible extensions of this paper.

2 Basic Definitions

2.1 First-Order Logic

First-order objects. We assume the reader is familiar with the basics of first-order logic, and just recall them here to define notations. Variables are denoted x, y, z and decorations thereof and constants denoted c etc. We let $\mathcal{X}$ and $\mathcal{C}$ be denumerable sets of variables and constants, respectively. Given a set $\mathcal{F}$ of function symbols with arity and a set S of nullary symbols, we denote $\mathcal{T}(\mathcal{F}, S)$ the set of terms over S, *i.e.* the least set containing S and closed by the application of functions in $\mathcal{F}$. The *set of terms* is $\mathcal{T}(\mathcal{F}, \mathcal{X} \cup \mathcal{C})$, and the *set of ground terms* is $\mathcal{T}(\mathcal{F}, \mathcal{C})$.

Given a set $\mathcal{P}$ of predicate symbols with arity, if $p \in \mathcal{P}$ is of arity n and $t_1, \ldots, t_n$ are terms then $p(t_1, \ldots, t_n)$ is an *atom*. A *literal* is either an atom A or its negation $\neg A$. We say the literal is positive in the first case and negative in the latter case. A *clause* is a multiset of literals denoting their disjunction. Accordingly the empty clause $\emptyset$ is always false. A theory is a set—or conjunction—of clauses. An atom, literal, clause, or theory is ground whenever all its atoms are constructed on ground terms. By skolemization every first-order logic theory can be presented by an equisatisfiable set of clauses. Finally we say a theory T is inconsistent, and write $T \models \emptyset$ if it has no model.

Unifier, resolution and factorization. A substitution σ is an idempotent function from variables to terms of finite support, *i.e.* $\sigma(x) = x$ for all but a finite number of variables. Substitutions are extended homomorphically (using $\sigma(f(t_1, \ldots, t_n) = f(\sigma(t_1), \ldots, \sigma(t_n))$ for function or predicate symbols and $\sigma(c) = c$ for constants) to terms, atoms, literals, and clauses. As usual, we prefer to use the infix notation, and write $t\sigma$ in lieu of $\sigma(t)$. A substitution σ is a *unifier* of two terms t, t' if $t\sigma = t'\sigma$. It is a most general unifier of t, t' if for every unifier τ of t, t' there exists a substitution θ such that $\sigma\theta = \tau$. As is well known the most general syntactic unifier of two terms is unique up to a renaming of variables. We extend this notion to atoms in the usual way. Resolution and factorization (see Fig. 2.1) were formally defined by Robinson [12]. A *deduction sequence* from T is a finite sequence $(C_1, \ldots, C_n)$ of clauses such that each C_i is either in T or the conclusion of a resolution or factorization inference with antecedents in $\{C_1, \ldots, C_{i-1}\}$. Whenever one such sequence exists we write $T \vdash C_n$. Resolution is complete for refutation, *i.e.* if $T \models \emptyset$ then $T \vdash \emptyset$. If Φ is a clause and T is not inconsistent, we have that $T \models \Phi$ is equivalent to $T, \neg\Phi \models \emptyset$, and thus by completeness for refutation $T, \neg\Phi \vdash \emptyset$. Trivially, if Φ is an atomic formula, *i.e.* contains only one literal, this implies $T \vdash \Phi$.

$$\frac{\Gamma \vee A \quad \neg B \vee \Delta}{(\Gamma \vee \Delta)\sigma}\ \sigma = \mathrm{mgu}(A, B) \qquad \frac{\Gamma \vee A \vee B}{(\Gamma \vee A)\sigma}\ \sigma = \mathrm{mgu}(A, B)$$

resolution factorisation

Fig. 1. Resolution and factorization inference rules

3 Agent Model

3.1 Modeling Reasoning

Agents are defined wrt to a logic $\mathcal{L}$ in which formulas are evaluated wrt to a set $\mathcal{M}$ of models. We let $\mathcal{S}_\mathcal{L}$ be the set of statements in $\mathcal{L}$, and $\mathcal{P}(\mathcal{S}_\mathcal{L})$ be the set of subsets of $\mathcal{S}_\mathcal{L}$, and Con be the set of consistent subsets of $\mathcal{L}$. First we define the introspection relation as a change of a representation of a theory T that yields a theory T' in which truth in a given model is preserved.

Definition 1. *(Introspection relation) An* introspection relation $\vdash^i$ *on the logic* $\mathcal{L}$ *is a truth-preserving relation on the powerset* $\mathcal{P}(\mathcal{S}_\mathcal{L})$ *, that is, for all* T, T' *, if* $T \vdash^i T'$ *, then for all* $m \in \mathcal{M}$ *we have* $m \models T$ *if and only if* $m \models T'$*.*

The $\vdash^i$ relation changes the world representation of an agent, presumably to better it. We now introduce a deduction relation that uses a representation of the world to reason about the truth of arbitrary statements in the language.

Definition 2. *(Deduction relation) A* deduction relation $\vdash^d$ *on the logic* $\mathcal{L}$ *is a relation* $\subseteq \mathcal{P}(\mathcal{S}_\mathcal{L}) \times \mathcal{L}$ *that represents a sound proof procedure for* $\mathcal{L}$*. Namely, if* $T \vdash^d \varphi$ *then for all* $m \in M$*, we have* $m \models T$ *implies* $m \models \varphi$*.*

In addition to being truth preserving, an introspection relation should, in order to be useful, reduce the difficulty of reasoning. We capture this intuition as follows. Given $T \in \mathrm{Con}$ and $\varphi \in \mathcal{L}$ such that $T \models \varphi$, we measure the difficulty of proving φ from T by the number of introspection steps needed to reach a set $T' \in \mathrm{Con}$ such that $T' \vdash^d \varphi$.

Definition 3. *(Distance to provability) Assume* $\vdash^i \subseteq \mathrm{Con} \times \mathrm{Con}$ *is an introspection relation,* $T \models \varphi$*, and that* $\vdash^d$ *is a deduction relation. Then we let:*

$$\mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T) = \min\{n \mid T \left(\vdash^i\right)^n T' \text{ and } T' \vdash^d \varphi\}$$

We aim to prove that introspection steps always make proving easier, or at least not more difficult. This implies a reasoning on all possible proofs, as is the case for unfailing completion (see above). The usual orderings on proofs (for example, a well-founded ordering on the clauses occurring in the proof [2, 3]) are not strictly decreasing, and the distance provides an admittedly coarser ordering sufficient for our purpose.

We capture the relation between introspection and deduction outlined previously by defining reasoning connections. These impose the constraints that can be summarized as follows:

– a *completeness* constraint (first point in the definition below) by imposing, when $T \models \varphi$ that at least one sequence of introspection steps will lead to a theory T' such that $T' \vdash^d \varphi$;
– a uniform *not-going-back* constraint imposing that whatever introspection step is taken from T, the distance of the resulting T' to any formula φ will be less than or equal to the distance between T and φ;
– a *progress* constraint stating that once a formula φ is chosen but not immediately provable, there exists an introspection step that will reduce the distance to provability.

These points are stated formally in the following definition.

Definition 4. *(Reasoning connection) A* reasoning connection *is a couple* $(\vdash^i, \vdash^d)$ *where* $\vdash^i$ *is an introspection relation,* $\vdash^d$ *is a deduction relation, and such that:*

1. completeness:

$$\forall \varphi \in \mathcal{L}, \forall T \in \mathrm{Con}, T \models \varphi \Rightarrow \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T) < +\infty$$

2. not-going-back:

$$\forall \varphi \in \mathcal{L}, \forall T, T' \in \mathrm{Con}, T \vdash^i T' \Rightarrow \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T') \leq \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T)$$

3. progress:

$$\forall T \in \mathrm{Con}, \forall \varphi \in \mathcal{L}, 0 < \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T) < +\infty \Rightarrow (\exists T', \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T') < \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T) \wedge T \vdash^i T')$$

The role of Def. 4 is to provide one with simple conditions that are sufficient to define reasoning agents in an arbitrary logic $\mathcal{L}$ who:

– start with imperfect reasoning yet are able to increase their reasoning power;
– and at the same time have a bounded deduction capacity as expressed by the (presumably incomplete in general) deduction relation.

3.2 Agent State and Trajectory

We now proceed to define an agent state. We consider a *naive* model of agent whose beliefs are statements *considered satisfiable*, *i.e.* that were not proven invalid. This choice of accepting all non-refuted statements transforms a sound but incomplete reasoning procedure into an unsound yet complete one, that will accept all valid entailments, but also may accept in the set of beliefs unsatisfiable statements.

Definition 5. *(Agent state) Let* $(\vdash^i, \vdash^d)$ *be a reasoning connection. A* $(\vdash^i, \vdash^d)$*-agent state is a couple* (T, B) *where* $T \in \mathrm{Con}$ *and* $B \subseteq \mathcal{L}$*, and for all* $\varphi \in B$ *we have* $T\neg \vdash^d \neg\varphi$*.*

In order to simplify notations in relation with this definition, we note $\mathrm{Ref}_{\vdash^d}(T)$ the set of statements φ that can be refuted from T given $\vdash^d$. That is,

$$\mathrm{Ref}_{\vdash^d}(T) = \{\varphi \mid T \vdash^d \neg\varphi\}$$

We present in this paper a first agent model that does not react to outside events nor with other agents. Accordingly, an agent evolves either by receiving a new aceptable belief or by changing her world's view through introspection.

Definition 6. *(Evolution of an agent) Let $A = (T, B)$ be a $(\vdash^i, \vdash^d)$-agent. An* evolution *of A is either:*

- *the reception of a new belief,*

$$(T, B) \xrightarrow[\varphi \notin \mathrm{Ref}_{\vdash^d}(T)]{?\varphi} (T, B \cup \{\varphi\})$$

- *an introspection step,*

$$(T, B) \xrightarrow[(\vdash^i T}{i}, B \setminus \mathrm{Ref}_{\vdash^d}(T'))$$

We note $(T, B) \rightsquigarrow (T', B')$ if there exists one evolution of (T, B) into (T', B').

We define *trajectories* as finite or infinite sequences of evolutions $\pi = (\pi_i = (T_i, B_i))_{0 \leq i}$. A trajectory is finite if one also has $i < n$ for some $n \in \mathbb{N}$ in the preceding definition. In that case it is said to be of length n, otherwise it is of infinite length. Finally, the *domain* of a trajectory π is the set of integers i for which π_i is defined, and is denoted $\mathrm{dom}(\pi)$.

Definition 7. *(Fair trajectories) A trajectory*

$$\pi = (T_0, B_0) \rightsquigarrow (T_1, B_1) \rightsquigarrow \ldots (T_n, B_n) \rightsquigarrow \ldots$$

is fair *if:*

$$\forall \varphi \in \mathcal{L}, \forall n \in \mathrm{dom}(\pi), +\infty > \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T_n) > 0 \Rightarrow \\ \exists m > n \in \mathrm{dom}(\pi), \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T)_n > \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi, T_m)$$

Given the unusual definition of fairness, we prove that fair trajectories always exist.

Lemma 1. *Let $(\vdash^i, \vdash^d)$ be a reasoning connection and (T, B) be a $(\vdash^i, \vdash^d)$-agent. Then there exists a fair trajectory starting from (T, B).*

Proof. Let $\varphi_0, \ldots, \varphi_n, \ldots$ be an enumeration of all statements in $\mathcal{L}$ such that $T \models \varphi_n$. By property 1. (completeness) we know that $\forall n \in \mathbb{N}, \mathrm{dist}_{\vdash^i, \vdash^d}(\varphi_n, T) < +\infty$. Let $\psi_0, \psi_1, \ldots$ be any sequence in which each φ_n appears $distdi\varphi_n T$ times. Let $T_0, \ldots, T_n, \ldots$ be constructed inductively as follows:

- $T_0 = T$

– if T_{i-1} constructed, $\text{dist}_{\vdash^i,\vdash^d}(\varphi_j, T_{i-1}) > 0$, and $\psi_i = \varphi_j$, then by property 3. (progress) choose T_i such that $\text{dist}_{\vdash^i,\vdash^d}(\varphi_j, T_i) < \text{dist}_{\vdash^i,\vdash^d}(\varphi_j, T_{i-1})$.

By property 3. (not-going-back) one easily proves that for all i the sequence $(\text{dist}_{\vdash^i,\vdash^d}(\varphi_i, T_j))_{j\in\mathbb{N}}$ is decreasing and thus if φ_i has been selected k times in the construction of T_j then $\text{dist}_{\vdash^i,\vdash^d}(\varphi_i, T_j) \leq \text{dist}_{\vdash^i,\vdash^d}(\varphi_i, T_0) - k$.

The interest of fair trajectories lies in the fact that they eventually prove any entailment.

Proposition 1. *Let $T \in \text{Con}$ and $\varphi \in \mathcal{L}$ be such that $T \models \varphi$, and let π be a fair trajectory starting with $(T_0 = T, B_0)$. Then there exists $N \geq 0$ such that for all $n \geq N$ we have $T_n \vdash^d \varphi$.*

Proof. Let $\pi = ((T_i, B_i)_{i\geq 0}$ and $u_n = \text{dist}_{\vdash^i,\vdash^d}(\varphi, T_n)$. The sequence of integers $u_0, u_1, \ldots$ contains only positive integers, and is decreasing by the not-going-back property. Furthermore, since the trajectory is fair, for each $n \in \text{dom}(\pi)$ such that $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T_n) > 0$ there exists $m > n$ such that $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T_n) > \text{dist}_{\vdash^i,\vdash^d}(\varphi, T_m)$. Thus there is a finite strictly decreasing subsequence of u_n which must converge to 0 at some indice N. By the not-going-back property, for each $n > N$ we have $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T_n) = 0$, *i.e.* $T_n \vdash^d \varphi$.

Note that since $\xrightarrow[\varphi\notin\text{Ref}_{\vdash^d}(T)]{?\varphi}$ steps have no impact on the first member of the couple, we have roundly ignored them. Prop. 1 can be applied to prove that any belief *varphi* $\in B$ inconsistent with T is eventually removed, and never reintroduced. The proof simply consists in applying Prop. 1 on $\neg\varphi$ for all $\varphi \in B$ inconsistent with T, and is thus omitted. It is however one of the main result of this paper, and is thus stated as a theorem.

Let (T, B) be a $(\vdash^i, \vdash^d)$-agent and let π be any fair trajectory starting from (T, B). Then there exists $N \geq 0$ such that for all $n \geq N$ and all $\varphi \in B$ such that $T \models \neg\varphi$, we have $\varphi \notin B_n$.

4 Example of Reasoning Connection

We continue in this section the example of subterm local deductions in first-order logic, and add to it ordered resolution. It is well known [2] that if a set of clauses is saturated for ordered resolution then linear order resolution is refutationally complete. The proof of this fact is based on the notion notion of redundancy employed in that article: a clause obtained by ordered resolution or factorisation is redundant if it is entailed by instances of clauses smaller than the premisses of the rule applied. Hence, intuitively, if a proof by ordered resolution and factorisation (denoted using $\vdash^o$ of $T, \neg\varphi \vdash^o \emptyset$ contains an inference between two clauses of T, this inference can be replaced by first, applying it on T to obtain T', and second, a case reasoning on the obtained clause: if it is redundant, then one can replace it in the original proof of $T, \neg\varphi \vdash^o \emptyset$ by the proof of redundancy,

and if it is not redundant, one can use the newly obtained clause to simplify the original proof.

Thus, in the original result, the completeness of linear ordered resolution for refutation is based on an ordering on proofs in which proofs are viewed as multisets of clauses. Thus, for every formula φ entailed by T, a saturation step can be translated into a decrease in the proof of φ. This approach is however not practical for our purpose, as we would to prove that the decrease happens for all formulas, and is strict for at least some. To prove this we need to consider the infinite number of minimal proofs, one for each entailed clause, and the order on this object is no longer well-founded. Hence our previous introduction of a distance with the *not-going-back* and *progress* properties. We also weaken slightly the redundancy criterion by considering the multiset of atoms resolved or factored upon in the redundancy proof, instead of the set of instances of clauses. If a proof is minimal in the sense of [2] then the multiset of atoms resolved or factored upon is minimal. However the converse does not hold: If $T = \{a \vee b, a \vee c, b, c\}$ and $a > b > c$ then we have $\{a \vee b, c\} > \{a \vee c, b\}$ though were these two sets usable interchangeably in a redundancy proof, they would yield the same multiset $\{a, b, c\}$ of atoms resolved upon.

In Sect. 4.1 we make this discussion more precise by introducing the orderings on terms, atoms, etc. In Sect. 4.2 we introduce the ordered resolution, factorisation, and redundancy rules we consider, and construct a $(\vdash^i, \vdash^d)$ reasoning connection based on these rules for introspection and linear ordered resolution for deduction. We prove that this is indeed a reasoning connection.

4.1 Orderings on Terms and Clauses

Let $<_t$ be a total, well-founded ordering on ground terms that extends the subterm relation, *i.e.* if t is a subterm of t' then $t <_t t'$.

An *atom ordering* is a total, well-founded ordering $<_a$ on all ground atoms of the signature. The atom ordering $<_a$ *is compatible with the term ordering* $<_t$ if $p(r_1, \ldots, r_n) <_a q(s_1, \ldots, s_m)$ implies that for $1 \leq j \leq n$ there exists $1 \leq i \leq m$ such that $r_j <_t s_i$. Terms and atoms ordering can be lifted into a partial ordering on terms and atoms with variables by setting $a \prec_a b$ if, for all substitution σ grounding both A and B we have $A\sigma <_a B\sigma$. Since this order is necessarily partial (variables don't compare, for example), we say that an atom A is (strictly) maximal wrt a set Γ of atoms if for every $B \in \Gamma$ we have $A \not\prec_a B$ ($A \not\preceq B$). Since a well-founded total ordering on a set E can always be extended to well-founded total ordering on multisets of elements of E, we extend atom orderings on literals by mapping a positive literal A to the multiset $\{\!|\, A \,|\!\}$ and a negative literal to the multiset $\{\!|\, A, A \,|\!\}$. Ground clauses are then ordered as multisets of multisets of atoms.

4.2 Ordered Resolution and Factorisation

Ordered resolution was introduced in [4]. We consider in this paper post-ordered resolution with an *ad hoc* redundancy rule. In short, *post-ordered* means that

once a unifier has been computed in the resolution of factorisation rule, we check that the instance of the atom that will be eliminated is maximal wrt to the other instantiated atoms (post ordered resolution). Regarding redundancy, if there is a proof π of $T \setminus \neg C \vdash^o \emptyset$ such that the multiset M_π of atoms resolved or factored upon is smaller than the multiset of atoms occurring in C we say that C is redundant in T, and introduce an introspection rule $T \vdash^i T \setminus \{C\}$.

The other introspection rules are presented formally in Fig. 2. In the ordered factorisation rule, Γ is the subset of negative literals while Δ is the subset of positive literals in the clause.

$$\frac{\Gamma \vee A \quad \neg B \vee \Delta}{(\Gamma \vee \Delta)\sigma}\ \sigma = \mathrm{mgu}(A, B)$$

ordered resolution

$A\sigma$ strictly maximal wrt $\Gamma\sigma$ and $A\sigma$ maximal wrt $\Delta\sigma$. In this case we have $T \vdash^i T \cup \{(\Gamma \vee \Delta)\sigma\}$ whenever $\Gamma \vee A, \neg B \vee \Delta \in T$

$$\frac{\Gamma \vee A \vee B}{(\Gamma \vee A)\sigma}\ \sigma = \mathrm{mgu}(A, B)$$

ordered factorisation

$A\sigma$ strictly maximal wrt $\Gamma\sigma$ and $A\sigma$ maximal wrt $\Delta\sigma$. In this case we have $T \vdash^i T \cup \{(\Gamma \vee A)\sigma\}$ whenever $\Gamma \vee A \vee B \in T$

Fig. 2. Ordered resolution and factorisation inference rules

Ordered resolution is complete for refutation: if $T \models \varphi$, then $T, \neg\varphi \models \mathit{emptyset}$, and in this case there exists a deduction sequence from $T \cup \{\neg\varphi\}$ to $\emptyset$. We denote $T \vdash^o \varphi$ the fact that such a sequence exists. Ordered resolution is complete for refutation: If T is an inconsistent theory then $T \vdash^o \emptyset$.

In this section we consider for simplicity reasons only first-order theories presented by a finite set of clauses, and note that it suffices to prove ground clauses φ. For such clauses we let $\neg\varphi$ be the set of the negation of the literals occurring in φ. Finally we denote $T \vdash^{lin} \emptyset$ the existence of a deduction sequence produced by linear ordered resolution and factorisation ending with the empty clause.

That is, in the rest of this section, we let:

- $\vdash^d$ be the relation $T \vdash^d \varphi$ if and only if $T, \neg\varphi \vdash^{lin} \emptyset$;
- $\vdash^i$ be the relation $T \vdash^i T'$ if T' can be obtained from T by either:
  - the addition of the conclusion of an ordered factorisation or ordered resolution rule;
  - the removal of a clause C such that $T \setminus \{C\} \vdash^d C$ with a deduction sequence for linear ordered resolution in which the multiset of atoms resolved or factored upon is smaller than the multiset of atoms occuring in C.

It is trivial that $\vdash^i$ is truth preserving and $\vdash^d$ is sound. So the rest of this section is devoted to proving that $(\vdash^i, \vdash^d)$ is a reasoning connection. We assume from now on that T is a satisfiable set of clauses. We let $\_g$ be a grounding operator mapping all variables of a clause to distinct constants not occurring

in T. In that case, it is well-known that $T \models \varphi$ if, and only if, $T \models \varphi^g$. As a consequence and in order to simplify notation, we consider only the entailment of ground Horn clauses, though the results lift to all clauses in the language.

First, let us note that we have taken an apparently weaker notion of redundancy than the one which is usually considered (*e.g.* in [3]) as usually there is no constraint on the atoms of C in the local proof. Requiring the proof redundancy to be linear implies in particular that we do not allow the application of an ordered factorisation on the instances of clauses of T. This constraint is necessary to prove the not-going-back property, and can always be enforced by preliminary saturation steps on T before eliminating the clause C by redundancy. That is, it may make $\mathrm{dist}_{\vdash^i,\vdash^d}(\varphi, T)$ greater than if it were not imposed, but has the advantage of ensuring a monotonous decrease of $\mathrm{dist}_{\vdash^i,\vdash^d}(\varphi, T)$ for all φ when performing an introspection step.

Since the definition employed are not exactly standard, we first prove that we still have refutational completeness, that is that the reasoning connection has the completeness property.

Lemma 2. *Let φ be a ground Horn clause, and assume $T \models \varphi$. Then* $\mathrm{dist}_{\vdash^i,\vdash^d}(\varphi, T) < +\infty$.

Proof. For $T \in \mathrm{Con}$ and φ such that $T, \neg\varphi \models \emptyset$, let $\mathit{\Pi}_{T,\varphi} = \{[C_1, \ldots, C_n] \,|\, C_n = \emptyset\}$ be the set of ground deduction sequences by ordered resolution and factoring from $T \cup \neg\varphi$ that ends with the empty clause. By completeness of ordered resolution and factoring, we know this set is not empty.

Also, let $\mathit{\Theta}_{T,\varphi}$ be the function mapping a proof to the multiset of atoms occuring in the proof and eliminated by ordered factorisation or resolution between clauses of T:

$$\begin{aligned} \mathit{\Theta}_{T,\varphi} : \quad & \mathit{\Pi}_{T,\varphi}(\to A \cup \bot) \times \mathbb{N} \\ & \pi \mapsto \{\!|\, a \,|\, a \text{ resolved or factored in } \pi \text{ from clauses in } T \,|\!\} \end{aligned}$$

By definition if $\mathit{\Theta}_{T,\varphi}(\pi) = \emptyset$ then π is a linear proof by ordered resolution and factorisation that $T, \neg\varphi \models \emptyset$, and thus $T, \varphi \vdash^i \emptyset$.

By contradiction let's assume there exists T, φ such that $T, \neg\varphi \models \emptyset$ but $\mathrm{dist}_{\vdash^i,\vdash^d}(\varphi, T)$ is not finite. The extension to multisets of the well-founded ordering on atoms is a well-founded ordering on the set $\mathit{\Theta}_{T,\varphi}(\mathit{\Pi}_{T,\varphi})$. By refutational completeness of ordered resolution and factoring this set is not empty, and thus has a minimal element π. If $\mathit{\Theta}_{T,\varphi}(\pi) = \emptyset$ then by definition pi is also a linear proof, and thus $T, \neg\varphi \vdash^i \emptyset$, and thus $\mathrm{dist}_{\vdash^i,\vdash^d}(\varphi, T) = 0$, a contradiction.

Thus $\mathit{\Theta}_{T,\varphi}(\pi) \neq \emptyset$. It is finite, and thus contains a maximal element a with at least 1 occurrence. By definition a is an atom either resolved upon or factored upon from clauses in T. Since this inference is by ordered resolution its result is an instance of a clause C with $T \vdash^d T \cup \{C\}$.

If C is redundant in T, in which case there exists a proof of redundancy of C in which the multiset of atoms resolved or factored upon is smaller than the multiset of atoms occurring in C. Since a is maximal among the atoms occurring in C none of these atoms are greater than a, and thus replacing C

with the redundancy proof yields a proof pi' with $\Theta_{T,\varphi}(\pi') < \Theta_{T,\varphi}(\pi)$, thus contradicting the minimality of π.

Thus C is not redundant in T. Let π' be the proof obtained from pi by replacing the inference on a by the introduction of C. We have:

$$\Theta_{T',\varphi}(\pi') < \Theta_{T,\varphi}(\pi)$$

Let $M_0 = \Theta_{T,\varphi}(\pi), M_1 = \Theta_{T',\varphi}(\pi'), \ldots$ be the sequence of multisets that can be obtained by iterating this construction. The ordering on multisets is well-founded, so this sequence is finite. Thus there exists n such that $T = T_0(\vdash_i)^n T_n$ and $\mathcal{T}_n, \neg\varphi \vdash^i \emptyset$. But then this means that $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T) \leq n$, a contradiction with the assumption that $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T)$ is not finite.

The proof of the preceding lemma can be adapted as follows. For each formula φ, order partially the set of theories T' reachable from T with the minimal number $n(T')$ of steps that will lead to a theory in which φ is provable with $\vdash^d$. Lemma 2 tells us that this number is always defined. Now, from T, instead of choosing the saturation step that will minimize the number of occurrences of the maximal atoms not occurring in φ, choose a T_1 with a minimal $n(T_1)$. By definition we have $n(T_1) = n(T) - 1$. Hence we have the following progress lemma.

Lemma 3. *Let φ be a ground Horn clause, and assume $T \models \varphi$ and and $T \not\vdash^d \varphi$. Then there exists T' such that $T \vdash^i T'$ and $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T') < \text{dist}_{\vdash^i,\vdash^d}(\varphi, T)$.*

Finally we prove the *not-going-back* property by proving that even the removal by redundancy of a clause in T does not increase the distance

Lemma 4. *Assume $T \models \varphi$ and $T \vdash^i T'$. Then $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T') \leq \text{dist}_{\vdash^i,\vdash^d}(\varphi, T)$.*

Proof. Since ordered resolution and factorisation simply add a new clause, the proof is trivial if T_0 is obtained from T using one of these inferences. Therefore, let us consider only the case $T \vdash^i T \setminus \{C\} = T'$ where C is a redundant clause in T, *i.e.* is such that $T \vdash^d C$. Letπ_C be a deduction sequence witnessing $T' \vdash^d C$ in which the mulstiset of atoms factored or resolved upon is smaller are equal to, for the atom ordering $<_a$, the multiset of atoms occurring in C. , and in which the multiset of atoms factored or resolved upon is smaller than the multiset of atoms occuring in C.

Given a deduction sequence π, let $R_C(\pi)$ be the deduction sequence in which every position in which a ground instance $C\sigma$ is introduced is replaced by the sequence of introductions of instances of clauses of T' in π_C . Since the multiset of atoms occurring in π_C is less or equal to the multiset of atoms in C, it is clear that for any proof π we have $Theta_{T',\varphi}(R_C(\pi)) \leq \Theta_{T',\varphi}(\pi)$. Let $T_0 = T \vdash^i T_1 \vdash^i \ldots \vdash^i T_n$ be a minimal sequence of introspections such that $T_n \vdash^d \varphi$. We can perform the replacement on all proofs in $\Pi_{T_i,\varphi}$ in this sequence and obtain a sequence of the same length from $T' = T'_0$ to T'_n. Thus $\text{dist}_{\vdash^i,\vdash^d}(\varphi, T') \leq \text{dist}_{\vdash^i,\vdash^d}(\varphi, T)$.

5 Conclusion and Future Works

The main contribution is the definition of an abstract reasoning couple that captures an interplay between introspection (as a change in the representation of the world not triggered by an external input) and deduction as a judgement on one's beliefs, a set of statements assumed satisfiable. The framework presented is a first step towards more interesting agent models, for which one would need at least:

- the possibility to change the knowledge, for example to reflect the agent's experience;
- the possibility for the agent to perform some actions, that is to influence the world based on her beliefs and knowledge.

Furthermore, there is no reasoning based on beliefs in the proposed framework. The proper extension in this case consists in introducing a generic structure such as a directed acyclic graph in which vertices or members are labelled with theories, and the couples (T, B) introduced in this paper would be modeled by a binary relation on vertices.

Also, on a more technical note, this works stems from previous work on the compilation of cryptographic protocols, that is from giving an operational semantics in a precise agent model to agents that will implement a protocol specification as well as possible. In that context, we hope that further extensions of the framework presented in this paper will be used as target agent models.

Acknowledgements. We would like to thank Philippe Besnard and Michal Rusinowitch for insightful comments on a first draft of this paper.

References

1. Armando, A., Ranise, S., Rusinowitch, M.: Uniform derivation of decision procedures by superposition. In: Fribourg, L. (ed.) Computer Science Logic: 15th International Workshop, CSL 2001 10th Annual Conference of the EACSL Paris, France, September 10–13, 2001, Proceedings. pp. 513–527. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
2. Bachmair, L., Ganzinger, H.: Rewrite-based equational theorem proving with selection and simplification. J. Log. Comput. 4(3), 217–247 (1994)
3. Basin, D., Ganzinger, H.: Automated complexity analysis based on ordered resolution. J. ACM 48(1) (Jan 2001)
4. Besnard, P., Quiniou, R., Quinton, P.: A theorem-prover for a decidable subset of default logic. In: Proceedings of the National Conference on Artificial Intelligence. Washington, D.C., August 22-26, 1983. pp. 27–30 (1983)
5. Dershowitz, N., Jouannaud, J.P.: Rewrite systems. In: Handbook of Theoretical Computer Science, Volume B, pp. 243–320. Elsevier (1990)
6. Knuth, D.E., Bendix, P.B.: Simple word problems in universal algebras. In: Siekmann, J.H., Wrightson, G. (eds.) Automation of Reasoning: 2: Classical Papers on Computational Logic 1967–1970. pp. 342–376. Springer Berlin Heidelberg, Berlin, Heidelberg (1983)

7. Konolige, K.: A deduction model of belief and its logics. Tech. Rep. 326, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025 (Aug 1984)
8. Kowalski, R., Kuehner, D.: Linear resolution with selection function. Artificial Intelligence 2 (1971)
9. Lynch, C., Morawska, B.: Automatic decidability. In: Proceedings of the *Logic in Computer Science Conference*. pp. 7–17. IEEE Computer Society (2002)
10. McAllester, D.A.: Automatic recognition of tractability in inference relations. J. ACM 40(2), 284–303 (1993)
11. de Nivelle, H.: Ordering refinements of resolution. Ph.D. thesis, Technische Universiteit Delft (1996)
12. Robinson, J.A.: A machine-oriented logic based on the resolution principle. Journal of the Association for Computing Machinery 12, 23–41 (1965)