A Defeasible Logic of Institutional Agency

Guido Governatori
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
Email: guido@itee.uq.edu.au

Antonino Rotolo
CIRSFID
University of Bologna
Via Galliera 3, 40121, Bologna, Italy
Email: rotolo@cirfid.unibo.it

Abstract

A non-monotonic logic of institutional agency is defined combining a computationally oriented non-monotonic system (Defeasible Logic) and intensional notions of agency.

1 Background and Motivation

In recent works on agents and on their societies, a specific normative line of research has been emerging. This research assumes that as in human societies, also in artificial societies normative concepts may play a decisive role, allowing for the flexible co-ordination of intelligent autonomous agents (see, e.g., [Conte and Dellarocas, 2001]).

Of course, there are number of ways that the issue of the role of normative concepts in MAS can be treated [Conte and Dellarocas, 2001]). Among them, a formal approach that makes use of a multi-modal logical setting seems to be promising. As recently pointed out regarding the design of computerised multi-agent systems, "modal logic [is] a means of supplying an intermediate level of description, falling somewhere between [...] ordinary-language account of what a system [...] is supposed to be able to do and [...] the level of implementation" [Jones, 2003]. In this perspective, a logical analysis of normative notions such as institutions, powers, obligations, responsibilities, delegation, etc., is one precondition for the development of norm-governed societies.

The background of this paper comes from the well-known Kanger-Lindahl-Pörn [Kanger, 1972; Lindahl, 1977; Pörn, 1977] logical theory to account for agency and organised interaction (see [Elgesem, 1997]). Our starting point is to take advantage of some recent contributions [Santos and Carmo, 1996; Santos *et al.*, 1997; Jones and Sergot, 1996; Jones, 2003], which have enriched this framework with some substantial refinements. Despite some well-known limitations (see [Elgesem, 1997; Segerberg, 1992; Royakkers, 2000]), such an approach allows to easily combine, e.g., actions with a number of other concepts, like powers, obligations, beliefs, etc. It also permits to provide a simple conceptual analysis of the structure of organisations of agents.

This paper is about how some intuitions from the above line of research can be embedded in a (computationally oriented) non-monotonic framework to account for the fundamental activities performed within an organisation. It is a preliminary work. In fact, our analysis does not deal directly with the multi-agent dimension of organisations. Rather it is confined to two main aspects: the modal notion of agency and that of institutionalised power. In this sense, we will explore them only with regard to the case of single-agent contexts. Notice also that the basic notions described in [Santos and Carmo, 1996; Santos *et al.*, 1997; Jones and Sergot, 1996; Jones, 2003] are simply reframed here to develop a computational treatment of institutional agency. No new concepts are added to the existing logical framework.

As regards agency, we will focus on two notions. The first is the idea of personal and direct action to realise a state of affairs. In the mentioned logical framework, it is formalised by the well-known modal operator E, such that a formula like E_iA means that the agent i brings it about that A. Different axiomatisations have been provided for it but almost all are characterised by $E_iA \rightarrow A$ (T, i.e., successfulness), $\neg E_i\top$ (No), $(E_iA \wedge E_iB) \rightarrow E_i(A \wedge B)$ (C, Agglomeration), and are closed under logical equivalence [Santos and Carmo, 1996]. The second notion is that of attempt, formalised by the operator H [Santos *et al.*, 1997; Jones, 2003]. H_iA says that i attempts to make it the case that A. The operator H is not necessarily successful. Besides that, it enjoyes Agglomeration as well and is also closed under logical equivalence. On the other hand, there are arguments for adopting [Jones, 2003] or not [Santos *et al.*, 1997] the analogous schema (No) for H: for the sake of simplicity, here we will follow the latter option. Finally, notice that we have $E_iA \rightarrow H_iA$.

One of the main limits of modal logic of agency is also one of its main advantages. In fact, such a logic is very general since actions are simply taken to be relationships between agents and states of affairs. This does not make the logic very expressive in itself, but, as we alluded to, allows flexibly for the combination of agency with a number of other concepts. In particular, this holds in order to characterise the idea of institutionalised power. Such a notion is central for describing norm-governed organisations of agents and comes from the distinction between the practical ability to realise a state of affairs —which is not considered in this paper [Elgesem, 1997]— and the institutional power to do this [Makinson, 1986]. For example, if in an auction i raises one hand, this implies that the act of making a bid is also obtained. In

principle, this kind of ability should be distinguished from the practical capacity to obtain a certain state of affairs. In fact, the attempt to make a bid may not be successful. Whether it is successful or not, within the institutional context (the auction), depends on whether that institution makes it effective. It is up to the institutional rules to establish whether *i*'s act, in the conditions in which it is made, makes so that a bid is effective or not. According to Searle [Searle, 1995], the rules through which institutions make effective these attempts are constitutive in character and have the form "*X* counts as *Y* in the context *C*". Their function is to create a special kind of facts, whose nature is institutional and conceptually distinct from that of the empirical facts.

In their seminal [Jones and Sergot, 1996], Jones and Sergot developed a formal approach to the notion of institutionalised power by introducing a new conditional connective "$\Rightarrow_s$". This connective expresses the "counts as" connection holding in the context of an institution *s*. In particular, when applied to action descriptions, a formula like $E_iA \Rightarrow_s E_iB$ represents *i*'s institutional power to produce *B* when *A* is realised (see [Jones and Sergot, 1996; Jones, 2003]). In a similar vein, but more closely to Searle's intuition, it has been argued [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] that the counts-as link is composed by a normative conditional $\Rrightarrow$ corresponding at least to cumulative logic (system **CU** [Artosi *et al.*, 2002]), plus a restricted form of Modus Ponens, and the modality D_s — introduced in [Jones and Sergot, 1996] but with a different meaning— to represent institutional facts. In this perspective, $A \Rightarrow_s B =_{def} (A \Rrightarrow D_sB) \wedge (D_sA \Rrightarrow D_sB)$, to capture the fact that counts-as rules may specify when (1) a brute fact (e.g., destroying the receipt) counts as a type of institutional act (e.g., freeing the debtor from his obligation), and (2) an institutional act (e.g., a contract made by person *j* in the name of person *k*) has the same effects of another institutional act (e.g., a contract made by *k*). D_s represents the domain of institutional facts and so it cannot be a normal modality. In fact, the weakening of counts-as consequents is not acceptable in the setting of [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] since, from $D_s(making\_a\_bid)$ should not follow $D_s(making\_a\_bid \vee drinking\_some\_water)$. In this sense, D_s is a non-normal modality closed under logical equivalence and satisfying Agglomeration and Consistency. Of course, necessitation does not hold: it sounds strange that $\top$ is an institutional fact for any institution *s*[1]. Finally, notice that the axiom $D_sE_iA \rightarrow D_sA$ was adopted to guarantee successfulness also within the domain of every institution *s*.

Basically, we will follow here the intuitions presented in [Governatori *et al.*, 2002a; Gelati *et al.*, 2002]. Though in different perspectives, however, an important point shared by [Jones and Sergot, 1996; Jones, 2003] and [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] is that the counts-as link is defeasible. This is a crucial feature of this notion. In fact, it is intuitive that, e.g., if the agent *i* raises one hand, this may count as making a bid but this does not hold if *i* raises one hand *and* scratches his own head.

The goal of this paper is to develop a computational framework, based on Defeasible Logic, able to treat this kind of institutional mechanisms. Although the above approaches provide an interesting analysis, they can hardly be used for implementation. This is clearly due at least to the well known computational limits of conditional logics (see, e.g., [Artosi *et al.*, 2002]). In this perspective, some basic patterns of defeasible reasoning will be re-framed and extended to account for the institutional dynamics insofar as they are interplayed with the notions of direct action and attempt.

2 Overview of Defeasible Logic

Defeasible Logic is a simple, efficient but flexible non-monotonic formalism which has been proven able to deal with many different intuitions of non-monotonic reasoning [Antoniou *et al.*, 2000b]. In the last few years it has been applied in many fields. Some of the particular applications require intensional notions; to this end some extensions of Defeasible Logic, designed to capture such intensional notions (usually described by modal operators), have been put forward [Governatori *et al.*, 2002b]. In this paper we propose a non-monotonic logic of agency based on the framework for Defeasible Logic proposed in [Antoniou *et al.*, 2000a].

It is not possible in this short paper to give a complete formal description of the logic. However, we hope to give enough information to make the discussion intelligible. We refer the reader to [Nute, 1987; Billington, 1993; Antoniou *et al.*, 2001] for more thorough treatments. As usual with non-monotonic reasoning, we have to specify 1) how to represent a knowledge base and 2) the inference mechanism.

Accordingly a defeasible theory *D* is a structure $(F, R, >)$ where *F* is a finite set of facts, *R* a finite set of rules (either strict, defeasible, or defeater), and $>$ a binary relation (superiority relation) over *R*.

Facts are indisputable statements. *Strict rules* are rules in the classical sense: whenever the premises are indisputable (e.g., facts) then so is the conclusion; *defeasible rules* are rules that can be defeated by contrary evidence; and *defeaters* are rules that cannot be used to draw any conclusions. Their only use is to prevent some conclusions. In other words, they are used to defeat some defeasible rules by producing evidence to the contrary. The *superiority relation* among rules is used to define priorities among rules, that is, where one rule may override the conclusion of another rule.

A *rule r* consists of its *antecedent* (or *body*) $A(r)$ ($A(r)$ may be omitted if it is the empty set) which is a finite set of literals, an arrow, and its *consequent* (or *head*) $C(r)$ which is a literal. Given a set *R* of rules, we denote the set of all strict rules in *R* by R_s, the set of strict and defeasible rules in *R* by R_{sd}, the set of defeasible rules in *R* by R_d, and the set of defeaters in

[1] In [Jones and Sergot, 1996] D_s corresponds to a normal **KD** modality. This is suggested to express all (logical, causal, institutional, etc.) constraints on *s* among which the link "counts as" is included. In other words, D_sA means that *A* is generically "recognised by the institution *s*". In [Governatori *et al.*, 2002a; Gelati *et al.*, 2002], as pointed out, the reading of this modality is different since it is meant strictly to represent the domain of institutional facts of a given *s*. In this sense, the English gloss for the expression D_sA is "*A* is an institutional fact holding within the institution *s*".

R by R_{dft}. $R[q]$ denotes the set of rules in R with consequent q. If q is a literal, $\sim q$ denotes the complementary literal (if q is a positive literal p then $\sim q$ is $\neg p$; and if q is $\neg p$, then $\sim q$ is p).

A *conclusion* of D is a tagged literal and can have one of the following four forms:

$+\Delta q$ which is intended to mean that q is definitely provable in D (i.e., using only facts and strict rules).

$-\Delta q$ which is intended to mean that we have proved that q is not definitely provable in D.

$+\partial q$ which is intended to mean that q is defeasibly provable in D.

$-\partial q$ which is intended to mean that we have proved that q is not defeasibly provable in D.

Provability is based on the concept of a *derivation* (or proof) in D. A derivation is a finite sequence $P = (P(1), \ldots P(n))$ of tagged literals satisfying four conditions (which correspond to inference rules for each of the four kinds of conclusion). $P(1..i)$ denotes the initial part of the sequence P of length i

$+\Delta$: If $P(i+1) = +\Delta q$ then
(1) $q \in F$ or
(2) $\exists r \in R_s[q]\ \forall a \in A(r) : +\Delta a \in P(1..i)$

$-\Delta$: If $P(i+1) = -\Delta q$ then
(1) $q \notin F$ and
(2) $\forall r \in R_s[q]\ \exists a \in A(r) : -\Delta a \in P(1..i)$

The definition of Δ describes just forward chaining of strict rules. For a literal q to be definitely provable we need to find a strict rule with head q, of which all antecedents have been definitely proved previously. And to establish that q cannot be proven definitely we must establish that for every strict rule with head q there is at least one antecedent which has been shown to be non-provable.

$+\partial$: If $P(i+1) = +\partial q$ then either
(1) $+\Delta q \in P(1..i)$ or
  (2.1) $\exists r \in R_{sd}[q]\ \forall a \in A(r) : +\partial a \in P(1..i)$ and
  (2.2) $-\Delta{\sim}q \in P(1..i)$ and
  (2.3) $\forall s \in R[\sim q]$ either
    (2.3.1) $\exists a \in A(s) : -\partial a \in P(1..i)$ or
    (2.3.2) $\exists t \in R_{sd}[q]$ such that $t > s$ and
      $\forall a \in A(t) : +\partial a \in P(1..i)$

Let us work through this condition. To show that q is provable defeasibly we have two choices: (1) We show that q is already definitely provable; or (2) we need to argue using the defeasible part of D as well. In particular, we require that there must be a strict or defeasible rule with head q which can be applied (2.1). But now we need to consider possible "attacks", i.e., reasoning chains in support of $\sim q$. To be more specific: to prove q defeasibly we must show that $\sim q$ is not definitely provable (2.2). Also (2.3) we must consider the set of all rules which are not known to be inapplicable and which have head $\sim q$ (note that here we consider defeaters, too, whereas they could not be used to support the conclusion q; this is in line with the motivation of defeaters given earlier). Essentially each such rule s attacks the conclusion q. For q to be provable, each such rule s must be counterattacked by a rule t with head q with the following properties: (i) t must be applicable at this point, and (ii) t must be stronger than s. Thus each attack on the conclusion q must be counterattacked by a stronger rule. In other words, r and the rules t form a team (for q) that defeats the rules s. In an analogous manner we can define $-\partial q$ as

$-\partial$: If $P(i+1) = -\partial q$ then
(1) $-\Delta q \in P(1..i)$ and
  (2.1) $\forall r \in R_{sd}[q]\ \exists a \in A(r) : -\partial a \in P(1..i)$ or
  (2.2) $+\Delta{\sim}q \in P(1..i)$ or
  (2.3) $\exists s \in R[\sim q]$ such that
    (2.3.1) $\forall a \in A(s) : +\partial a \in P(1..i)$ and
    (2.3.2) $\forall t \in R_{sd}[q]$ either $t \not> s$ or
      $\exists a \in A(t) : -\partial a \in P(1..i)$

The purpose of the $-\partial$ inference rules is to establish that it is not possible to prove $+\partial$. This rule is defined in such a way that all the possibilities for proving $+\partial q$ (for example) are explored and shown to fail before $-\partial q$ can be concluded. Thus conclusions tagged with $-\partial$ are the outcome of a constructive proof that the corresponding positive conclusion cannot be obtained.

Sometimes all we want to know is whether a literal is *supported*, that is if there is a chain of reasoning that would lead to a conclusion in absence of conflicts. This notion is captured by the following proof conditions:

$+\Sigma$: if $P(i+1) = +\Sigma p$ then
(1) $+\Delta p \in P(1..i)$ or
(2) $\exists r \in R_{sd}[p]\ \forall a \in A(r) : +\Sigma a \in P(1..i)$

$-\Sigma$: if $P(i+1) = -\Sigma p$ then
(1) $-\Delta p \in P(1..i)$ and
(2) $\forall r \in R_{sd}[p]\ \exists a \in A(r) : -\Sigma a \in P(1..i)$

The notion of support corresponds to monotonic proofs using both the monotonic (strict rules) and non-monotonic (defeasible rules) parts of defeasible theories.

Notice that all the proof conditions satisfy the Principle of Strong Negation introduced in [Antoniou *et al.*, 2000a]. The strong negation of a formula is closely related to the function that simplifies a formula by moving all negations to an innermost position in the resulting formula and replace the positive tags with the respective negative tags and viceversa. In what follows we will assume that all inference rules are defined according to this principle; consequently we list only the positive version of the inference rules.

3 A Defeasible Logic of Institutional Agency

As we have seen in Section 1 modal logics have been put forward to capture the intensional nature of (institutional) agency[2]. Usually modal logics are extensions of classical propositional logic with some intensional operators. Thus any modal logic should account for two components: (1)

[2]In this context multi-modal logics have been adopted to cope with multi-agent institutions. As we said, however, we limit ourselves to single-agent contexts, since the main aim of the paper is to demonstrate how to capture the non-monotonic nature of agency in the proposed framework, and not the relationships among agents in societies.

the underlying logical structure of the propositional base and (2) the logic behavior of the modal operators. Alas, as is well-known, classical propositional logic is not well suited to deal with real life scenarios. The main reason is that the descriptions of real-life cases are, very often, partial and somewhat unreliable. In such circumstances classical propositional logic might produce counterintuitive results insofar as it requires complete, consistent and reliable information. Hence any modal logic based on classical propositional logic is doomed to suffer from the same problems.

On the other hand the logic should specify how modalities can be introduced and manipulated. Some common rules for modalities are, e.g., Necessitation and RM [Chellas, 1980]. Both dictate conditions for introducing modalities in contrast with the analysis of institutional agency as outlined in Section 1. To comply with the properties of this notion, in the setting provided by Defeasible Logic we have to set 1) the rules describing the logical inferences and 2) the rules to introduce the modal operators of agency E (*brings it about*), and H (*attempts*). Accordingly we will consider two types of rules: a set of rules (strict, defeasible, and defeaters) for the notion of *counts-as*, and a set of rules (strict, defeasible, and defeaters) for the notion of *results-in*.

Since we want to be able to reason about actions we extend the language of Defeasible Logic with a set of action symbols; we will use α, β, γ to denote atomic actions. The intended meaning of an action symbol, for example α, is that the action corresponding to it has been performed, while we use $\neg\alpha$ to denote that the action described by α has not been performed. Given the modal operators E and H we form new literals as follows: if l is a propositional literal then El, $\neg El$, Hl and $\neg Hl$ are modal literal. A literal is either a propositional literal or a modal literal.

In this perspective a defeasible institutional action theory is a structure

$$I = (F, R^c, R^r, >)$$

where, as usual F is a set of facts, R^c is a set of counts-as rules (i.e., $\rightarrow_c$, $\Rightarrow_c$, $\rightsquigarrow_c$), R^r is a set of results-in rules (i.e., $\rightarrow_r$, $\Rightarrow_r$, $\rightsquigarrow_r$), and $>$, the superiority relation, is a binary relation over the set of rules (i.e., $> \subseteq (R^c \cup R^r)^2$).

The intuition is that, given an institution, F consists of the description of the raw institutional facts, either in form of states of affairs (literal and modal literal) and actions that have been performed. R^c describes the basic inference mechanism internal to an institution, while R^r encodes the transitions from state to state occurring as the results of actions. Technically R^r is used to introduce modal operators. In order to correctly capture these notions we impose some restrictions on the form of rules: modal literals can occur only in the antecedent of rules, while actions symbols are not permitted in the consequent of results-in rules. The first restriction is motivated from the fact that 1) results-in rules are the rules to introduce the modalities and in the present context nested literals are meaningless 2) counts-as rules make possibile the derivation of institutional actions (modalised literals) only when they follow from specific actions (intentionally) performed by the agent. The second restriction is due to the idea that results-in rules describe, as their name suggests, the results of actions, not actions themselves.

Let us see by means of some examples the intuition behind this formalism. We focus here on defeasible rules but similar remarks can be applied to the other kinds of rules. Suppose the agent i is acting in the context of an auction. Then we may have cases like the following[3]:

$$\textbf{bids}, \textit{auction\_begun} \Rightarrow_r \textit{offer}$$

This rule is an example corresponding to the introduction of the modality E. In fact, i's fulfilment of the conditions in the antecedent produces the occurrence of *offer*: i's action of bidding has the result that i has made an offer. As we will see, if *offer* can be derived, this permits the introduction of $E_i(\textit{offer})$.

$$\textit{auction\_begun} \Rightarrow_r \neg \textit{offer}$$

The example above does not specify any action in the antecedent (empty action). This means that, when the auction is begun, i's refraining from doing any action has the result to have no offer. In logical terms, also this case can lead to the introduction of E[4].

Let us consider examples of counts-as rules.

$$\textbf{raises\_hand}, \textit{auction\_begun} \Rightarrow_c \textbf{bids}$$

This rules says that that i's action of raising one hand counts as i's action of bidding, when the auction is begun.

$$\textit{auction\_begun}, E_i(\textit{offer}) \Rightarrow_c \neg \textbf{raises\_offer}$$

Also here we have i's generic refraining from doing any action in the antecedent. This example represents the institutional connection linking such refraining, and *the fact* that i made an offer when the auction is begun, to i's specific refraining from raising a new offer. Notice that the same meaning is assigned to counts-as rules where the antecedent contains only non-modal literals.

$$\textit{auction\_begun}, \textbf{raises\_hand} \Rightarrow_c \textit{offer}$$

This rule is an example of the institutional analogous of results-in rules, where an action and a state of affairs occur respectively in their antecedent and consequent. However, in this case the result is an institutional fact and follows by convention only within the institution. In fact, that an offer is a consequence of i's raising one hand is not a simple matter of i's action results. The attempt of i to make an offer by raising the hand is effective only if the institution recognises this.

It is worth noting that no explicit reference is made here to the modality D_s as introduced in [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] and recalled in Section 1. In fact, the present setting accounts for the idea of institution in terms a special kind of defeasible theory. Each institutional action theory I encodes in itself all possible inferences that can be drawn within the domain of institutional facts relative to a given s. This means that s may be identified with

[3] Bold type expressions correspond to action symbols, the italicized ones to state of affairs.

[4] The ideas of empty action and refraining from doing a specific action should not be confused with what it is expressed by $\neg E_i A$. As we will see, this last corresponds to the non-derivability of A within I, which can depend also on reasons that have nothing to do with i's refraining from acting to realise A.

I since all action results are obtained within such a domain of facts. In other words, the introduction of the modality D_s corresponds here to the general definition of derivability using counts-as and results-in rules. Technically, counts-as rules are meant to capture the case $D_sA \Rrightarrow D_sB$ mentioned in Section 1. Roughly speaking, on the other hand, the case $A \Rrightarrow D_sB$ will be treated as a special kind of results-in rule, where the manipulation of the consequent is made under the constraints designed to account for the idea of institutional consequence. This is just a technical device to differentiate the two cases: the logic behaviour of the counts-as link as described in [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] is here encoded in the whole formal machinery corresponding to the definitions of the proof conditions.

Such proof conditions are as follows. For counts-as derivability (R^c) we assume the basic conditions of Defeasible Logic given in Section 2. Thus $\pm\Delta_c$, $\pm\partial_c$ correspond, respectively, to $\pm\Delta$ and $\pm\partial$.

The conditions for derivations involving results-in rules are more complicated since we have to cater for more possibilities. First of all we have that $I \vdash Ep$ if either $I \vdash +\Delta_r p$ or $I \vdash +\partial_r p$, and $I \vdash Hp$ if $I \vdash +\Sigma_r p$. In other words it is possible to derive Ep if we have either a strict of defeasible derivation of p using both results-in and counts-as rules, and that an agent (in an institution I) attempts p (Hp) if I supports p using counts-as ad results-in rules. The output of a results-in rule produces E modal literals, and we have seen in Section 1 that the E operator is a success operator; therefore we add the conditions that it is possible to derive $+\Delta_c p$ from $+\Delta_r p$ and $+\partial_c p$ from $+\partial_r p$.

In the same way we have that $-\partial_r p$ corresponds to $\neg Ep$ and $-\Sigma p$ to $\neg Hp$. This is in agreement with the principle of strong negation used to define the inference conditions.

$+\Delta_r$: if $P(i+1) = +\Delta_r p$ then
(1) $Ep \in F$; or
(2) $\exists r \in R_s^c[p]$ $\forall a, \alpha \in A(r)$:
$+\Delta_r a, +\Delta_c \in \alpha \in P(1..i)$; or
(3) $\exists r \in R_s^r[p]$ $\forall a, Eb, \alpha \in A(r)$:
$+\Delta_c a, +\Delta_r b, +\Delta_c \alpha \in P(1..i)$.

To prove an indefeasible brings it about, we need either that it is given as a fact (1), or that we have a strict rule for results-in (an irrevocable policy) whose antecedent is indisputable (3). However we have another case (2): if an agent knows that B is an indisputable consequence of A in the institution (it always is the case that A counts as B), and it produces A, then it must realise B. This is in contrast with the NML interpretation whereby the agent has to brings it about all the consequences of his/her actions.

$+\Sigma_r$: if $P(i+1) = +\Sigma_r p$ then
(1) $Ep \in F$; or
(2) $\exists r \in R_{sd}^c[p]$ $\forall a, \alpha \in A(r)$:
$+\Sigma_r a, +\Sigma_c \alpha \in P(1..i)$; or
(3) $\exists r \in R_{sd}^r[p]$ $\forall a, Eb, \alpha \in A(r)$:
$+\Sigma_c a, +\Sigma_r b, +\Sigma_c \alpha \in P(1..i)$.

The inference conditions for H are very similar to those for strong brings it about; essentially they are monotonic proofs using both the monotonic part (strict rules) and the supportive non-monotonic part (defeasible rules) of a defeasible institutional action theory.

To capture the results of defeasible actions we have to use the superiority relations to resolve conflicts. Thus we can give the following definition for the inference rules for $+\partial_r$.

$+\partial_r$: if $P(i+1) = +\partial_r p$ then
(1) $+\Delta_r p \in P(1..i)$ or
(2.1) $-\Delta_c \sim p, -\Delta_r \sim p \in P(1..i)$ and
(2.2) $\exists r \in R_{sd}^c[p]$ $\forall a \in A(r) : +\partial_r a \in P(1..i)$, or
$\exists r \in R_{sd}^r[p]$ $\forall Eb, a, \alpha \in A(s)$:
$+\partial_r a, +\partial_c a, +\partial_c \alpha \in P(1..i)$; and
(2.3) $\forall s \in R[\sim p]$ either
(2.3.1) $\exists \alpha \in A(s) : -\partial_c \alpha \in P(1..i)$, or
(2.3.2) if $s \in R^c[\sim p]$ then
$\exists a \in A(s) : -\partial_c a \in P(1..i)$; and
if $s \in R^r[\sim p]$ then either
$\exists Ea \in A(s) : -\partial_r a \in P(1..i)$ or
$\exists a \in A(s) : -\partial_c a \in P(1..i)$; or
(2.3.3) $\exists t \in R[p]$ such that $t > s$ and
$\forall \alpha \in A(t) : +\partial_c \alpha \in P(1..i)$ and
if $t \in R^c[p]$ then $\forall a \in A(t) : +\partial_c a$; and
if $t \in R^r[p]$ then
$\forall a, Eb \in A(t) : +\partial_c a, +\partial_r a \in P(1..i)$.

The conditions for proving the results of defeasible actions are essentially the same as those given for defeasible derivations in Section 2. The only difference is that at each stage we have to check for two cases, namely: (1) the rule used is a results-in rule; (2) the rule is a counts-as rule. In the first case we have to verify that factual antecedents are defeasibly proved/disproved using counts-as ($\pm\partial_c$), and brings it about antecedents are defeasibly proved/disproved using results-in rules ($\pm\partial_r$). In the second case we have to remember that a conclusion of a institutional counts-as rule can be transformed into a results-in if all the literals in the antecedent are defeasibly executed.

Let us examine the above conditions at work with the help of some examples. We assume the following theory:

$$F = \{\alpha, p, Eq\}$$
$$R = \{r_1 : \alpha, p, Eq \Rightarrow_r s,$$
$$r_2 : s \Rightarrow_r r,$$
$$r_3 : r \Rightarrow_c t\}.$$

In this theory we are able to prove Et. The facts fire r_1, thus we can prove $+\partial_r s$ (Es). Now, since s has been brought about, s is the case. We can use this to fire the rule r_2. Hence we obtain $+\partial_r r$, which is Er. This implies that all the requisites of r_3 have been brought about; but r_3 states that r counts as t; this means that t has been brought about, hence $+\partial_r t$ and Et.

Let us replace r_3 with

$$r_3' : p, r \Rightarrow_c t$$

This time we can prove $+\partial_c t$, but not Et ($+\partial_r t$). The reason is that p is the case without a specific "intention" of the agent to bring it about. Similarly, if we replace r_3 by

$$r_3'' : Er \Rightarrow_c t$$

we can no longer derive Et. In this case Er is understood as a mere institutional fact, and not as the successful intention of the agent to realise r in order to realise t.

In the previous example we have seen how we can argue in favour of Ep (for same literal p). Let us examine the conditions to attack it. Let I be the following institutional defeasible theory

$$F = \{\alpha, p, q\}$$
$$R = \{r_1 : \alpha, p \Rightarrow_r s,$$
$$r_2 : q \Rightarrow_c r,$$
$$r_3 : p, r \Rightarrow_c \neg s\}.$$

Clearly Es ($+\partial_r s$) is not derivable from the given theory since there is an applicable rule for $\neg s$. r_3 is applicable since we can derive $+\partial_c r$. Similarly, if we replace r_2 with $q \Rightarrow_r r$, r_3 is still applicable. We can prove $+\partial_r r$: this means that there is a successful action resulting in r. In general to discard a rule we have to show that some of the premises cannot be derived. With a factual literal we have to show that the literal is not the case (or, in other terms, that there are no literals that count as it), and that the literal is not the result of a successful action: results of successful actions are indeed the case.

To illustrate the role of the superiority relation we use the following set of rules

$$R = \{r_1 : a \Rightarrow_c p,$$
$$r_2 : b \Rightarrow_r \neg p\}$$

in relation with some scenarios.

In the first case we have $F = \{a, b\}$. Here the two rules are conflicting both at the counts-as level and at the results-in level, and there is no way to solve the conflict. Hence we derive $-\partial_c p$ and $-\partial_r \neg p$ ($\neg E_i \neg p$).

In the second scenario we assume a and b as facts and $r_1 > r_2$. This means that we consider the counts-as rule stronger that the results-in rule. Accordingly we obtain $+\partial_c p$ and $\partial_r \neg p$. Consider the following concrete example

$$r_1 : \mathit{minor} \Rightarrow_c \neg \mathit{legallyResponsible}$$
$$r_2 : \textbf{signDocument} \Rightarrow_r \mathit{legallyResponsible}$$
$$r_1 > r_2$$

where we derive $+\partial_c \neg \mathit{legallyResponsible}$

In the third scenario we have the same facts as before but we reverse the order of the superiority relation, i.e., $r_2 > r_1$. In this case the consequences of the facts are $-\delta_c p$, $+\partial_c \neg p$ and $+\partial_r \neg p$ ($E_i \neg p$).

$$r_1 : \mathit{minor} \Rightarrow_c \neg \mathit{legallyResponsible}$$
$$r_2 : \textbf{signDocument}, \mathit{tutorApproval} \Rightarrow_r \mathit{legallyResponsible}$$
$$r_2 > r_1$$

Here we have that i brings it about that she is legally responsible.

The last case involves $E_i a$ and b as facts, and $r_1 > r_2$. The effect of $E_i a$ is that now r_1 can be conceived as a results-in rule, thus we have $+\partial_r p$, $+\partial_c p$; remember that the E_i operator is a success operator, and $-\partial_r \neg p$.

Let us consider the following chess example: the chess rules state that a drawn can be claimed after 50 consecutive moves without a capture or without moving a pawn (article 9.5, comma a of the Laws of Chess). This comma can be represented as follows:

$$r_1 : \mathit{50moves} \Rightarrow_c \mathit{draw}$$

On the other hand it is known that a position with King, Knight and Bishop vs King leads to a victory, if properly played, in less that 35 moves. Hence we have

$$r_2 : \mathit{KNBvsK} \Rightarrow_r \neg \mathit{draw}$$

Moreover since the first rule is an official chess rule it is stronger than the second, thus the superiority relation is $r_1 > r_2$. If player i makes a mistake in the crucial point of the sequence leading to checkmate the opponent king can momentarily escape checkmate gaining precious moves. In this case the mistake made by player i results in 50 consecutive moves without capture and without moving a pawn. In this case we have KNBvsK and can assume $E(\mathit{50moves})$ as a fact. With those premises we derive $E(\mathit{draw})$.

The superiority relation does not play a very relevant role in the derivation of results-in conclusions. It is the usual superiority relation of defeasible logic. Moreover, in general, the superiority relation does not increase the expressive power of defeasible logic. In fact it can be simulated, in a modular and incremental way, in terms of the other components [Antoniou *et al.*, 2001]. Although we do not have a formal results for the present variant of defeasible logic, we are confident that the techniques of [Antoniou *et al.*, 2001] can be applied successfully in the present case.

4 Discussion and Future Work

In this paper we have presented a possible way to combine intensional notions of agency and a (computationally oriented) non-monotonic system. The resulting system seems to provide a sound theoretical and practical framework to reason about actions and the states resulting from them in a non-monotonic setting.

Our aim is just to make a step forward in the development of computational treatments of the notion of institutional agency. In this perspective, our contribution does not include any explicit refinement (e.g., in terms of articulating new axioms) of what has been already proposed in [Governatori *et al.*, 2002a; Gelati *et al.*, 2002]. This does not mean, however, that the model presented here cannot be a potential starting point to achieve new proof-theoretical results. Let us recall that the propositional base of the modal logic of agency is the classical propositional logic [Santos and Carmo, 1996; Elgesem, 1997]. On the other hand, any refinement to introduce non-monotonic reasoning as a crucial aspect of institutional agency has been confined both in [Governatori *et al.*, 2002a; Gelati *et al.*, 2002] and in [Jones and Sergot, 1996; Jones, 2003] to account only for the counts-as link. Although this paper provides a formal machinery to reason about actions only with regard to institutional domains, it proposes some inferential mechanisms that may be generalised to define a non-monotonic theory of agency. How to do this and which is the axiomatisation resulting from such a generalisation is a matter of future research.

The logic presented here is just one of the many logics that can be defined using the main idea of the paper (see Section 3). Non-monotonic reasoning is a complex phenomenon with many facets. Several variants of defeasible logic have been put forward to deal with different (sometimes incompatible)

intuitions behind non-monotonic reasoning. Accordingly a designer of a defeasible logic of agency has to chose the most appropriate defeasible inference mechanism and the degree of provability corresponding to the modalities at hand for the intended application. In a similar way the designer can chose more or less liberal conditions to use counts-as rules to derive brings-it-about literals. For example in this paper we have assumed that we can use a counts-as rule to derive a brings-it-about literal if all the literal in the antecedent of the rule can be derive as results-in. A more liberal condition could just require that only one of them is derived in such a way.

Finally, we suggest some refinements that could enrich the logical framework presented in this paper. In fact, the whole story of institutional agency is not treated in the model presented here.

First of all, the model should deal with the multi-agent dimension of the institutions. In this perspective, it could be extended to capture at least the axiom schema $E_iE_jA \rightarrow \neg E_iA$ (EE$\neg$E) [Santos *et al.*, 1997]. Such a schema corresponds to the idea that the brings-it-about operator expresses actions performed directly and personally, and provides a principle of rationality for modeling co-ordination in institutional organisations: it is counterintuitive that the same agent brings it about that A and brings it about that somebody else achieves A. A first tentative extension to deal with this is in course of preparation [Governatori and Rotolo, 2003].

This extension, however, does not exhaust the matter. Our framework may be articulated to take into account other action concepts. For example, in [Santos *et al.*, 1997; Jones, 2003] the operator G has been also defined to express the idea of indirect successful action. The reading of G_iA is that i ensures that A. G enjoyes the same general properties of E. However, instead of (EE$\neg$E), it is adopted $G_iG_jA \rightarrow G_iA$ (GGG). (GGG) differentiates G from E insofar as the former is meant to represent indirect actions. Moreover, a concept that deserves special attention is that of "practical capability" [Elgesem, 1997; Jones, 2003]. As we alluded to, this should be distinguished from the institutionalised power to achieve a state of affairs. However, such a concept can play an important role in modeling the co-ordination of agents. In fact, the inference of institutional facts may be conditioned by the practical capability of an agent to do things that generate by convention these facts.

Another issue to be investigated concerns the formal treatment of relations between different institutions. These relations are relevant when an action takes place in different institutional contexts and produces diverse, and possibly contradictory, results. Following [Governatori *et al.*, 2002a; Gelati *et al.*, 2002], multi-institutional contexts are captured by stipulating that $A \Rightarrow_s B =_{def} (A \Rrightarrow D_sB) \wedge (D_sA \Rrightarrow D_sB) \wedge (D_sA \Rrightarrow D_{s'}B)$. In the present setting, they could be represented by introducing counts-as rules indexed by different institutions. Of course, the superiority relations would play an important role in settling possible contradictions between different institutional contexts. But that is not all since the matter is concerned with the complex problem of the relation between normative systems [Prakken, 1997].

Finally, one important issue, which we could not address here, is how to deal with conflicting institutional results arising from the exercise of different powers (such conflicts are implicit in certain types of acts, such as when an agent renounces a power). This a crucial question that requires in general to develop a dynamic account of the institutional mechanisms. In this perspective, an important reference is still Goldman's theory of actions generating actions [Goldman, 1970]. It has been pointed out to us that the generation of institutional facts via counts-as rules is quite close to the idea of causality. If so, counts-as relations cannot be reflexive since it may be argued that "it is precisely the property of non-reflexiveness that distinguishes a generation relation as such" [Jones *et al.*, 2003]. This point seems to be problematic in the present framework insofar as defeasible logics are closely related to cumulative reasoning, which includes the property of reflexivity [Billington, 1993]. Also this question will be a matter of future research.

Acknowledgements

We would like to thank an anonymous referee for her helpful criticisms on an earlier version of this paper. This work has been partially supported by the EU IST FET UIE project ALFEBIITE (IST-1999-10298), and this support is gratefully acknowledged. Thanks are due to the partners in this project for providing the context for the current work. However, the authors themselves are solely responsible for any opinion or mistake contained in this document.

References

[Antoniou *et al.*, 2000a] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. A flexible framework for defeasible logics. In *Proc. American National Conference on Artificial Intelligence (AAAI-2000)*, pages 401–405, Menlo Park, CA, 2000. AAAI/MIT Press.

[Antoniou *et al.*, 2000b] Grigoris Antoniou, David Billington, Guido Governatori, Michael J. Maher, and Andrew Rock. A family of defeasible reasoning logics and its implementation. In Werner Horn, editor, *ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence*, pages 459–463, Amsterdam, 2000. IOS Press.

[Antoniou *et al.*, 2001] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Representation results for defeasible logic. *ACM Transactions on Computational Logic*, 2(2):255–287, 2001.

[Artosi *et al.*, 2002] Alberto Artosi, Guido Governatori, and Antonino Rotolo. Labelled tableaux for non-monotonic reasoning: Cumulative consequence relations. *Journal of Logic and Computation*, 12(6):1027–1060, 2002.

[Billington, 1993] David Billington. Defeasible logic is stable. *Journal of Logic and Computation*, 3:370–400, 1993.

[Chellas, 1980] B. Chellas. *Modal Logic. An Introduction*. Cambridge University Press, Cambridge, 1980.

[Conte and Dellarocas, 2001] R. Conte and Ch. Dellarocas. *Social Order in Multiagent Systems*. Kluwer Academic Publishers, Boston, 2001.

[Elgesem, 1997] D. Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2:1–48, 1997.

[Gelati *et al.*, 2002] Jonathan Gelati, Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Declarative power, representation, and mandate: A formal anaysis. In Trevor Bench-Capon, Aspassia Daskalopulu, and Radboudb Winkels, editors, *Legal Knowledge and Information Systems*, pages 41–52. IOS Press, Amsterdam, 2002.

[Goldman, 1970] A. Goldman. *A Theory of Human Action*. Prentice Hall, Princeton, 1970.

[Governatori and Rotolo, 2003] Guido Governatori and Antonino Rotolo. Non-monotonic agency and multi-agent systems. Submitted, 2003.

[Governatori *et al.*, 2002a] Guido Governatori, Jonathan Gelati, Antonino Rotolo, and Giovanni Sartor. Actions, institutions, powers. preliminary notes. In Gabriela Lindemann, Daniel Moldt, Mario Paolucci, and Bin Yu, editors, *International Workshop on Regulated Agent-Based Social Systems: Theories and Applications(RASTA'02)*, pages 131–147. Fachbereich Informatik, Universität Hamburg, 2002.

[Governatori *et al.*, 2002b] Guido Governatori, Vineet N. Padmanabhan, and Abdul Sattar. A defeasible logic of policy-based intention. In Harald Søndergaard, editor, *Australasian Workshop on Computational Logic 2002: Proceedings*, pages 9–20. Department of Computer Science, The University of Melbourne, 2002.

[Jones and Sergot, 1996] A.J.I. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of IGPL*, 3:427–443, 1996.

[Jones *et al.*, 2003] A.J.I. Jones, X. Parent, and A. Stolpe. Private communication, 2003.

[Jones, 2003] Andrew J.I. Jones. A logical framework. In Jeremy Pitt, editor, *Open Agent Societies: Normative Specifications in Multi-Agent Systems*, chapter 3. John Wiley and Sons, Chichester, 2003.

[Kanger, 1972] S. Kanger. Law and logic. *Theoria*, 38:105–32, 1972.

[Lindahl, 1977] L. Lindahl. *Position of change: A Study in law and logic*. Reidel, Dordrecht, 1977.

[Makinson, 1986] D. Makinson. On the formal representation of rights relations. *Journal of Philosophical Logic*, 15:403–25, 1986.

[Nute, 1987] Donald Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 353–395. Oxford University Press, 1987.

[Pörn, 1977] I. Pörn. *Action Theory and Social Science: Some Formal Models*. Reidel, Dordrecht, 1977.

[Prakken, 1997] H. Prakken. *Logical Tools for Modelling Legal Argument*. Kluwer, Dordrecht, 1997.

[Royakkers, 2000] L. Royakkers. Combining deontic and action logics for collective agency. In J. Breuker et al., editor, *Legal Knowledge and Information Systems (Jurix)*. IOS Press, Amsterdam, 2000.

[Santos and Carmo, 1996] F.A.A. Santos and J.M.C.L.M. Carmo. Indirect action. influence and responsibility. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*. Springer, Berlin, 1996.

[Santos *et al.*, 1997] F.A.A. Santos, A.J.I. Jones, and J.M.C.L.M. Carmo. Action concepts for describing organised interaction. In *Thirtieth Annual Hawaii International Conference on System Sciences*. IEEE Computer Society Press, Los Alamitos, 1997.

[Searle, 1995] J.R. Searle. *The Construction of Social Reality*. Penguin Press, Harmondsworth, 1995.

[Segerberg, 1992] K. Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51:347–58, 1992.