The logic of acceptance: grounding institutions on agents' attitudes

Emiliano Lorini, Dominique Longin, Benoit Gaudou, Andreas Herzig

The logic of acceptance: grounding institutions on agents' attitudes

Emiliano Lorini, Dominique Longin, Benoit Gaudou, Andreas Herzig
Institut de Recherche en Informatique de Toulouse (IRIT)
118 Route de Narbonne, F-31062, Toulouse, France
{lorini,longin,gaudou,herzig}@irit.fr

December 1, 2008

Abstract

In the recent years, several formal approaches to the specification of normative multi-agent systems and artificial institutions have been proposed. The aim of this paper is to advance the state of the art in this area by proposing an approach in which a normative multi-agent system is conceived to be autonomous, in the sense that it is able to create, maintain, and eventually change its own institutions by itself, without the intervention of an external designer in this process. In our approach the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances.

In order to meet this objective, we propose the logic $\mathcal{AL}$ (*Acceptance Logic*) in which the acceptance of a proposition by the agents *qua* members of an institution is introduced. Such propositions are true w.r.t. an institutional context and correspond to facts that are instituted in an attitude-dependent way.

The second part of the paper is devoted to the logical characterization of some important notions in the theory of institutions. We provide a formalization of the concept of *constitutive rule*, expressed by a statement of the form "X counts as Y in the context of institution x". Then, we formalize the concepts of obligation and permission (so called *regulative rules*). In our approach constitutive rules and regulative rules of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of the institution.

Keywords

Modal logic, institutions, acceptance, normative systems, multi-agent systems

Contents

1 Introduction

The problem of devising artificial institutions and modeling their dynamics is a fundamental problem in the multi-agent system (MAS) domain [Dignum and Dignum, 2001]. Following [North, 1990, p. 3], artificial institutions can be conceived as "the rules of the game in a society or the humanly devised constraints that structure agents' interaction". Starting from this concept of institution, many researchers working in the field of normative MAS have been interested in developing models which describe the different kinds of rules and norms that agents have to deal with. In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system [Esteva et al., 2001]. More sophisticated models of institutions leave to the agents' autonomy the decision whether to comply or not with the specified rules and norms of the institution [Ågotnes et al., 2007, Lopez y Lopez et al., 2004]. However, all previous models abstract away from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members. More precisely, while it is widely shared in the MAS field that, in order to face complex and dynamical problems, individual agents must be autonomous, less emphasis is devoted to the fact that MASs themselves for exactly the same reasons should be conceived and designed to be autonomous. In fact, etymologically, autonomous means self-binding ('auto' and 'nomos'), and an autonomous MAS should be the vision of an artificial society that is able to create, maintain, and eventually change its own institutions by itself, without the intervention of the external designer in this process.

The aim of this work is to advance the state of the art on artificial institutions and normative multi-agent systems by proposing a logical model in which the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the individual and collective attitudes of the agents which identify themselves as members of the institution. In particular, we propose a model in which an institution is grounded on the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances. On this aspect we agree with [Mantzavinos et al., 2004], when the authors say that (p. 77):

> "only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions." [Emphasis added].

This relationship between acceptance and institutions has been emphasized in the philosophical doctrine of Legal Positivism [Hart, 1992]. According to Hart, the foundations of a normative system or institution consist of adherence to, or acceptance of, an ultimate rule of recognition by which the validity of any rule of the institution may be evaluated. [1]

[1] In Hart's theory, the rule of recognition is the rule which specifies the ultimate criteria of validity in a legal system.

Other authors working in the field of multi-agent systems have advocated the need for a bottom up approach to the explanation of the origin and the evolution of institutions. According to these authors, institutions and their dynamics should be anchored in the agents' attitudes [Conte et al., 1998, Boella and van der Torre, 2007]. For instance, in agreement with Hart's theory, [Conte et al., 1998] have stressed that the existence of a norm in an institution (but also in a group, organization, *etc.*) depends on the recognition and acceptance of the norm by the members of the institution. In their perspective, agents in a multi-agent system contribute to the enforcement and the propagation of the norm in the social context.

The fundamental concept in our paper is that of acceptance *qua* member of an institution. This notion will be informally presented in Section 2. In Section 3 we will introduce a modal logic (called $\mathcal{AL}$ for *Acceptance Logic*) which enables to reason about acceptances of agents and groups of agents. We call the former *individual acceptances*, and the latter *collective acceptances*. In Section 4 we will study the logical properties of the notion of acceptance and its interactions with classical notions such that of individual (private) belief and that of mutual belief. On the basis of the concept of acceptance *qua* member of an institution, we will specify how a group of agents can create and maintain normative and institutional facts which hold only in an attitude-dependent way. That is, it is up to the agents, and not to the external designer, to support such facts (Section 5). Then, we will distinguish regulative components and non-regulative components of an institution [Searle, 1995] (Section 6). On the one hand, we will formalize the concept of *constitutive rule*, that is, the kind of rules accepted by the members of an institution which express classifications between different concepts and establish the relations between "brute" physical facts and institutional facts within the context of the institution (Section 6.1). Since [Searle, 1995, Searle, 1969] and [Jones and Sergot, 1996], these rules have been expressed in terms of assertions of the form "X counts as Y in the context of institution x" (*e.g.* in the institutional context of US, a piece of paper with a certain shape, color, *etc.* counts as a five-dollar bill). On the other hand, *regulative rules* will be formalized through a notion of obligation and a notion of permission by studying a reduction of deontic logic to the logic of acceptance (Section 6.2). Section 7 will be devoted to show how the logic of acceptance $\mathcal{AL}$ can be appropriately refined in order to capture some essential properties of legal institutions in which a special kind of agents called *legislators* are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. Finally, in Section 8, we will compare our proposal with related logical works on institutions and normative systems. Special emphasis will be devoted to the comparison between our approach and the modal logic of normative systems and "counts-as" proposed by Grossi et al. [Grossi et al., 2006]. Proofs of the main theorems presented in the paper are collected in the annex.

2 The concept of acceptance

Some conceptual clarifications of the concept of acceptance *qua* member of an institution are needed because of the crucial role it plays in explaining the maintenance of social institutions.

Several authors have emphasized the difference between acceptance and belief as particular kinds of *individual* attitudes. Whereas private beliefs have been studied for decades [Hintikka, 1962] as representative of doxastic mental states, acceptances have only been examined since [Stalnaker, 1984] and since [Cohen, 1992]. Some authors (*e.g.* [Clarke, 1994]) claim that acceptance implies belief (at least to some minimal degree as argued in [Tollefsen, 2003]). On the contrary, in [Stalnaker, 1984] acceptance is considered to be stronger than belief. Although belief and acceptance seem very close, several authors [Bratman, 1992, Cohen, 1992, Tuomela, 2000] have argued for the importance of keeping the two notions independent. We here agree with this point of view (see Section 4.3).

For the aims of this paper we are particularly interested in a particular feature of acceptance, namely the fact that acceptance is context-dependent (on this point see also [Engel, 1998]). In our approach, this feature is directly encoded in the formal definition of acceptance (see Section 3.1). In fact, one can decide (say for prudential reasons) to reason and act by "accepting" the truth of a proposition in a specific context, and reject the very same proposition in a different context. We will explore the role of acceptance in institutional contexts. Institutional contexts are conceived here as rule-governed social practices on the background of which the agents reason. For example, take the case of a game like Clue. The institutional context is the rule-governed social practice which the agents conform to in order to be competent players. On the background of such contexts, we are interested in the agents' attitudes that can be formally captured. In the context of Clue, for instance, an agent accepts that something has happened *qua* player of Clue. The state of acceptance *qua* member of an institution is the kind of acceptance one is committed to when one is "functioning as a member of the institution" [Tuomela, 2002, Tuomela, 2007]. In these situations it may happen that the agent's acceptances are in conflict with his/her beliefs. For instance, a lawyer who is trying to defend a client in a murder case accepts *qua* lawyer that his/her client is innocent, even she/he believes the contrary.

There exist others differences between belief and acceptance that are not encoded in our formalization of acceptance. According to [Hakli, 2006], the key difference between belief and acceptance is that the former is aimed at truth, whilst the latter depends on an agent's decision. More precisely, while a belief that p is an attitude constitutively aimed at the truth of p, an acceptance is the output of "a decision to treat p as true in one's utterances and actions" without being *necessarily* (see [Tuomela, 2000] for instance) connected to the actual truth of the proposition.

In the present paper the notion of acceptance *qua* member of an institution is also applied to the collective level named *collective acceptance*. The idea of collective attitudes is developed by Searle [Searle, 1995] among others: without supposing the existence of any collective consciousness, he argues that attitudes can be ascribed to a group of agents and that "the forms of collective intentionality cannot (...) be reduced to something else" [Searle, 1995][2].

Collective attitudes such as collective acceptance have been studied in social philosophy in opposition to the traditional notions of *mutual belief* and *mutual knowl-*

[2] A deeper discussion on this point remains out of the scope of this paper. Some interesting arguments for collective intentionality can be found in [Tollefsen, 2002].

edge that are very popular in artificial intelligence and theoretical computer science [Fagin et al., 1995, Lewis, 1969]. It has been stressed that, while mutual belief is strongly linked to individual beliefs and can be reduced to them, collective attitudes such as collective acceptance cannot be reduced to a composition of individual attitudes. This aspect is particularly emphasized by Gilbert [Gilbert, 1987] who follows Durkheim's non-reductionist view of collective attitudes [Durkheim, 1982]. According to Gilbert, any proper group attitude cannot be defined only as a label on a particular configuration of individual attitudes, as mutual belief is. In [Gilbert, 1989, Tuomela, 2007] it is suggested that a collective acceptance of a set of agents C is based on the fact that the agents in C identify themselves as members of a certain group, institution, team, organization, *etc.* and recognize each other as members of the same group, institution, team, organization, *etc.* (this is the view that we adopt in our formalization of acceptance, see Section 3). But mutual belief (and mutual knowledge) does not entail this aspect of mutual recognition and identification with respect to the same social context.

In accordance with [Tuomela, 2002, Tuomela, 2007], in this paper we consider collective acceptance with respect to institutional contexts as an attitude that is held by a set of agents *qua* members of the same institution. A collective acceptance held by a set of agents C *qua* members of a certain institution x is the kind of acceptance the agents in C are committed to when they are "functioning together as members of the institution x", that is, when the agents in C identify and recognize each other as members of the institution x. For example, in the context of the institution Greenpeace agents (collectively) accept that their mission is to protect the Earth *qua* members of Greenpeace. The state of acceptance *qua* members of Greenpeace is the kind of acceptance these agents are committed to when they are functioning together as members of Greenpeace, that is, when they identify and recognize each other as members of Greenpeace.

3 Acceptance logic

The logic $\mathcal{AL}$ (*Acceptance Logic*) enables expressing that some agents identify themselves as members of a certain institution and what (groups of) agents accept while functioning together as members of an institution. The principles of $\mathcal{AL}$ clarify the relationships between individual acceptances (acceptances of individual agents) and collective acceptances (acceptances of groups of agents).

3.1 Syntax

The syntactic primitives of $\mathcal{AL}$ are the following: a finite non-empty set of agents AGT; a countable set of atomic formulas ATM; and a finite set of labels $INST$ denoting institutions. We note $2^{AGT\star} = 2^{AGT} \setminus \{\emptyset\}$ the set of all non-empty subsets of AGT. The language $\mathcal{L}_{\mathcal{AL}}$ of the logic $\mathcal{AL}$ is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathcal{A}_{C:x}\varphi$$

where p ranges over ATM, C ranges over $2^{AGT\star}$ and x ranges over $INST$. We define $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ from $\vee$, $\neg$ and $\bot$ in the usual manner.

The formula $\mathcal{A}_{C:x}\varphi$ reads "the agents in C accept that φ while functioning together as members of the institution x". For notational convenience, we write $i{:}x$ instead of $\{i\}{:}x$.

For example, $\mathcal{A}_{C:Greenpeace}protectEarth$ expresses that the agents in C accept that the mission of Greenpeace is to protect the Earth while functioning together as activists in the context of Greenpeace; and $\mathcal{A}_{i:Catholic}PopeInfallibility$ expresses that agent i accepts that the Pope is infallible while functioning as a member of the Catholic Church.

The formula $\mathcal{A}_{C:x}\bot$ has to be read "agents in C are not functioning together as members of the institution x", because we assume that functioning as a member of an institution is, at least in this minimal sense, a rational activity. Conversely, $\neg\mathcal{A}_{C:x}\bot$ has to be read "agents in C are functioning together as members of the institution x". Thus, $\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi$ stands for "agents in C are functioning together as members of the institution x and they accept that φ while functioning together as members of x" or simply "agents in C accept that φ *qua* members of the institution x". Therefore $\neg\mathcal{A}_{C:x}\varphi$ has to be read "agents in C do not accept that φ be true *qua* members of x".

3.2 $\mathcal{AL}$ frames

We use a standard possible worlds semantics. Let the set of all couples of non-empty subsets of agents and institutional contexts be

$$\Delta = 2^{AGT\star} \times INST.$$

A *frame* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *frame*) is a couple

$$\mathcal{F} = \langle W, \mathscr{A} \rangle$$

where:

- W is a non-empty set of possible worlds;
- $\mathscr{A} : \Delta \rightarrow W \times W$ maps every $C{:}x \in \Delta$ to a relation $\mathscr{A}_{C:x}$ between possible worlds in W.

We note $\mathscr{A}_{C:x}(w) = \{w' : \langle w, w' \rangle \in \mathscr{A}_{C:x}\}$ the set of worlds that the agents in C accept at w while functioning together as members of the institution x.

We impose the following constraints on $\mathcal{AL}$ frames, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

(S.1) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(S.2) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

(S.3) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \subseteq \mathscr{A}_{C:x}(w)$

(S.4) if $w' \in \mathscr{A}_{C:x}(w)$ then $w' \in \bigcup_{i \in C} \mathscr{A}_{i:x}(w')$

(S.5) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \neq \emptyset$

The constraint **S.1** is a generalized version of transitivity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C accept at w' while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.2** is a generalized version of euclideanity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C accept at w while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w' while functioning together as members of the institution x.

The constraint **S.3** is a property of conditional inclusion: given two sets of agents C, B such that $B \subseteq C$, if there exists a world w'' that the agents in C accept at w while functioning together as members of the institution x and w' is a world that the agents in B accept at w while functioning together as members of the institution x then, w' is also a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.4** is a sort of weak reflexivity: if w' is a world that the agents in C accept at w while functioning together as members of the institution x then, there exists some agent $i \in C$ such that w' is a world that agent i accepts at w', while functioning as a member of the institution x.

According to the last constraint **S.5**, given two sets of agents C, B such that $B \subseteq C$, if there exists a world w' that the agents in C accept at w while functioning together as members of the institution x then, there exists a world w'' that the agents in B accept at w while functioning together as members of the institution x.

3.3 $\mathcal{AL}$ models and validity

A *model* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *model*) is a couple

$$\mathcal{M} = \langle \mathcal{F}, \mathscr{V} \rangle$$

where:

- $\mathcal{F}$ is a $\mathcal{AL}$ frame;
- $\mathscr{V} : ATM \rightarrow 2^W$ is valuation function associating a set of possible worlds $\mathscr{V}(p) \subseteq W$ to each atomic formula p of ATM.

Given $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V} \rangle$ and $w \in W$, the couple $\langle \mathcal{M}, w \rangle$ is a *pointed* $\mathcal{AL}$ *model*. Given a formula φ, we write $\mathcal{M}, w \models \varphi$ and say that φ is *true* at world w in $\mathcal{M}$. The notation $\mathcal{M}, w \not\models \varphi$ means that φ is *false* at world w in $\mathcal{M}$. The truth conditions for the formulas of the logic $\mathcal{AL}$ are:

- $\mathcal{M}, w \not\models \bot$;
- $\mathcal{M}, w \models p$ iff $w \in \mathscr{V}(p)$;
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$;

- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in \mathscr{A}_{C:x}(w)$.

A formula φ is *true in a $\mathcal{AL}$ model* $\mathcal{M}$ if and only if $\mathcal{M}, w \models \varphi$ for every world w in $\mathcal{M}$. φ is $\mathcal{AL}$ *valid* (noted $\models_{\mathcal{AL}} \varphi$) if and only if φ is true in all $\mathcal{AL}$ models. φ is $\mathcal{AL}$ *satisfiable* if and only if $\neg\varphi$ is not $\mathcal{AL}$ valid.

3.4 Axiomatization

The axiomatization of $\mathcal{AL}$ is as follows:

(**ProTau**)	All principles of propositional calculus
(**K**)	$\mathcal{A}_{C:x}(\varphi \to \psi) \to (\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{C:x}\psi)$
(**PAccess**)	$\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**NAccess**)	$\neg\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**Inc**)	$(\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \to \mathcal{A}_{B:x}\varphi$ if $B \subseteq C$
(**Unanim**)	$\mathcal{A}_{C:x}(\bigwedge_{i \in C} \mathcal{A}_{i:x}\varphi \to \varphi)$
(**Mon**)	$\neg\mathcal{A}_{C:x}\bot \to \neg\mathcal{A}_{B:x}\bot$ if $B \subseteq C$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$
(**Nec**)	From $\vdash \varphi$ infer $\vdash \mathcal{A}_{C:x}\psi$

This axiomatization includes all tautologies of propositional calculus (**ProTau**) and the rule of inference *modus ponens* (**MP**). Axiom **K** and rule of *necessitation* (**Nec**) define a minimal normal modal logic. (See [Chellas, 1980, chap. 4].)

Axioms **PAccess** and **NAccess** express that a group of agents has always access to what is accepted (resp. not accepted) by its supergroups.

Axiom **PAccess** concerns the (positive) access to what is accepted by a supergroup: when the agents in a set C function together as members of the institution x, then for all $B \subseteq C$ the agents in B have access to all facts that are accepted by the agents in C. That is, if the agents in C accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C accept that φ.

Axiom **NAccess** concerns the (negative) access to what is not accepted by a supergroup: if the agents in C do not accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C do not accept that φ.

Example 1. *Suppose that three agents i, j, k, while functioning together as members of the UK trade union, accept that their mission is to increase teachers' wages, but they do not accept* qua *members of the trade union that their mission is to increase railway workers' wages:*

$\mathcal{A}_{\{i,j,k\}:Union}\, increaseTeacherWage$ *and* $\neg\mathcal{A}_{\{i,j,k\}:Union}\, increaseRailwayWage$.

*By Axiom **PAccess** we infer that, while functioning as a UK citizen, i accepts that i, j, k accept that their mission is to increase teachers' wages, while functioning together as members of the trade union:*

$$\mathcal{A}_{i:UK}\mathcal{A}_{\{i,j,k\}:Union}increaseTeacherWage.$$

*By Axiom **NAccess** we infer that, while functioning as a UK citizen, i accepts that i, j, k do not accept,* qua *members of the trade union, that their mission is to increase railway workers' wages:*

$$\mathcal{A}_{i:UK}\neg\mathcal{A}_{\{i,j,k\}:Union}increaseRailwayWage.$$

Axiom **Inc** says that, if the agents in C accept that φ *qua* members of x then for every subset B of C the agents in B accept φ while functioning together as members of x. This means that the facts accepted by the agents in C *qua* members of a certain institution x are necessarily accepted by the agents in all of C's subsets with respect to the same institution. Therefore Axiom **Inc** describes the *top down* process leading from C's collective acceptance to the individual acceptances of the agents in C.

Example 2. *Imagine three agents i, j, k that,* qua *players of the game Clue, accept that someone called Mrs. Red, has been killed:*

$$\neg\mathcal{A}_{\{i,j,k\}:Clue}\bot \wedge \mathcal{A}_{\{i,j,k\}:Clue}killedMrsRed.$$

*By Axiom **Inc** we infer that also the two agents i, j, while functioning as Clue players, accept that someone called Mrs. Red has been killed:*

$$\mathcal{A}_{\{i,j\}:Clue}killedMrsRed.$$

Axiom **Unanim** expresses a unanimity principle according to which the agents in C, while functioning together as members of x, accept that if each of them individually accepts that φ while functioning as a member of x, then φ is the case. This axiom describes the *bottom up* process leading from the individual acceptances of the members of C to the collective acceptance of the group C.

Finally, Axiom **Mon** expresses an intuitive property of monotonicity about institution membership. It says that, if the agents in C are functioning together as members of the institution x then, for every subset B of C, the agents in B are also functioning together as members of the institution x. As emphasized in Section 2, "the agents in C function together as members of institution x" means for us that "the agents in C identify and recognize each other as members of the same institution x". Thus, Axiom **Mon** can be rephrased as follows: if the agents in a set C identify and recognize each other as members of the institution x then, for every subset B of C, the agents in B also identify and recognize each other as members of x.

The following correspondences (in the sense of correspondence theory, see for instance [van Benthem, 2001, Blackburn et al., 2001]) exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames given in Section 3.2 (see also proof of Theorem 1 in the Annex): Axiom **PAccess** corresponds to the constraint **S.1**, **NAccess** corresponds to **S.2**, **Inc** corresponds to **S.3**, **Unanim** corresponds to **S.4** and **Mon** corresponds to **S.5**.

We call $\mathcal{AL}$ the logic axiomatized by the principles given above: **ProTau**, **K**, **PAccess**, **NAccess**, **Inc**, **Unanim**, **Mon**, **MP**, **Nec**. We write $\vdash_{\mathcal{AL}} \varphi$ if formula φ is a theorem of $\mathcal{AL}$ and $\nvdash_{\mathcal{AL}} \varphi$ if formula φ is not a theorem.

We can prove that $\mathcal{AL}$ is sound and complete with respect to the class of $\mathcal{AL}$ frames.

Theorem 1. *$\vdash_{\mathcal{AL}} \varphi$ if and only if $\models_{\mathcal{AL}} \varphi$.*

By the standard filtration method we can also prove that the logic $\mathcal{AL}$ is decidable.

Theorem 2. *The logic $\mathcal{AL}$ is decidable.*

In the following section the properties of the concepts of individual acceptance, collective acceptance and institution membership will be studied. We will also study the relationships between acceptance and belief in a more formal way than in Section 2.

4 General properties

4.1 Properties of acceptance and institution membership

The following theorem highlights some interesting properties of collective acceptance and institution membership.

Theorem 3. *For every $x, y \in INST$ and $B, C \in 2^{AGT^\star}$ such that $B \subseteq C$:*

(3a) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \neg \mathcal{A}_{C:x} \bot$

(3b) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \bigwedge_{i \in C} \neg \mathcal{A}_{i:x} \bot$

(3c) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \mathcal{A}_{C:x} \varphi \leftrightarrow \mathcal{A}_{C:x} \varphi$

(3d) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \neg \mathcal{A}_{C:x} \varphi \leftrightarrow (\mathcal{A}_{B:y} \bot \vee \neg \mathcal{A}_{C:x} \varphi)$

(3e) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} (\mathcal{A}_{C:x} \varphi \rightarrow \varphi)$

(3f) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C:x} \bigwedge_{i \in C} \mathcal{A}_{i:x} \varphi) \leftrightarrow \mathcal{A}_{C:x} \varphi$

Theorem 3a expresses a property of institution membership. It says that the agents in a group C, while functioning together as members of the institution x, accept that they are functioning together as members of the institution x. Theorem 3b is another way to express the property of institution membership: it expresses that the agents in a group, while functioning together as members of a certain institution, accept that everyone of them is functioning as a member of the institution.

Example 3. *Suppose that, during a concert, the agents in C are functioning together as members of the Philharmonic Orchestra. Then, according to Theorem 3a, this fact is accepted by the group C. That is, while functioning together as members of the Philharmonic Orchestra, the agents in C accept that they are functioning together as members of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \neg \mathcal{A}_{C:Orchestra} \bot$. Moreover, they accept that everyone of them is functioning as a member of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \bigwedge_{i \in C} \neg \mathcal{A}_{i:Orchestra} \bot$.*

Theorem 3c and Theorem 3d together express that a group of agents B can never be wrong in ascribing a collective acceptance to its supergroup C and in recognizing that its supergroup C does not accept something. Furthermore, a group of agents B

has always correct access to what is accepted (resp. not accepted) by its supergroups. The right to left direction of Theorem 3c is Axiom **PAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C accept φ while functioning together as members of institution x then, the agents in C accept φ while functioning together as members of institution x. The right to left direction of Theorem 3d is Axiom **NAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C do not accept φ *qua* members of institution x then, either the agents in B do not function as members of y or the agents in C do not accept φ *qua* members of institution x.

Theorem 3e and Theorem 3f are variants of the unanimity Axiom **Unanim**. Theorem 3e says that for every set of agents C, the agents in C, while functioning together as members of x, accept that if they accept that φ while functioning together as members of x, then φ is the case. Theorem 3f expresses that: if the agents in C, while functioning together as members of x, accept that each of them individually accepts that φ while functioning as a member of x, then the agents in C, while functioning together as members of x, accept that φ is the case.

The following theorem highlights the relationship between the acceptance of a group of agents and the acceptances of its subgroups.

Theorem 4. *For every $x \in INST$ and $C_1, C_2, C_3 \in 2^{AGT}$ such that $C_3 \subseteq C_2 \subseteq C_1$ and $C_3 \neq \emptyset$:*

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C_1:x}(\mathcal{A}_{C_2:x}\varphi \rightarrow \mathcal{A}_{C_3:x}\varphi)$$

Theorem 4 expresses that every group of agents has to accept the principle of inclusion formalized by Axiom **Inc**.

4.2 Discussion around the unanimity principle

Let us consider more in detail the unanimity property of our logic of acceptance expressed by Axiom **Unanim** (and Theorems 3e,3f). This property says that collective acceptances emerge from consensus. This is for us a necessary requirement for a notion of collective acceptance which is valid for all institutions and groups. We did not include stronger principles which explain how a collective acceptance of a group of agents C might be constructed. Nevertheless, one might go further and consider other kinds of principles which are specific to certain institutions and groups.

For example, one might want to extend the analysis to formal (legal) institutions in which special agents with the power to affect the acceptances of the other members of the institution are introduced. In legal institutions, one can formalize the rule according to which all facts that are accepted by the legislators of an institution must be universally accepted by all members of this institution. Suppose that x denotes a legal institution (*e.g.* EU, Association of Symbolic Logic, *etc.*) which has a non-empty set of agents called legislators, noted $Leg(x) \in 2^{AGT^\star}$. (See Section 7 for a precise definition of the function $Leg()$ and a more elaborate analysis of the concepts of legislator and legal institution.) From this, one can formalize a principle stating that everything

that the legislators of the legal institution x accept is universally accepted in the legal institution x:

(**Legislators**) $$\mathcal{A}_{C:x}(\bigwedge_{i\in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Legislators** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case.

Another interesting principle for the construction of collective acceptance is majority. (In this case, unanimity is not required to obtain a consensus.) This kind of principle applies both to informal and formal institutions. The principle of majority could be introduced as a logical axiom for two specific sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$ (*i.e.* B represents the majority of agents in C):

(**Majority**) $$\mathcal{A}_{C:x}(\bigwedge_{i\in B} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Majority** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the majority of them accept that φ, then φ is the case. The following example by Pettit [Pettit, 2001] shows how the majority principle would work.

Example 4. *Imagine a three-member court which has to make a judgment on whether a defendant is liable (noted l) for a breach of contract. The three judges i, j and k accept a majority rule to decide on the issue. That is, i, j and k, while functioning together as members of the court, accept that if the majority of them accepts that the defendant is liable (resp. not liable), then the defendant is liable (resp. not liable). Formally, for any B such that $B \subseteq \{i, j, k\}$ and $|B| = 2$ we have:*

$$\mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i\in B} \mathcal{A}_{i:court}l \rightarrow l) \wedge \mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i\in B} \mathcal{A}_{i:court}\neg l \rightarrow \neg l)$$

Therefore, if the three judges accept that two of them accept that the defendant is liable, i.e. $\mathcal{A}_{\{i,j,k\}:court}(\mathcal{A}_{i:court}l \wedge \mathcal{A}_{j:court}l)$*, by the Principle* ***Majority*** *and Axiom* ***K*** *it follows that the three judges have to accept that the judge is liable,* i.e. $\mathcal{A}_{\{i,j,k\}:court}l$*).*

It has to be noted that the previous principle of majority cannot be generalized to all sets of agents without incurring the following very counterintuitive consequence.

Proposition 1. *If we suppose that the Principle* ***Majority*** *is valid for any B, C such that $B \subseteq C$ and $|C \setminus B| < |B|$ then, the following consequence is derivable, for $i \neq j$:*

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

This means that, when the majority principle is generalized to all sets of agents, we can infer that: if all agents, *qua* members of institution x, accept that two of them accept φ while functioning together as members of institution x then, the acceptances of the two agents propagate to all agents in such a way that all agents accept φ *qua* members of institution x.

4.3 Relationships between acceptance and belief

As said in Section 2, there is a large literature about the distinction between belief and acceptance. For us, belief and acceptance are clearly different concepts in several senses. In this section we focus on the distinction between acceptance, individual belief and mutual belief. Our aim is to provide further clarifications of the concept of acceptance in terms of its relationships with other kinds of agents' attitudes rather than proposing an extension of the logic $\mathcal{AL}$ with individual belief and mutual belief and studying its mathematical properties. Here, we just show how modal operators for belief and mutual belief can be integrated into the logic $\mathcal{AL}$ on the basis of some intuitive interaction principles relating acceptance and belief.

For convenience, we note $Bel_i\varphi$ the formula that reads "the agent i believes that φ is true", and we suppose that belief operators of type Bel_i are defined as usual in a KD45 modal logic [Hintikka, 1962]. Belief operators Bel_i are interpreted in terms of accessibility relations $\mathscr{B}_i$ on the set of possible worlds W. These accessibility relations are supposed to be serial, transitive and euclidean. We write $\mathscr{B}_i(w)$ for the set $\{w' : \langle w, w'\rangle \in \mathscr{B}_i\}$. $\mathscr{B}_i(w)$ is the set of worlds that are possible according to agent i. The truth condition is:

$$\mathcal{M}, w \models Bel_i\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_i(w)$$

Moreover we introduce the notion of mutual belief which has been extensively studied both in the computer science literature [Fagin et al., 1995] and in the philosophical literature [Lewis, 1969]. Given a set of agents $C \subseteq AGT$, $\mathcal{MB}_C\varphi$ reads "there is a mutual belief in C that φ", that is, "everyone in C believes that φ, everyone in C believes that everyone in C believes that φ, everyone in C believes that everyone in C believes that everyone in C believes that φ, and so on". The mutual belief of a set of agents C is interpreted in terms of the transitive closure $\mathscr{B}_C^+$ of the union of the accessibility relations $\mathscr{B}_i$ for every agent $i \in C$, that is:

$$\mathcal{M}, w \models \mathcal{MB}_C\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_C^+(w)$$

Let the concept of "everybody in group C believes φ" be defined as follows:

$$E_C\varphi \stackrel{def}{=} \bigwedge_{i \in C} Bel_i\varphi$$

As shown in [Fagin et al., 1995], the following axioms and rules of inference provide a sound and complete axiomatization of the logic of individual belief and mutual belief:

(**KD45$_{\mathbf{Bel}}$**)	All KD45-principles for the operators Bel_i
(**FixPoint**)	$\vdash \mathcal{MB}_C\varphi \leftrightarrow E_C(\varphi \wedge \mathcal{MB}_C\varphi)$
(**InductionRule**)	From $\vdash \varphi \rightarrow E_C(\varphi \wedge \psi)$ infer $\vdash \varphi \rightarrow \mathcal{MB}_C\psi$

The first interesting thing to note is that, although collective acceptance and mutual belief have different natures (see the discussion in Section 2), they share the Fix Point property. The following Theorem 5 highlights this aspect.

Theorem 5.

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\varphi \leftrightarrow \bigwedge_{i \in C} \mathcal{A}_{i:x}(\varphi \wedge \mathcal{A}_{C:x}\varphi)$$

Nevertheless we cannot argue that our concept of collective acceptance is stronger than the concept of mutual belief, in particular because the **InductionRule** does not hold in $\mathcal{AL}$. This is due to the non-reductionist feature of the collective acceptance: it cannot be reduced to a particular configuration of individual acceptances.

The following two sections are devoted to discuss other interesting relations between acceptance and belief. We will first provide an analysis of the shared aspect of collective acceptance expressed in terms of mutual belief. Then, we will briefly consider the problem of the incompatibility between acceptance and belief.

4.3.1 The shared nature of collective acceptance

As emphasized in the philosophical literature [Gilbert, 1989, Tuomela, 1992], a collective acceptance of the agents in a set C must not be confused with (nor reduced to) the sum of the individual acceptance of the agents in C. On the contrary, when the agents in C accept some fact φ to be true *qua* members of a certain institution, it means that every agent in C declares to the other agents of the group C that she/he is willing to accept φ to be true. This aspect of acceptance can be formally derived by supposing the following two principles relating individual beliefs with collective acceptances.

(**PIntrAccept**)	$\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$	if $i \in C$
(**NegIntrAccept**)	$\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$	if $i \in C$

The first principle says that: if the agents in C accept that φ while functioning together as members of the institution x then, every agent in C believes this. The second principle says that: if the agents in C do not accept φ *qua* members of x then every agent in C believes this.

We can easily prove that, under the previous two principles, collective acceptance is always shared so much that the group C accepts φ if and only if the agents in C mutually believe this. More formally:

Proposition 2. *For any $C{:}x \in \Delta$, the following formulas are derivable from the axiom **D** for belief (following from **KD45**$_{\mathbf{Bel}}$), Axiom **FixPoint** and Rule of inference **InductionRule** for mutual belief, and the interaction Principles **PIntrAccept** and **NegIntrAccept** for acceptance and belief.*

(2a)	$\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$
(2b)	$\neg\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\neg\mathcal{A}_{C:x}\varphi$

According to Proposition 2a, the agents in C accept that φ while functioning together as members of the institution x if and only if there is a mutual belief in C that they accept that φ while functioning together as members of the institution x. According to Proposition 2b, the agents in C do not accept that φ *qua* members of x if and

only if there is a mutual belief in C that they do not accept that φ *qua* members of x. Hence, accepting (resp. not accepting) a proposition while functioning as members of an institution is always a *mutually believed* fact (for the members of the group) which is out in the open and that is used by all the members to reason about each other in the institutional context.

4.3.2 Acceptance and belief might be incompatible

Individual belief and individual acceptance are both private mental attitudes but: an individual belief does not depend on context, whilst an individual acceptance is a context-dependent attitude which is entertained by an agent *qua* member of a given institution. Therefore, an agent can privately disbelieve something she/he accepts while functioning as a member of a given institution. Formally: $Bel_i\varphi \wedge \mathcal{A}_{i:x}\neg\varphi$ may be true. In a similar way, as emphasized in [Tuomela, 1992], a collective acceptance that φ by a group of agents C (*qua* members of a given institution) might be compatible with the fact that none of the agents in C believes that φ (and even that every agent in C believes that $\neg\varphi$). The following example, inspired by [Tuomela, 1992, p. 285], illustrates this point.

Example 5. *At the end of the 80s, the Communist Party of Ruritania accepted that capitalist countries will soon perish (but none of its members really believed so).*

This means that the agents in C accept that capitalist countries will perish ($ccwp$) *qua* members of the Communist Party of Ruritania (CPR) but nobody in C (privately) believes this. Thus, formally: $\neg\mathcal{A}_{C:CPR}\bot \wedge \mathcal{A}_{C:CPR}ccwp \wedge \bigwedge_{i\in C}\neg Bel_i ccwp$.

In the following Section 5 we will show how institutional facts can be grounded on agents' acceptances in such a way that the existence of the former depends on the latter.

5 Truth in an institutional context

Recent theories of institutions [Lagerspetz, 2006, Searle, 1995, Tuomela, 2002] share at least the following two theses.

Performativity: the acceptance that a certain fact is true shared by the members of a certain institution may contribute to the truth of this fact within the context of the institution.

Reflexivity: if a certain fact is true within the context of a certain institution, the acceptance of this fact by the members of the institution is present.

More precisely, a certain fact φ is true within the context of an institution x if and only if the fact φ is accepted to be true by the members of the institution x. Therefore, a *necessary* condition for the existence of a fact within the context of an institution is that this fact is accepted to exist by the members of the institution. Moreover, the acceptance of a certain fact by the members of an institution is a *sufficient* condition for the existence of this fact within the context of the institution.

Example 6. *If the agents,* qua *European citizens, accept a certain piece of paper with a certain shape, color,* etc. *as money, then, within the context of EU, this piece of paper is money (performativity). At the same time, if it is true that a certain piece of paper is money within the context of EU, then the agents* qua *European citizens accept the piece of paper as money (reflexivity).*

Our aim here is to represent in $\mathcal{AL}$ those facts that are true within the context of an institution, that is, to define the concept of truth with respect to an institutional context (*institutional truth*) in a way that respects the previous two principles of reflexivity and performativity. We formalize the notion of institutional truth by means of the operator $[x]$. A formula $[x]\,\varphi$ is read "within the institutional context x, it is the case that φ". We take the latter to be synonymous of "for every set of agents C, the agents in C accept that φ while functioning together as members of the institution x". Formally, for every $x \in INST$:

$$[x]\,\varphi \stackrel{def}{=} \bigwedge_{C \in 2^{AGT\star}} \mathcal{A}_{C:x}\varphi$$

According to our definition, a fact φ is true within the context of institution x if and only if, for every group C, the agents in C accept φ, while functioning together as members of x. Hence the performativity and the reflexivity principles mentioned above are guaranteed.

It is worth noting that this formal definition of truth with respect to an institution is perfectly adequate to characterize informal institutions in which there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. It is a peculiar property of informal institutions the fact that they are based on the general consensus of all their members [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. In Section 7 we will show how the operator $[x]$ can be appropriately redefined in order to characterize formal (legal) institution and to distinguish them from informal institutions. For the moment, we just suppose that our model only applies to the basic informal institutions of a society in which no legislator is given.

It is straightforward to prove that $[x]$ is a normal modal operator satisfying Axiom K and the necessitation rule.

Theorem 6. *For every $x \in INST$:*

(6a) $\qquad \vdash_{\mathcal{AL}} \quad [x]\,(\varphi \to \psi) \to ([x]\,\varphi \to [x]\,\psi)$

(6b) $\qquad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]\,\varphi$

Nevertheless, institutional operators of type $[x]$ fail to satisfy Axiom 4 and Axiom 5. That is, $[x]\,\varphi \wedge \neg\,[x]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \wedge \neg\,[x]\,\neg\,[x]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$ for any $x \in INST$. This means that for every institution x, the members of x might accept φ while they do not accept that they accept φ and, it might be the case that the members of x do not accept φ, while they do not accept that they do not accept φ. The operator $[x]$ does to satisfy these two properties because of the restriction imposed on Axioms **PAccess** and **NAccess** according to which, the agents in a group B have access to all facts accepted (resp. not accepted) by the agents in another group C, *only if* B is

a subgroup of C. Therefore, in the logic $\mathcal{AL}$, a certain fact φ might be accepted by all groups of members of a certain institution x, while some group of members of x does not have access to the fact that all groups of members of x accept φ. (See Section 8.1 for a discussion about a different point of view.)

The following operator $[Univ]$ is defined in order to express facts which are true in all institutions:

$$[Univ]\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]\,\varphi$$

where $[Univ]\,\varphi$ is meant to stand for "φ is universally accepted as true". The operator $[Univ]$ is also a normal modal operator satisfying Axiom K and the necessitation rule:

Theorem 7.

(7a) $\quad \vdash_{\mathcal{AL}} \quad [Univ]\,(\varphi \rightarrow \psi) \rightarrow ([Univ]\,\varphi \rightarrow [Univ]\,\psi)$

(7b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [Univ]\,\varphi$

The operator $[Univ]$ too fails to satisfy Axiom 4 and Axiom 5. Indeed, $[Univ]\,\varphi \wedge \neg\,[Univ]\,[Univ]\,\varphi$ and $\neg\,[Univ]\,\varphi \wedge \neg\,[Univ]\,\neg\,[Univ]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$. This means that: φ might be universally accepted, while it is not universally accepted that φ is universally accepted and; it might be the case that φ is not universally accepted, while it is not universally accepted that φ is not universally accepted.

In the following section operators of institutional truth of type $[x]$ and the operator of universal truth $[Univ]$ will be used to define the concepts of *constitutive rule* and *regulative rule*. These two concepts are indeed fundamental for a theory of institutions.

6 Constitutive rules and regulative rules

According to many philosophers [Rawls, 1955, Alchourrón and Bulygin, 1971] working on social theory and researchers in the field of normative multi-agent systems [Boella and van der Torre, 2004b], institutions are based both on regulative and non-regulative components. In particular, institutions are not only defined in terms of sets of permissions, obligations, and prohibitions (*i.e. norms of conduct* [Bulygin, 1992]) but also in terms of rules which specify and create new forms of behavior and concepts. Several terms such as *constitutive rule* [Searle, 1969, Searle, 1995], *conceptual rule* [Bulygin, 1992] or *determinative rule* [Von Wright, 1963] have been used to identify this non-regulative dimension of institutions. According to Searle for instance "(...) regulative rules regulate antecedently or independently existing forms of behavior (...). But constitutive rules do not merely regulate, they create or define new forms of behavior" [Searle, 1969, p. 33]. In Searle's theory of institutions [Searle, 1969, Searle, 1995], constitutive (*i.e.* non-regulative) rules are expressed by means of "counts-as" statements of the form "X counts as Y in context x" where the context x refers to the institution/normative system in which the rule is specified. As emphasized in [Grossi et al., 2006], "counts-as" statements are used to express classifications and subsumption relations between different concepts, that is, they assert just that a concept X is a subconcept of a concept Y. These classifications are fundamental for establishing the relations between "brute" physical facts and objects on the one

hand, and institutional facts and objects on the other hand (*e.g.* money, private property, *etc.*). For example, in the institutional context of Europe, a piece of paper with a certain shape, color, *etc.* (a physical object) counts as a five-euro bill (an institutional object).

6.1 Constitutive rules

From the concept of institutional truth presented above, a notion of constitutive rule of the form "φ counts as ψ in the institutional context x" can be defined in the logic $\mathcal{AL}$. We conceive a constitutive rule as a material implication of the form $\varphi \rightarrow \psi$ in the scope of an operator $[x]$. Thus, "φ counts as ψ in the institutional context x" only if every group of members of institution x accepts that φ entails ψ. Furthermore, we suppose that a constitutive rule is *intrinsically contextual*, which means that the rule is not universally valid while it is accepted by the members of a certain institution. More precisely, we exclude situations in which $[Univ](\varphi \rightarrow \psi)$ is true (*i.e.* situations in which it is universally accepted that φ entails ψ).

In this perspective, "counts-as" statements with respect to a certain institutional context x do not just express that the members of institution x classify φ as ψ in virtue of their acceptances, but also that this classification is proper to the institution, *i.e.* it is not universally accepted that φ entails ψ. (See [Grossi et al., 2006] for a similar perspective.) In this sense, the notion of "counts-as" presented here is aimed at capturing the proper meaning of the term "constitutive rule", that is, a rule which constitutes something new within the context of an institution.

Thus, for every $x \in INST$ the following abbreviation is given:

$$\varphi \overset{x}{\rhd} \psi \overset{def}{=} [x](\varphi \rightarrow \psi) \wedge \neg [Univ](\varphi \rightarrow \psi)$$

where $\varphi \overset{x}{\rhd} \psi$ stands for "φ counts as ψ in the institutional context x".

Example 7. *Let consider the institutional context of gestural language. There exists a constitutive rule in this language according to which, the nodding gesture counts as an endorsement of what the speaker is suggesting,* i.e. $nodding \overset{gesture}{\rhd} yes$. *This means that every group of speakers using gestural language accepts that making the nodding gesture entails endorsing what the speaker is suggesting,* i.e. $[gesture](nodding \rightarrow yes)$, *and there are members of other institutions (*e.g. *different cultural contexts in which the same gesture does not express the same fact) who do not accept this,* i.e. $\neg [Univ](nodding \rightarrow yes)$.

Note that a stronger version of the concept of constitutive rule could be given by supposing that "φ counts as ψ in the institutional context x" if and only if φ entails ψ within the institutional context x, *i.e.* $[x](\varphi \rightarrow \psi)$, and for every institution y, if $y \neq x$ then it is not the case that φ entails ψ within the institutional context y, *i.e.* $\bigwedge_{y \in INST, y \neq x} \neg [y](\varphi \rightarrow \psi)$. The latter condition implies the condition $\neg [Univ](\varphi \rightarrow \psi)$ in the definition of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. This stronger version of the concept of constitutive rule is not analyzed in the present paper.

The following two theorems highlight some valid and invalid properties of "counts-as" operators of the form $\overset{x}{\rhd}$. Similar properties of "counts-as" have been isolated in [Jones and Sergot, 1996] and [Grossi et al., 2006].

The invalidities 8a-8e show that operators $\overset{x}{\rhd}$ do not satisfy reflexivity (invalidity 8a), transitivity (invalidity 8b), strengthening of the antecedent (invalidity 8c), weakening of the consequent (invalidity 8d) and cautious monotonicity (invalidity 8e).

On the contrary, operators $\overset{x}{\rhd}$ satisfy the properties of right logical equivalence (Theorem 9a), left logical equivalence (Theorem 9b), conjunction of the consequents (Theorem 9c), disjunction of the antecedents (Theorem 9d), cumulative transitivity (Theorem 9e).

Theorem 8.

(8a) $\nvdash_{\mathcal{AL}} \varphi \overset{x}{\rhd} \varphi$

(8b) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(8c) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow ((\varphi_1 \wedge \varphi_3) \overset{x}{\rhd} \varphi_2)$

(8d) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \vee \varphi_3))$

(8e) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)$

Theorem 9. *For every $x \in INST$:*

(9a) *From* $\vdash_{\mathcal{AL}} (\varphi_2 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(9b) *From* $\vdash_{\mathcal{AL}} (\varphi_1 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_3 \overset{x}{\rhd} \varphi_2)$

(9c) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3))$

(9d) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_3 \overset{x}{\rhd} \varphi_2)) \rightarrow ((\varphi_1 \vee \varphi_3) \overset{x}{\rhd} \varphi_2)$

(9e) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

The invalidities 8a-8e are due to the local nature of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. For instance, the fact that $\varphi_1 \overset{x}{\rhd} \varphi_2$ and $\varphi_2 \overset{x}{\rhd} \varphi_3$ are constitutive rules of the institution x does not necessarily entail that $\varphi_1 \overset{x}{\rhd} \varphi_3$ is a constitutive rule of x since it does not necessarily entail $\neg [Univ] (\varphi_1 \rightarrow \varphi_3)$. This is the reason why $\overset{x}{\rhd}$ fails to satisfy transitivity.

Example 8. *In the US state of Texas, "to commit a murder counts as to be punishable by the Death Penalty", and "to be punishable by the Death Penalty counts as to be liable to indictment". As the Death Penalty is not universally accepted in all institutions, both these rules are constitutive rules of Texas,* i.e. $murder \overset{Texas}{\rhd} DeathPenalty$ *and* $DeathPenalty \overset{Texas}{\rhd} indictable$. *From this, it does not follow that it is a constitutive rule of Texas that "to commit a murder counts as to be liable to indictment". Indeed,* $\neg(murder \overset{Texas}{\rhd} indictable)$ *is true. This is due to the fact that "to commit a murder counts as to be liable to indictment" in all countries and institutions, and it is not constitutive of Texas,* i.e. $[Univ] (murder \rightarrow indictable)$.

Similarly, $\overset{x}{\rhd}$ fails to satisfy reflexivity. Indeed, all agents in all possible institutions accept the tautology $\varphi \rightarrow \varphi$ so that "φ counts as φ" cannot be intrinsically contextual

with respect to a certain institution. For similar reasons, strengthening of the antecedent is not a valid property of the operator $\overset{x}{\rhd}$.[3] The following example clarifies this aspect.

Example 9. *It is an accepted custom in the US that a person must leave a tip to the waiter that served him/her at a restaurant. That is, it is a constitutive rule of US that "not leaving a tip to the waiter counts as a violation",* i.e. $\neg leaveTip \overset{US}{\rhd} viol$. *From this, it does not follow that it is a constitutive rule of US that "not leaving a tip to the waiter and not paying the bill counts as a violation". Indeed,* $\neg((\neg leaveTip \wedge \neg payBill) \overset{US}{\rhd} viol)$ *is true. This is because "not leaving a tip and not paying the bill counts as a violation" in all countries and institutions, and it is not constitutive of US,* i.e. $[Univ]((\neg leaveTip \wedge \neg payBill) \rightarrow viol)$.

Discussion

The formal analysis of "counts-as" presented in this section is in agreement with the formal analysis of "counts-as" proposed in [Grossi et al., 2006], where a notion of *proper classificatory rule* is introduced. A proper classificatory rule is represented by the construction $\varphi \Rightarrow_x^{cl+} \psi$ which is meant to stand for "φ counts as ψ in the normative system x". Proper classificatory rules are distinguished by Grossi et al. from (non-proper) *classificatory rules* of type $\varphi \Rightarrow_x^{cl} \psi$. In a way similar to our concept of constitutive rule, *proper classificatory rules* have the specific property of not being universally valid (*i.e.* valid in all institutional contexts). That is, differently from non-proper *classificatory rules*, *proper classificatory rules* are rules which would not hold without the normative system/institution stating them.[4]

Non-proper classificatory rules could be expressed in our logical framework by constructions of the form $[x](\varphi \rightarrow \psi)$, that is, by removing the condition $\neg[Univ](\varphi \rightarrow \psi)$ from the definition of $\varphi \overset{x}{\rhd} \psi$. In agreement with Grossi et al., we would be able to prove that, differently from constitutive rules of the form $\varphi \overset{x}{\rhd} \psi$, such a kind of rules satisfy reflexivity, transitivity, strengthening of the antecedent, weakening of the consequent, and cautious monotonicity. Indeed, the following formulas are all theorems of our logic $\mathcal{AL}$:

Theorem 10.

(10a) $\vdash_{\mathcal{AL}} [x](\varphi \rightarrow \varphi)$

(10b) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow \varphi_3)$

(10c) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x]((\varphi_1 \wedge \varphi_3) \rightarrow \varphi_2)$

(10d) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x](\varphi_1 \rightarrow (\varphi_2 \vee \varphi_3))$

(10e) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_1 \rightarrow \varphi_3)) \rightarrow [x]((\varphi_1 \wedge \varphi_2) \rightarrow \varphi_3)$

In Section 8.1 a more elaborate and detailed analysis of the logic presented in [Grossi et al., 2006] will be provided and its formal relationships with our logic of acceptance will be studied.

[3] Other authors have defended the idea that strengthening of the antecedent and transitivity should not be valid properties of "counts-as" conditionals (*e.g.* [Gelati et al., 2004]).

[4] See [Grossi et al., 2008] for a refinement of this typology of rules.

Another important aspect to be discussed about our formalization of "counts-as" is the problem of contraposition. Indeed, at the present stage, $\varphi \overset{x}{\rhd} \psi$ is logically equivalent to $\neg\psi \overset{x}{\rhd} \neg\varphi$ which can be counterintuitive in some situations. However, the problem of contraposition could be solved by distinguishing in the language of the logic $\mathcal{AL}$ formulas denoting "brute" physical facts from formulas denoting institutional facts and by imposing that the consequent ψ of a "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$ is always a formula denoting an institutional fact. Under this assumption, if the negation of the antecedent in the "counts-as" conditional is not an institutional fact (*i.e.* formula $\neg\varphi$ does not denote an institutional fact), contraposition is not allowed. That is, $\varphi \overset{x}{\rhd} \psi$ does not imply $\neg\psi \overset{x}{\rhd} \neg\varphi$.[5] It is worth noting that, this distinction between formulas denoting "brute" physical facts and formulas denoting institutional facts would enable us to account for an aspect of "counts-as" that our current formalization is not able to capture, namely: the function of "counts-as" statements of establishing the relations between physical facts and objects on the one hand (the antecedent of the "counts-as"), and institutional facts and objects on the other hand (the consequent of the "counts-as"), *e.g.* a certain piece of paper counts as a five-euro bill.

6.2 Regulative rules

Constitutive rules as defined in the previous Section 6.1 are still not sufficient for a characterization of institutional reality. An institution is indeed connected to a deontic dimension that up to now is still missing in our analysis. This deontic dimension consists in several concepts such as obligation, permission, prohibition, *etc.* which are aimed at regulating agents' behaviors and social interactions within the context of the institution.

In order to capture this deontic dimension of institutions, our logic $\mathcal{AL}$ can be appropriately extended by introducing a *violation* atom $viol$ as in Anderson's reduction of deontic logic to alethic logic [Anderson, 1958] and in dynamic deontic logic [Meyer, 1988]. A similar approach has been recently taken in [Grossi, 2008]. By means of the new formal construct $viol$ we can specify the concepts of obligation and that of permission in a way that respects their being also a kind of attitude-dependent facts holding in a specific institutional context.

As far as obligations are concerned, we introduce operators of the form O_x which are used to specify what is obligatory in the context of a certain institution x:

$$O_x\varphi \overset{def}{=} \neg\varphi \overset{x}{\rhd} viol$$

According to this definition, "φ is obligatory within the institutional context x" if and only if "$\neg\varphi$ counts as a violation within the institutional context x".

Example 10. *The formula* $(driveCar \wedge RightSide) \overset{UK}{\rhd} viol$ *which is equivalent to* $O_{UK}(driveCar \rightarrow \neg RightSide)$ *expresses that in the UK it is obligatory to drive on*

[5] See also [Grossi, 2008] for a different solution on how to solve the problem of contraposition in a normal modal logic of "counts-as".

*the left side of the street (*i.e. *"driving a car on the right side of the street counts as violation in UK").*

As the following theorem highlights, our O_x operators satisfy axiom K (Theorem 11a) and do not allow obligations about tautologies (Theorem 11b).

Theorem 11. *For every $x \in INST$:*

(11a) $\vdash_{\mathcal{AL}} \quad O_x(\varphi \rightarrow \psi) \rightarrow (O_x\varphi \rightarrow O_x\psi)$

(11b) $\vdash_{\mathcal{AL}} \neg O_x \top$

On the contrary, obligation operators do not satisfy the necessitation rule. This is due to the negative condition $\neg\,[Univ]\,(\neg\varphi \rightarrow viol)$ in the definition of $O_x\varphi$. Indeed, in order to have a normal modal operator for obligation, it is sufficient to remove the negative condition $\neg\,[Univ]\,(\varphi \rightarrow \psi)$ from the definition of the "counts-as" conditional $\varphi \overset{x}{\triangleright} \psi$ given in Section 6.1. The following theorem highlights other interesting invalidities of the obligation operators O_x.

Theorem 12.

(12a) $\nvdash_{\mathcal{AL}} \neg O_x \bot$

(12b) $\nvdash_{\mathcal{AL}} O_x\varphi \rightarrow O_x(\varphi \vee \psi)$

(12c) $\nvdash_{\mathcal{AL}} O_x(\varphi \wedge \psi) \rightarrow O_x\varphi$

According to the invalidity 12a, obligation operators do not satisfy the axiom D of Standard Deontic Logic (SDL) [Åqvist, 2002]. For instance, in the logic $\mathcal{AL}$ institutions might be empty, that is, for every $C \in 2^{AGT\star}$, $\mathcal{A}_{C:x}\bot$. If institution x is empty, it does not have any obligation (*i.e.* $O_x\bot$). According to the other two invalidities we have that: if φ is obligatory within the context of institution x then, it is not necessarily the case that φ or ψ is obligatory within the context of the same institution (invalidity 12b) and if φ and ψ are obligatory within the context of institution x then, it is not necessarily the case that φ is obligatory within the context of the same institution (invalidity 12c). Thus, our obligation operators O_x do not incur two classical problems of Standard Deontic Logic which are commonly referred to as "Ross paradox" and "Good Samaritan paradox" [Carmo and Jones, 2002]. On the one hand, it seems rather odd to say that the *obligation to mail a certain letter* entails an *obligation to mail the letter or to burn it* which can be fulfilled simply by burning the letter (something presumably forbidden) ("Ross paradox"). On the other hand, it seems rather odd to say that if *it is obligatory that Mary helps John who has had an accident*, then *it is obligatory that John has an accident* ("Good Samaritan paradox"). Here we do not consider other well-known paradoxes of deontic logic (such as Chisholm paradox for instance) which require an elaborate and detailed analysis of contrary-to-duty obligations and defeasible conditional obligations (on this see [Prakken and Sergot, 1997, Makinson, 1993, Hansen et al., 2007] for instance). Indeed, this issue goes beyond the objectives of the present work.

As far as permissions are concerned we say that "φ is permitted within the institutional context x" (noted $P_x\varphi$) if and only if $\neg\varphi$ is not obligatory within the institutional

context x. Formally:

$$P_x\varphi \stackrel{def}{=} \neg O_x \neg \varphi$$

That is, we define the permission operator in the standard way as the dual of the obligation operator.[6]

Before concluding this section, it is important to stress again that in our approach regulative rules of type $O_x\varphi$ and $P_x\varphi$ as well as constitutive rules of type $\varphi \overset{x}{\triangleright} \psi$ of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of a certain institution.

7 Towards legal institutions

In Section 5 we have supposed that φ is true within the context of institution x if and only if all members of this institution accept φ to be true. At this point, it might be objected that there are facts which are true in an institutional context but only "special" members of the institution are aware of them. For instance, there are laws in every country that are known only by the specialists of the domain (lawyers, judges, members of the Parliament, *etc.*). Aren't these facts true notwithstanding that many members of the institution are not aware of them?

In order to resist to this objection recall that until now our model applied to the basic informal institutions of a society, that is, *rule-governed social practices* [Tuomela, 2002] in which no member with "special" powers is introduced.

It is a peculiar property of informal institutions to be based on general consensus [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. Relative to this restriction, the assumption made in Section 5 is justified because, with respect to informal institutions, there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. For instance, in the informal institution of common language, nobody has the power to change the rules for promising. (See [Searle, 1969] for more details.) On the contrary, it is a specificity of legal (formal) institutions to have such specialized agents with special powers to interpret and modify the institution itself. This distinction between informal and formal (legal) institutions has been stressed by many authors working in the field of social and legal theory [Castelfranchi, 2003, North, 1990, Lorini and Longin, 2008, Von Wright, 1963]. Consider for instance the following quotation from Von Wright where the terms *prescription* and *custom* respectively correspond to the terms *formal institution* and *informal institution* used here: "(...) Prescriptions are *given* or *issued* by someone. They 'flow' from or have their 'source' in the will of norm-giver (...) Customs, first of all, are not *given* by any authority to subjects. If we can speak of an authority behind the customs at all this authority would be the community itself" [Von Wright, 1963, p. 7–9].

[6] We do not consider here the classical distinction between *weak permission* and *strong permission* [Alchourrón and Bulygin, 1971, Raz, 1975, Von Wright, 1963]. According to legal theory, a weak permission corresponds to the absence in a normative system of a norm prohibiting φ (this is represented by our permission operator P_x). A strong permission corresponds to the existence in the normative system of an explicit norm, issued by the legislators, according to which φ is permitted. For a logical analysis of the distinction between weak and strong permission see our related work [Lorini and Longin, 2008].

In the rest of this section we will show how the logic $\mathcal{AL}$ can be appropriately refined in order to move beyond informal institutions and to capture some essential properties of formal (legal) institutions in which legislators are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. For the sake of simplicity and readability of the article, these principles will not be included in the axiomatization of the logic $\mathcal{AL}$ and their semantic counterparts will not be studied.

In order to distinguish formal from informal institutions, we introduce a total function Leg which assigns a (possibly empty) set of agents to every institution x:

$$Leg : INST \longrightarrow 2^{AGT}$$

$Leg(x)$ denotes the set of legislators of institution x, that is, the set of agents legally responsible over institution x and which are entitled to modify its structure. The function Leg allows distinguishing formal from informal institutions in a simple way. It is indeed reasonable to suppose that informal institutions are those institutions that do not have legislators, that is, x is an informal institution if and only if $Leg(x) = \emptyset$. On the contrary, if $Leg(x) \neq \emptyset$, x is a legal or formal institution. In this sense, the cardinality of $Leg(x)$ provides an important property: it allows us to distinguish between legal institutions and informal institutions.

It seems reasonable to suppose that the legislators of a certain legal institution x must function together as members of institution x. This assumption is expressed by the following principle. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$\neg \mathcal{A}_{Leg(x):x} \bot$$

As emphasized in Section 5, legislators are "special" agents who have the power to affect the acceptances of the other members of the institution. In legal institutions, all facts that are accepted by the legislators must be universally accepted by all members of the institution. In this perspective, legal institutions are characterized by the following principle which explains how the collective acceptance of a set C of members of institution x is affected by the acceptance of the legislators of the institution. For every $C \in 2^{AGT\star}$ and $x \in INST$ such that $Leg(x) \neq \emptyset$:

(Legislators) $$\mathcal{A}_{C:x}(\bigwedge_{i \in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

According to **Legislators**, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case. As emphasized in Section 4.1, the Principle **Legislators** can be conceived as an additional specification of how collective acceptances of groups of agents are built within the context of an institution. It is worth noting that **Legislators** is perfectly compatible with the general principle of unanimity of the logic $\mathcal{AL}$ described by Axiom **Unanim** (and the related Theorems 3e, 3f). Indeed, we can reasonably suppose that the members of an institution might accept certain things on the basis of a criterion of unanimity and, at the same time, accept what the legislators accept and

decide.[7]

We conclude by showing how the concept of institutional truth proposed in Section 5 can be appropriately refined in order to deal with legal institutions. Differently from informal institutions, legal institutions do not necessarily depend on the general consensus of all their members. More precisely, if a certain fact φ is true within the context of the legal institution x then, it is not necessarily the case that for every set of agents C, the agents in C accept φ while functioning together as members of the legal institution x. In a legal institution it is sufficient that the legislators accept φ to be true to make it true for the institution. This means that the notion of institutional truth for legal institutions should be defined as follows. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$[x]^L \varphi \stackrel{def}{=} \mathcal{A}_{Leg(x):x}\varphi$$

This means that "within the context of the legal institution x it is the case that φ" if and only if "the legislators of institution x accept that φ".

From the principles of $\mathcal{AL}$ and the definition of the function $Leg()$, it follows that the operators $[x]^L$ are also normal. Moreover, differently from the $[x]$ operators, which adequately characterize the notion of institutional truth for informal institutions, $[x]^L$ operators satisfy axioms 4 and 5 of modal logic, that is: if the legislators of institution x accept φ then, they accept that they accept φ (Theorem 13c); if the legislators of an institution x do not accept φ, then they accept that they do not accept φ (Theorem 13d).[8]

Theorem 13. *For every $x \in INST$:*

(13a) $\quad \vdash_{\mathcal{AL}} \quad [x]^L (\varphi \rightarrow \psi) \rightarrow ([x]^L \varphi \rightarrow [x]^L \psi)$

(13b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]^L \varphi$

(13c) $\quad \vdash_{\mathcal{AL}} \quad [x]^L \varphi \rightarrow [x]^L [x]^L \varphi$

(13d) $\quad \vdash_{\mathcal{AL}} \quad \neg [x]^L \varphi \rightarrow [x]^L \neg [x]^L \varphi$

It is worth noting that the analysis of constitutive rules and regulative rules proposed in Sections 6.1 and 6.2 could be refined in the light of this distinction between informal and legal institutions. In particular, a new form of "counts-as" and two related concepts of obligation and permission could be defined in terms of the previous operator $[x]^L$. This is in order to characterize a notion of constitutive rule and a notion of regulative rule which apply straightforwardly to the context of legal institutions, and which go beyond the notions of constitutive rule and regulative rule for informal institutions given in Sections 6.1 and 6.2 and based on the operator $[x]$. We postpone this kind of analysis to future works.

[7]Note that a further principle which seems reasonable for legal institutions is a majority principle for legislators: the legislators of a certain legal institution x accept that if the majority of them accept φ, then φ is true. This should be conceived as a particular case of the majority principle discussed in Section 4.1. Formally, for any $x \in INST$ such that $Leg(x) \neq \emptyset$, if $B \subseteq Leg(x)$ and $|Leg(x) \setminus B| < |B|$ (*i.e.* B represents the majority of the legislators of the institution x) then: $\mathcal{A}_{Leg(x):x}(\bigwedge_{i \in B} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$.

[8]Note that the operator $[x]$ is stronger than the operator $[x]^L$, that is, $[x]\,\varphi$ implies $[x]^L\,\varphi$.

8 Comparison with other logical approaches to normative systems

In the following two sections our logic $\mathcal{AL}$ will be compared with two approaches to normative systems and institutions which have been recently proposed in the multi-agent system domain.

8.1 Embedding Grossi et al.'s logic of "counts-as" into $\mathcal{AL}$

Because of the interesting formal similarities, we will first compare $\mathcal{AL}$ with the modal logic of normative systems proposed in [Grossi et al., 2006], henceforth abbreviated $\mathcal{GMD}$ logic.

In the $\mathcal{GMD}$ logic a set of contexts CXT denoting normative systems is introduced. $\mathcal{GMD}$ logic is based on a set of modal operators $[\![x]\!]$ (one for every context x in CXT). Operators $[\![x]\!]$ are similar to our operators $[x]$ defined in Section 5.[9] A formula $[\![x]\!]\,\varphi$ approximately stands for "in the institutional context/normative system x it is the case that φ". It is supposed that CXT contains a special context $Univ$, where the operator $[\![Univ]\!]$ is used for denoting facts which universally hold. We note $CXT_0 = CXT \setminus \{Univ\}$. The language of the $\mathcal{GMD}$ logic is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid [\![x]\!]\,\varphi$$

where p ranges over ATM and x ranges over CXT. $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ are defined from $\vee$, $\neg$ and $\bot$ in the usual manner.

As noted in Section 6.1, operators $[\![x]\!]$ and $[\![Univ]\!]$ are exploited in Grossi et al.'s logic to define contextual conditionals called *proper classificatory rules*, noted $\varphi \Rightarrow_x^{cl+} \psi$, which are an abbreviation of $[\![x]\!]\,(\varphi \rightarrow \psi) \wedge \neg\,[\![Univ]\!]\,(\varphi \rightarrow \psi)$ and which read "φ counts as ψ in the normative system x". The construction $\varphi \Rightarrow_x^{cl+} \psi$ is similar to our $\varphi \overset{x}{\rhd} \psi$.

The most striking difference between our logic of acceptance $\mathcal{AL}$ and the $\mathcal{GMD}$ logic is that in the logic $\mathcal{AL}$ the contextual operators $[x]$ are built on the notion of collective acceptance, whereas in the $\mathcal{GMD}$ logic the contextual operators $[\![x]\!]$ are given as primitive operators.

Frames of the $\mathcal{GMD}$ logic are called *multi-context frames*. A multi-context frame has the following form:

$$\mathcal{F}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0} \rangle$$

where:

- S is a set of possible worlds;
- $\{S_x\}_{x \in CXT_0}$ is a family of subsets of S, one for every institutional context $x \in CXT_0$.

A *multi-context model* is a tuple

$$\mathcal{M}^{\mathcal{GMD}} = \langle \mathcal{F}^{\mathcal{GMD}}, \pi \rangle$$

where:

[9] Here we use the notation $[\![x]\!]$ in order to distinguish their operators from ours.

- $\mathcal{F}^{\mathcal{GMD}}$ is a multi-context frame;
- $\pi : ATM \rightarrow 2^S$ is a valuation function associating a set of possible worlds $\pi(p) \subseteq S$ to each atomic formula p of ATM.

The truth conditions for formulas of the $\mathcal{GMD}$ logic are just standard for contradiction, atomic propositions, negation and disjunction. The following are the truth conditions for $[\![x]\!]\,\varphi$ and $[\![Univ]\!]\,\varphi$.

- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![x]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S_x$;
- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![Univ]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S$.

A formula φ is *true in a $\mathcal{GMD}$ model* $\mathcal{M}^{\mathcal{GMD}}$ iff $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ for every world w in $\mathcal{M}^{\mathcal{GMD}}$. φ is $\mathcal{GMD}$ *valid* (noted $\models_{\mathcal{GMD}} \varphi$) if and only if φ is true in all $\mathcal{GMD}$ models. φ is $\mathcal{GMD}$ *satisfiable* iff $\neg\varphi$ is not $\mathcal{GMD}$ valid.

The $\mathcal{GMD}$ logic is axiomatized by the following principles, where x and y denote elements of the set CXT_0 :

(**ProTau**)	All principles of propositional calculus
($\mathbf{K}_{[\![x]\!]}$)	$[\![x]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![x]\!]\,\varphi \rightarrow [\![x]\!]\,\psi)$
($\mathbf{K}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\psi)$
($\mathbf{4}_{[\![x]\!],[\![y]\!]}$)	$[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,[\![x]\!]\,\varphi$
($\mathbf{5}_{[\![x]\!],[\![y]\!]}$)	$\neg\,[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,\neg\,[\![x]\!]\,\varphi$
($\mathbf{4}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,[\![Univ]\!]\,\varphi$
($\mathbf{5}_{[\![Univ]\!]}$)	$\neg\,[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\neg\,[\![Univ]\!]\,\varphi$
($\mathbf{T}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow \varphi$
($\subseteq_{[\![Univ]\!],[\![x]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![x]\!]\,\varphi$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$
($\mathbf{Nec}_{[\![x]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![x]\!]\,\psi$
($\mathbf{Nec}_{[\![Univ]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![Univ]\!]\,\psi$

We write $\vdash_{\mathcal{GMD}} \varphi$ if formula φ is a theorem of $\mathcal{GMD}$.

Axiom $\mathbf{K}_{[\![x]\!]}$ and Rule $\mathbf{Nec}_{[\![x]\!]}$ express that the operators $[\![x]\!]$ are normal modal operators. Axioms $\mathbf{K}_{[\![Univ]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$, $\mathbf{T}_{[\![Univ]\!]}$ and the rule of inference $\mathbf{Nec}_{[\![Univ]\!]}$ express that the universal modality $[\![Univ]\!]$ is defined in the modal logic system S5. According to the Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, truth and falsehood in institutional contexts/normative systems are absolute because they remain invariant even if they are evaluated from another institutional context/normative system. This means that every normative system y has full access to all facts which are true in a different normative system x. In our view, these two principles are criticizable because they rely on a strong assumption of perfect information, *i.e.* a normative system has perfect information about the facts that are true in the other normative systems. Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ expresses the relationship between the universal modality and the contextual modalities.

In [Grossi, 2007] it is proved that the $\mathcal{GMD}$ logic is sound and complete with respect to the class of $\mathcal{GMD}$ frames.

It is easy to show that the principles of the acceptance logic $\mathcal{AL}$ given in Section 3 are not sufficient to derive the principles of the $\mathcal{GMD}$ logic. In particular, Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$, $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$ and $\mathbf{T}_{[\![Univ]\!]}$ are not derivable in $\mathcal{AL}$.

In order to embed $\mathcal{GMD}$ we need to slightly modify the properties of the logic $\mathcal{AL}$. On the one hand, we need to generalize Axioms **PAccess** and **NAccess** by supposing that they **also** hold for the case $B \not\subseteq C$. This is in order to infer the formulas $[x]\,\varphi \rightarrow [y]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \rightarrow [y]\,\neg\,[x]\,\varphi$ in the augmented logic $\mathcal{AL}$. Thus, we need to assume that, given two arbitrary sets of agents B and C, the agents in B have access to all facts that the agents in C accept (do not accept), while functioning together as members of a certain institution x. On the other hand, we need to add the principle $[Univ]\,\varphi \rightarrow \varphi$ to the logic $\mathcal{AL}$. The way to embed the $\mathcal{GMD}$ logic into our logic $\mathcal{AL}$ is illustrated in the following paragraph.

An embedding of $\mathcal{GMD}$ logic. Let us slightly modify the logic of acceptance $\mathcal{AL}$ in order to provide a correct embedding of $\mathcal{GMD}$. We call $\mathcal{AL}^+$ the modified logic of acceptance.

$\mathcal{AL}^+$ has the same language as $\mathcal{AL}$ (see Section 3.1). $\mathcal{AL}^+$ frames are tuples $\mathcal{F} = \langle W, \mathscr{A} \rangle$ where W and $\mathscr{A}$ are defined as for $\mathcal{AL}$ frames, except that the constraints **S.1** and **S.2** given in Section 3.2 are supposed to hold also for the case $B \not\subseteq C$ and the following additional constraint **S.6** is imposed. That is, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT\star}$ we suppose:

(**S.1'**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(**S.2'**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

Furthermore, for any world $w \in W$ we suppose:

(**S.6**) $\exists C \in 2^{AGT\star}$, $\exists x \in INST$ such that $w \in \mathscr{A}_{C:x}(w)$

The axiomatization of $\mathcal{AL}^+$ is given by the axiom schemes and rules of inference of $\mathcal{AL}$, except that an Axiom corresponding to the Axiom $\mathbf{T}_{[\![Univ]\!]}$ of the $\mathcal{GMD}$ logic is added, and the Axioms **PAccess** and **NAccess** of the logic $\mathcal{AL}$ are generalized in such a way that they also for hold for the case $B \not\subseteq C$. That is, for any sets of agents $C, B \in 2^{AGT\star}$, we suppose:

(**PAccess**$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$

(**NAccess**$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$

Furthermore, we suppose:

($\mathbf{T}_{Univ}$) $[Univ]\,\varphi \rightarrow \varphi$

Axioms **PAccess**$^+$ and **NAccess**$^+$ respectively correspond to the semantic constraints **S.1'** and **S.2'**, whilst Axiom $\mathbf{T}_{Univ}$ corresponds to the semantic constraint **S.6**.

The definitions of validity and satisfiability in $\mathcal{AL}^+$ are given accordingly. We write $\models_{\mathcal{AL}^+} \varphi$ if formula φ is *valid* in all $\mathcal{AL}^+$ models satisfying the semantic constraints **S.3**, **S.4**, **S.5** given in Section 3.2 and the constraints **S.1'**, **S.2'**, **S.6** given here. We call $\mathcal{AL}^+$ the logic axiomatized by Axiom $\mathbf{T}_{Univ}$ and the principles of the logic $\mathcal{AL}$ (Section 3.4), where Axioms **PAccess** and **NAccess** are generalized to **PAccess**$^+$ and **NAccess**$^+$. We write $\vdash_{\mathcal{AL}^+} \varphi$ if formula φ is a theorem of $\mathcal{AL}^+$.

We can prove that $\mathcal{AL}^+$ as well is sound and complete. More precisely:

Theorem 14. *$\vdash_{\mathcal{AL}^+} \varphi$ if and only if $\models_{\mathcal{AL}^+} \varphi$.*

Consider the following translation tr from $\mathcal{GMD}$ to the new logic $\mathcal{AL}^+$:

- $tr(\bot) = \bot$
- $tr(p) = p$
- $tr(\neg\varphi) = \neg tr(\varphi)$
- $tr(\varphi \vee \psi) = tr(\varphi) \vee tr(\psi)$
- $tr([\![x]\!]\,\varphi) = [x]\,tr(\varphi)$
- $tr([\![Univ]\!]\,\varphi) = [Univ]\,tr(\varphi)$,

As the following Theorem 15 shows, tr is a correct embedding of the $\mathcal{GMD}$ logic.

Theorem 15. *Let $INST = CXT$ and φ be a formula of the $\mathcal{GMD}$ logic. Then, φ is $\mathcal{GMD}$ satisfiable if and only if $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

REMARK. It is worth noting that $\mathcal{GMD}$ logic can also be embedded into the variant of $\mathcal{AL}$ with legislators presented in Section 7 by the translations $tr([\![x]\!]\,\varphi) = [x]^L\,\varphi$ and $tr([\![Univ]\!]\,\varphi) = [Univ]^L\,\varphi$, after defining

$$[Univ]^L\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]^L\,\varphi.$$

To obtain a correct embedding of the $\mathcal{GMD}$ logic, it is sufficient to add to $\mathcal{AL}$ the three axioms $[x]^L\,\varphi \rightarrow [y]^L\,[x]^L\,\varphi$, $\neg\,[x]^L\,\varphi \rightarrow [y]^L\,\neg\,[x]^L\,\varphi$ and $[Univ]^L\,\varphi \rightarrow \varphi$ and the two corresponding semantic constraints over $\mathcal{AL}$ frames:

$$\text{if } w' \in \mathscr{A}_{Leg(y):y}(w) \text{ then } \mathscr{A}_{Leg(x):x}(w') = \mathscr{A}_{Leg(x):x}(w), \text{ and}$$
$$\exists x \in INST \text{ such that } w \in \mathscr{A}_{Leg(x):x}(w).$$

8.2 A conceptual comparison with Boella & van der Torre's model

The formal approach to institutions and normative systems proposed by Boella & van der Torre [Boella and van der Torre, 2004a, 2004b, 2007] is similar in some respect to ours. Here we just provide a *conceptual* comparison between the two approaches. We are not able to provide a more *technical* comparison. Indeed, our formalism based on modal logic and their formalism based on input-output logic [Makinson and van der Torre, 2000] are too different to be compared in the fashion followed in Section 8.1.

Boella & van der Torre emphasize the relevance of the concept of acceptance for a formal model of institutions. In their model, individual agents *accept* a norm, together

with its associated sanctions and rewards, when they recognize that this norm serves to achieve their desires and believe that the other agents will conform to it. According to them, for a norm to be really effective it must be respected due to its acceptance, and not only to the fear of sanctions. Although they take the concept of acceptance into consideration, they do not analyze it in detail. In particular, in their model there is no distinction between individual acceptance and collective acceptance. On the contrary, this distinction is fundamental in our $\mathcal{AL}$ logic in which we clarify the relationships between individual acceptances and collective acceptances and we provide an explanation of how the collective acceptance of a group of agents C is built from the individual acceptances of the agents in C.

Moreover, in Boella & van der Torre's approach, normative systems and institutions are conceived as agents and mental attitudes such as beliefs and goals are ascribed to them. Differently from them, we do not claim that institutions can be conceived as agents. In our approach, we only defend the idea that the institutional reality is built on the top of the agents' attitudes. In particular, we claim that institutions are grounded on the individual and collective *acceptances* of their members and groups of members, and their dynamics depend on the dynamics of these acceptances.

9 Conclusion

We have presented in this article a logic of acceptance and applied it to the analysis of institutions. Our logic of acceptance allows to express that agents accept something to be true *qua* members of a certain institution. Given the properties of this demystified notion of acceptance, we have provided an analysis of the kind of attitude-dependent facts which are typical of institutions. We have formalized the concept of constitutive rule expressed by statements of the form "X counts as Y in the context of institution x". Then, we have introduced a notion of obligation and a notion permission with respect to an institutional context (*i.e.* so-called regulative rules). While constitutive rules and regulative rules are usually defined from the external perspective of a normative system or institution, in the present work we have anchored these rules in the agents' acceptances.

Directions for future research are manifold. For instance, future works will be devoted to integrate modalities expressing agents' goals and preferences, such as the ones provided in [Cohen and Levesque, 1990], into the logical framework presented in this paper. This is in order to investigate the decision to join (resp. not to join) a given institution and the related decision to accept (resp. not to accept) the norms of the institution with its associated sanctions and rewards. These kinds of decisions are indeed influenced by the inconsistency between the agent's goals and the current norms and rules of the institution. For instance, if the agent's goals conflict with the norms proclaimed by the legislators then, the agent will probably decide not to join the institution.

Another interesting topic to be investigated in future works is the dynamics of individual and collective acceptances in institutional contexts. We have already started to study this topic in a recent work [Herzig et al., 2008]. The idea is to extend the logic of acceptance $\mathcal{AL}$ by events of type $x!\varphi$ and corresponding dynamic operators

of the form $[x!\varphi]$. A formula $[x!\varphi]\psi$, means that ψ is true after every announcement of formula φ in the context of institution x. Operators of type $[x!\varphi]$, which are similar to the operators of announcements in dynamic epistemic logic [Baltag et al., 1998, Gerbrandy and Groeneveld, 1997, van Ditmarsch et al., 2007], express that the members of an institution x learn that φ is true in that institution in such a way that their acceptances, *qua* members of institution x, are updated. Such operators can also be used to describe how the acceptances of the members of institution x change, after that a certain norm (e.g. obligation, permission) is *issued* or *promulgated* within the context of this institution.

References

[Ågotnes et al., 2007] Ågotnes, T., van der Hoek, W., Rodriguez-Aguilar, J., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1181–1186. AAAI Press.

[Alchourrón and Bulygin, 1971] Alchourrón, C. and Bulygin, E. (1971). *Normative systems*. Springer, New York.

[Anderson, 1958] Anderson, A. (1958). A reduction of deontic logic to alethic modal logic. *Mind*, 22:100–103.

[Åqvist, 2002] Åqvist, L. (2002). Deontic Logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, 2nd edition.

[Baltag et al., 1998] Baltag, A., Moss, L., and Solecki, S. (1998). The logic of public announcements, common knowledge and private suspicions. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK'98)*, pages 43–56, San Francisco, CA. Morgan Kaufmann Publishers Inc.

[van Benthem, 2001] van Benthem, J. (2001). Correspondence theory. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Kluwer Academic Publishers, 2nd edition.

[Blackburn et al., 2001] Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge.

[Boella and van der Torre, 2004b] Boella, G. and van der Torre, L. (2004b). Regulative and constitutive norms in normative multiagent systems. In *Proceedings of the 9th International Conference on Principles on Principles of Knowledge Representation and Reasoning(KR 2004)*, pages 255–266. AAAI Press.

[Boella and van der Torre, 2007] Boella, G. and van der Torre, L. (2007). Norm negotiation in multiagent systems. *International Journal of Cooperative Information Systems*, 16(1):97–122.

[Bratman, 1992] Bratman, M. E. (1992). Practical reasoning and acceptance in context. *Mind*, 101(401):1–15.

[Bulygin, 1992] Bulygin, E. (1992). On norms of competence. *Law and Philosophy*, 11(3):201–216.

[Carmo and Jones, 2002] Carmo, J. and Jones, A. (2002). Deontic logic and contrary-to-duties. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 265–343. Kluwer Academic Publishers, 2nd edition.

[Castelfranchi, 2003] Castelfranchi, C. (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic*, 1(1-2):47–92.

[Chellas, 1980] Chellas, B. F. (1980). *Modal Logic: an Introduction*. Cambridge University Press, Cambridge.

[Clarke, 1994] Clarke, D. (1994). Does acceptance entail belief? *American Philosophical Quarterly*, 31(2):145–155.

[Cohen, 1992] Cohen, L. J. (1992). *An essay on belief and acceptance*. Oxford University Press, New York, USA.

[Cohen and Levesque, 1990] Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42:213–261.

[Coleman, 1990] Coleman, J. (1990). *Foundations of Social Theory*. Harvard University Press, Cambridge.

[Conte et al., 1998] Conte, R., Castelfranchi, C., and Dignum, F. (1998). Autonomous norm acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNCS*, pages 99–112, Berlin. Springer Verlag.

[Dignum and Dignum, 2001] Dignum, V. and Dignum, F. (2001). Modelling agent societies: Coordination frameworks and institutions. In Brazdil, P. and Jorge, A., editors, *Proceedings of the Tenth Portuguese Conference in Artificial Intelligence (EPIA'01)*, volume 2258 of *LNAI*, pages 191–204, Berlin. Springer-Verlag.

[van Ditmarsch et al., 2007] van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer.

[Durkheim, 1982] Durkheim, E. (1982). *The rules of Sociological Method*. Free Press, New York. first published in French in 1895.

[Engel, 1998] Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations*, 1(2):140–151.

[Esteva et al., 2001] Esteva, M., Padget, J., and Sierra, C. (2001). Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, pages 348–366, Berlin. Springer Verlag.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press, Cambridge.

[Gaudou et al., 2008] Gaudou, B., Longin, D., Lorini, E., and Tummolini, L. (2008). Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In Padgham, L. and Parkes, D. C., editors, *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, pages 728–735. ACM Press.

[Gelati et al., 2004] Gelati, J., Rotolo, A., Sartor, G., and Governatori, G. (2004). Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law*, 12(1-2):53–81.

[Gerbrandy and Groeneveld, 1997] Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–196.

[Gilbert, 1987] Gilbert, M. (1987). Modelling collective belief. *Synthese*, 73(1):185–204.

[Gilbert, 1989] Gilbert, M. (1989). *On Social Facts*. Routledge, London and New York.

[Goldblatt, 1992] Goldblatt, R. (1992). *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California.

[Grossi, 2007] Grossi, D. (2007). *Designing Invisible Handcuffs*. PhD thesis, University of Utrecht, The Netherlands.

[Grossi, 2008] Grossi, D. (2008). Pushing Anderson's envelope: the modal logic of ascription. In *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON'08)*, number 5076 in LNAI, pages 263–277. Springer Verlag.

[Grossi et al., 2006] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2006). Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation*, 16(5):613–643.

[Grossi et al., 2008] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2008). The many faces of counts-as: A formal analysis of constitutive rules. *Journal of Applied Logic*, 6(2):192–217.

[Hakli, 2006] Hakli, P. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7:286–297.

[Hansen et al., 2007] Hansen, J., Pigozzi, G., and van der Torre, L. (2007). Ten philosophical problems in deontic logic. In Boella, G., van der Torre, L., and Verhagen, H., editors, *Normative Multi-agent Systems*, number 07122 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

[Hart, 1992] Hart, H. L. A. (1992). *The concept of law*. Clarendon Press, Oxford. new edition.

[Herzig et al., 2008] Herzig, A., de Lima, T., and Lorini, E. (2008). What do we accept after an announcement? In Meyer, J.-J. and Broersen, J., editors, *Proceedings of the First Workshop on Knowledge Representation for Agents and Multi-Agent Systems (KRAMAS 2008)*, pages 81–94.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca.

[Jones and Sergot, 1996] Jones, A. and Sergot, M. J. (1996). A formal characterization of institutionalised power. *Journal of the IGPL*, 4:429–445.

[Lagerspetz, 2006] Lagerspetz, E. (2006). Institutional facts, performativity and false beliefs. *Cognitive Systems Research*, 7(2-3):298–306.

[Lewis, 1969] Lewis, D. K. (1969). *Convention: a philosophical study*. Harvard University Press, Cambridge.

[Lopez y Lopez et al., 2004] Lopez y Lopez, F., Luck, M., and d'Inverno, M. (2004). Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, pages 732–739. ACM Press.

[Lorini and Longin, 2008] Lorini, E. and Longin, D. (2008). A logical account of institutions: from acceptances to norms via legislators. In Brewka, G. and Lang, J., editors, *Proceedings of the Eleventh International Conference on Principles on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 38–48. AAAI Press.

[Makinson, 1993] Makinson, D. (1993). Five faces of minimality. *Studia Logica*, 52(3):339–379.

[Makinson and van der Torre, 2000] Makinson, D. and van der Torre, L. (2000). Input-output logics. *Journal of Philosophical Logic*, 29:383–408.

[Mantzavinos et al., 2004] Mantzavinos, C., North, D., and Shariq, S. (2004). Learning, institutions, and economic performance. *Perspectives on Politics*, 2:75–84.

[Meyer, 1988] Meyer, J. J. (1988). A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136.

[North, 1990] North, D. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge.

[Pettit, 2001] Pettit, P. (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11:268–99.

[Prakken and Sergot, 1997] Prakken, H. and Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In Nute, D., editor, *Defeasible Deontic Logic*, pages 223–262. Synthese Library.

[Rawls, 1955] Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, 64:3–32.

[Raz, 1975] Raz, J. (1975). *Practical reason and norms*. Hutchinson, London.

[Sahlqvist, 1975] Sahlqvist, H. (1975). Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., editor, *Proceedings of the 3rd Scandinavian Logic Symposium*, volume 82 of *Studies in Logic*.

[Searle, 1969] Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York.

[Searle, 1995] Searle, J. R. (1995). *The Construction of Social Reality*. The Free Press, New York.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. MIT Press, Cambridge.

[Tollefsen, 2002] Tollefsen, D. P. (2002). Challenging epistemic individualism. *Protosociology*, 16:86–117.

[Tollefsen, 2003] Tollefsen, D. P. (2003). Rejecting rejectionism. *Protosociology*, 18–19:389–405.

[Tuomela, 1992] Tuomela, R. (1992). Group beliefs. *Synthese*, 91:285–318.

[Tuomela, 2000] Tuomela, R. (2000). Belief versus Acceptance. *Philosophical Explorations*, 2:122–137.

[Tuomela, 2002] Tuomela, R. (2002). *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge.

[Tuomela, 2007] Tuomela, R. (2007). *The Philosophy of Sociality*. Oxford University Press, Oxford.

[Von Wright, 1963] Von Wright, G. H. (1963). *Norm and Action*. Routledge and Kegan, London.

A Annex: proofs of some theorems

This Annex contains some selected proofs of the theorems presented in the paper.

Proof of Theorem 1

Axiom **K** and rule of inference **Nec** define a minimal normal modal logic. Thus, they do not have an associated semantic constraint. It is a routine task to check that the Axioms **PAccess**, **NAccess**, **Inc**, **Unanim** and **Mon** of the logic $\mathcal{AL}$ correspond to their semantic counterparts **S.1**-**S.5** over $\mathcal{AL}$ models. In particular, the following correspondences exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames.

- Axioms **PAccess** corresponds to the constraint **S.1**.
- Axiom **NAccess** corresponds to the constraint **S.2**.
- Axiom **Inc** corresponds to the constraint **S.3**.
- Axiom **Unanim** corresponds to the constraint **S.4**.
- Axiom **Mon** corresponds to the constraint **S.5**.

It is a routine, too, to check that all of axioms of the logic $\mathcal{AL}$ are in the Sahlqvist class, for which a general completeness result exists.
(See [Sahlqvist, 1975, Blackburn et al., 2001].)

Proof of Theorem 2

For notational convenience, we will use the following abbreviation in the proof:

$$\widehat{\mathcal{A}}_{C:x}\varphi \stackrel{def}{=} \neg\mathcal{A}_{C:x}\neg\varphi$$

We have to prove that if φ is $\mathcal{AL}$ *satisfiable* then it is satisfiable in a finite $\mathcal{AL}$ model.

Suppose that $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V}\rangle$ is a $\mathcal{AL}$ model which satisfies φ. Our aim is to build a finite $\mathcal{AL}$ model which satisfies φ. To do this, we use a filtration method [Blackburn et al., 2001, Goldblatt, 1992].

Let us introduce the following definition.

Definition 1. *A set of formulas Σ is closed under subformulas (cus) if for all formulas φ, φ': if $\varphi \vee \varphi' \in \Sigma$ then so are φ and φ'; if $\neg\varphi' \in \Sigma$ then so is φ; for any $x \in INST$ and $C \in 2^{AGT\star}$ if $\mathcal{A}_{C:x}\varphi \in \Sigma$ then $\varphi \in \Sigma$.*

Let us now consider an arbitrary finite set of formulas Σ_φ which is *closed under subformulas* and which contains φ. From Σ_φ we define the set Σ^+_φ as follows.

Σ^+_φ is defined as the smallest superset of Σ_φ such that:

1. for all $x, y \in INST$ and $C, B \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then $\mathcal{A}_{B:y}\varphi \in \Sigma^+_\varphi$;

2. for all $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$ then $\neg\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$;

3. for all $x \in INST$ and $C \in 2^{AGT\star}$, $\mathcal{A}_{C:x}\bot \in \Sigma_\varphi^+$;

4. $\bot \in \Sigma_\varphi^+$.

The following proposition follows straightforwardly due to the fact that the sets AGT and $INST$ are supposed to be finite.

Proposition 3. *Σ_φ^+ is finite and closed under subformulas.*

We define the relation $\leftrightsquigarrow$ between the worlds in W of the model $\mathcal{M}$. For every two worlds $w, v \in W$:

- $w \leftrightsquigarrow v$ iff for all $\varphi \in \Sigma_\varphi^+$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}, v \models \varphi$.

For every world $w \in W$, we note $|w|$ the equivalence class of world w of $\mathcal{M}$ with respect to $\leftrightsquigarrow$. Moreover, let $W_{\Sigma_\varphi^+} = \{|w| \mid w \in W\}$.

Now, we have to build a filtrated model $\mathcal{M}^f = \langle W^f, \mathscr{A}^f, \mathscr{V}^f \rangle$ of the model $\mathcal{M}$.

Definition 2. *We define $\mathcal{M}^f$ as follows.*

A. $W^f = W_{\Sigma_\varphi^+}$*;*

B. *for every $B \in 2^{AGT\star}$ and $x \in INST$, $|v| \in \mathscr{A}^f_{B:x}(|w|)$ if and only if:*

1. $\forall \mathcal{A}_{B:x}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*;*

2. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:y}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \mathcal{A}_{C:y}\varphi$*;*

3. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \widehat{\mathcal{A}}_{C:y}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \widehat{\mathcal{A}}_{C:y}\varphi$*;*

4. $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:x}\varphi, \widehat{\mathcal{A}}_{C:x}\top \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top \wedge \mathcal{A}_{C:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*;*

5. $\exists i \in B$ *such that* $\forall \mathcal{A}_{i:x}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, v \models \mathcal{A}_{i:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*.*

C. $\mathscr{V}^f(p) = \{|w| \mid \mathcal{M}, w \models p\}$*, for all propositional atoms in* Σ_φ^+*.*

It is straightforward to prove that the model $\mathcal{M}^f$ is indeed a filtration of $\mathcal{M}$ through Σ_φ^+.

Lemma 1. *$\mathcal{M}^f$ is a filtration of $\mathcal{M}$ through Σ_φ^+.*

The next step consists in proving that $\mathcal{M}^f$ is a $\mathcal{AL}$ model.

Lemma 2. *$\mathcal{M}^f$ is a $\mathcal{AL}$ model.*

Proof. We have to prove that the model $\mathcal{M}^f$ satisfies the five semantic constraints **S.1**-**S.5** over $\mathcal{AL}$ models.

Let us start with constraint **S.1**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ then $|w''| \in \mathscr{A}^f_{C:y}(|w|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.1** we just need to prove that the previous items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ when $B \subseteq C$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $B \subseteq C$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

This proves that **S.1** holds.

Let us now consider constraint **S.2**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$ then $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.2** we just need to prove that items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{C:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq C$, it follows that $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. By construction of Σ^+_φ we have $\mathcal{A}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. By construction of Σ^+_φ we have $\mathcal{A}_{D:y}\bot, \widehat{\mathcal{A}}_{D:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

This proves that **S.2** holds.

As a next step we have to prove the model $\mathcal{M}^f$ satisfies the semantic condition **S.3**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \subseteq \mathscr{A}^f_{C:x}(|w|)$.

The following proposition is needed to prove that $\mathcal{M}^f$ satisfies the condition **S.3**.

Proposition 4. *For every $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$.*

Proof. Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, and $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ for all $w \in |w|$. We are going to show that the two facts are inconsistent.

Condition $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies that $\exists |w'| \in W^f$ such that: if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As we have $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$ (by construction of Σ^+_φ) and we have supposed $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we can infer that $\mathcal{M}, w' \models \bot$. $\square$

Let us now prove that $\mathcal{M}^f$ satisfies the condition **S.3**. Consider an arbitrary $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$. Suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $w' \in \mathscr{A}^f_{B:x}(|w|)$. We have to prove that $w' \in \mathscr{A}^f_{C:x}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

So, to prove that $\mathcal{M}^f$ satisfies the condition **S.3** we just need to prove that items **1**-**5** are consequences of $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$. By construction of Σ^+_φ, we have $\widehat{\mathcal{A}}_{C:x}\top \in \Sigma^+_\varphi$. From $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ it follows that $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$ (by Proposition 4). Thus, by definition of $|w|$, we can conclude that $\forall w \in |w|$ it holds that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. Then, in particular, $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. As $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ and $B \subseteq C$, from the latter it follows that $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ (by Axiom **Inc** of the logic $\mathcal{AL}$). As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $\mathcal{A}_{B:x}\varphi \in \Sigma^+_\varphi$ (from $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, by construction of Σ^+_φ), from the latter we conclude $\mathcal{M}, w' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:x}\varphi \wedge \widehat{\mathcal{A}}_{D:x}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \varphi$.

Item 5. From $w' \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\exists i \in B$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As $B \subseteq C$, the latter implies that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.3** holds.

Now, we prove that the model $\mathcal{M}^f$ satisfies the semantic condition **S.4**. That is, we prove that for any $x \in INST$ and $C \in 2^{AGT\star}$:

- if $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ then $|w'| \in \bigcup_{i \in C} \mathscr{A}_{i:x}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$. We have to prove that $|w'| \in \bigcup_{i\in C} \mathscr{A}_{i:x}(|w'|)$. By Definition 2, the latter is equivalent to the fact that $\exists i \in C$ such that:

1. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

Thus, we have to suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ and prove that $\exists i \in C$ which satisfies items **1**-**5**. Items **2** and **3** trivially hold for all $\exists i \in C$. Moreover, items **1** and **5** are the same condition. Therefore, we just need to prove that $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ implies that $\exists i \in C$ which satisfies items **1** and **4**.

From $|w'| \in \mathscr{A}^f_{C:x}(|w|)$, we can infer that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

By Axiom **Inc** of the logic $\mathcal{AL}$ and by construction of Σ^+_φ the following property holds for all $i \in C$. For all $D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ and $\mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$.

From the previous two facts, we conclude that $\exists i \in C$ such that: $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$; $\forall D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.4** holds.

It remains to be proved that the model $\mathcal{M}^f$ satisfies the semantic condition **S.5**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

In order to prove this, we prove first that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ with $B \subseteq C$. We show that these facts are inconsistent.

From $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ we infer $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ (by Axiom **Mon** of the logic $\mathcal{AL}$ and the fact that $B \subseteq C$). From Definition 2 and $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, we can infer that $\exists |w'|$ such that $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. By construction of Σ^+_φ we have that $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$. Thus, as we have $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we conclude that $\exists |w'|$ such that $\mathcal{M}, w' \models \bot$.

This proves that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Now, we have to show that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

$\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies that $\exists w'$ such that $w' \in \mathscr{A}_{B:x}(w)$. As $\mathcal{M}^f$ is a filtration of $\mathcal{M}$ (Lemma 1), from the latter we conclude that $\exists |w'|$ such that $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

This proves that **S.5** holds.

□

Lemma 3. *The model $\mathcal{M}^f$ contains at most 2^n worlds where n denotes the size of Σ^+_φ.*

Proof. From Lemma 1 and Proposition 2.38 given in [Blackburn et al., 2001, p. 79]. □

Lemma 4. *$\mathcal{M}^f$ is a finite model.*

Proof. From Lemma 3 and Proposition 3. □

Lemma 5. *Formula φ is satisfiable in $\mathcal{M}^f$.*

Proof. From Lemma 1 and Proposition 3, the fact that φ is satisfiable in $\mathcal{M}$, the fact that $\varphi \in \Sigma^+_\varphi$ and the Filtration Theorem given in [Blackburn et al., 2001, p. 79]. □

Lemma 6. *The logic $\mathcal{AL}$ has the finite model property.*

Proof. We have started with an arbitrary formula φ which is satisfiable in a $\mathcal{AL}$ model $\mathcal{M}$. We have built a model $\mathcal{M}^f$ and proved that $\mathcal{M}^f$ is a finite $\mathcal{AL}$ model (Lemma 4). Finally, we have proved that φ is satisfiable in $\mathcal{M}^f$ (Lemma 5). Thus, we can conclude that for every formula φ, if φ is $\mathcal{AL}$ satisfiable then, φ is satisfiable in a finite $\mathcal{AL}$ model. □

Theorem 2 is a direct consequence of Lemma 6.

Proof of Theorem 14

As for the logic $\mathcal{AL}$, it is a routine to prove soundness, whereas completeness is again obtained by Sahlqvist completeness theorem. Indeed, all axioms of $\mathcal{AL}^+$ are in the Sahlqvist class, for which a general completeness result exists [Sahlqvist, 1975, Blackburn et al., 2001].

Proof of Theorem 15

In order to prove Theorem 15, it is sufficient to prove that if $INST = CXT$ and φ is a formula of the $\mathcal{GMD}$ logic then: if φ is a theorem of $\mathcal{GMD}$ then $tr(\varphi)$ is a theorem of $\mathcal{AL}^+$ and, if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.

Proposition 5. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if $\vdash_{\mathcal{GMD}} \varphi$ then $\vdash_{\mathcal{AL}^+} tr(\varphi)$.*

Proof. We only need to prove that the translations of the axioms of the $\mathcal{GMD}$ logic are theorems of $\mathcal{AL}^+$ and that the translated rules of inference of $\mathcal{GMD}$ preserves validity.

It is straightforward to show that the translation of the rules of inference $\mathbf{Nec}_{[\![x]\!]}$, $\mathbf{Nec}_{[\![Univ]\!]}$ and **MP** preserve validity. As the $\mathcal{AL}^+$ operators $[x]$ and $[Univ]$ are normal, it is a routine to verify that the translation of the $\mathcal{GMD}$ Axioms $\mathbf{K}_{[\![x]\!]}$ and $\mathbf{K}_{[\![Univ]\!]}$ are

theorems of $\mathcal{AL}^+$. Furthermore, by the definitions of $[x]\,\varphi$ and $[Univ]\,\varphi$, it is just trivial to prove that that the translation of the $\mathcal{GMD}$ Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ is a theorem of $\mathcal{AL}^+$. The translation of the $\mathcal{GMD}$ Axiom $\mathbf{T}_{[\![Univ]\!]}$ is a theorem of $\mathcal{AL}^+$ as well. Indeed, this corresponds to the Axiom $\mathbf{T}_{Univ}$ of the logic $\mathcal{AL}^+$. By Axioms **PAccess**$^+$ and **NAccess**$^+$ we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$ are theorems of $\mathcal{AL}^+$. By the same principles, we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![Univ]\!]}$ and $\mathbf{5}_{[\![Univ]\!]}$ are theorems of $\mathcal{AL}^+$. □

Proposition 6. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

Proof. Suppose that φ is $\mathcal{GMD}$ satisfiable. Thus, there exists a $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0}, \pi \rangle$ which satisfies φ. We prove that we can build a $\mathcal{AL}^+$ model $\mathcal{M}$ which satisfies the same formulas as $\mathcal{M}^{\mathcal{GMD}}$.

As we have supposed $INST = CXT$, the $\mathcal{AL}^+$ model $\mathcal{M}$ associated with the $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}}$ can be defined as follows.

- $W = S$;
- $\forall w \in W, \forall x \in CXT_0, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:x}(w) = S_x$;
- $\forall w \in W, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:Univ}(w) = S$;
- $\forall w \in W, \forall p \in ATM, w \in \pi(p)$ if and only if $w \in \mathscr{V}(p)$.

It is a routine to verify that the previous conditions ensure that the model $\mathcal{M}$ is indeed a $\mathcal{AL}^+$ model. By structural induction on φ, it is also a routine to prove that the previous $\mathcal{AL}^+$ model satisfies the same formulas as the $\mathcal{GMD}$ model it is associated. That is, $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ if and only if $\mathcal{M}, w \models tr(\varphi)$.

Theorem 15 is an immediate corollary of Proposition 5 and Proposition 6.

□

Proof of Theorems 3, 4, 5, 6 and 7

Theorems 3 and 4 can be syntactically proved using $\mathcal{AL}$ logic axiomatization. Theorem 5 proof is based on Theorem 3. As every $\mathcal{A}_{C:x}$ operator is normal, Theorems 6 and 7 can be proved by iteration of the Axiom (**K**) and the Rule of Necessitation (**Nec**) for every group of $2^{AGT\star}$ and every institution of $INST$. We provide in the sequel only the complete proof for Theorems (3a) and (3e).

Proof. Theorem (3a):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \vee \neg\mathcal{A}_{C:x}\bot$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**ProTau**), (**Nec**) and (**K**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, from (1), (2) and (3) by (**ProTau**)

□

Proof. Theorem (3e):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \mathcal{A}_{i:x}\varphi)$, for every $i \in C$, from Axiom (**Inc**) by inference rule (**Nec**),

(2) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C}\mathcal{A}_{i:x}\varphi)$, from (1), by K principles

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\bigwedge_{i\in C}\mathcal{A}_{i:x}\varphi \rightarrow \varphi)$, from (**Unanim**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \varphi)$, from (2) and (3) by (**ProTau**) and (**K**)

(5) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (4) by (**ProTau**) and (**K**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(7) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (5) and (6) by (**ProTau**)

(8) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, by (**ProTau**), (**Nec**) and (**K**)

(9) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (7) and (8) by (**ProTau**)

□

Proof of Proposition 1

Proof. Let suppose that the majority Principle (**Majority**) holds for any sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$. We will prove by induction on the set C_n, that there exists a set C_n such that:

$$(P_n) \qquad (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$$

where $i, j \in AGT$, $C_n \subseteq AGT$, $|C_n| = n$ and $n \geq 2$.

We begin by showing that (P_2) holds.

(1) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\mathcal{A}_{\{i,j\}:x}\varphi$, by (**Inc**).

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\varphi$, from (1) by Theorem (3c)

(2) entails that (P_2) holds.

We suppose that (P_n) holds for any n such that $C_n \subset AGT$. Under this hypothesis we will show that (P_{n+1}) holds. We suppose that C_{n+1} is defined as: $C_{n+1} = C_n \cup \{i\}$, with $i \in AGT$ and $i \notin C_n$ (thus $C_n \subset C_{n+1}$).

(3) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$, by induction hypothesis (P_n)

(4) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi$, from (3) by (**Nec**), (**K**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(5) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_{n+1}:x}\bot$, from (4) by (**PAccess**), (**NAccess**), (**Mon**) and (**ProTau**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \neg\mathcal{A}_{C_n:x}\bot$, by (**Mon**)

(7) $\vdash_{\mathcal{AL}} \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (6) by (**Nec**), (**K**)

(8) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (7) by (**NAccess**) and (**ProTau**)

(9) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_n:x}\bot)$, from (5) and (8) by (**ProTau**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(10) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\bigwedge_{k\in C_n}\mathcal{A}_{k:x}\varphi)$, from (9) by (**Inc**), (**K**), (**Nec**) and (**ProTau**)

(11) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\varphi$, from (10) by (**Majority**), (**K**), (**Nec**) and (**ProTau**)

Thus (11) entails (P_{n+1}).

As (P_2) holds and from (P_n) we can infer that (P_{n+1}) for $n < |AGT|$, we can thus deduce by induction that (P_n) holds for $n \leq |AGT|$. In particular, we can deduce from the extension of the Principles (**Majority**) for every set of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$, that the following counterintuitive formula holds:

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

□

Proof of Proposition 2

Lemma 7.

(7a) $\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

(7b) $\neg\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

Proof. Lemma (7a) and (7b):

(1) $\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$, by (**NegIntrAccept**), for $i \in C$

(2) $Bel_i\neg\mathcal{A}_{C:x}\varphi \rightarrow \neg Bel_i\mathcal{A}_{C:x}\varphi$, by Axiom (**D**) for Bel_i

(3) $Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), and (**ProTau**), for $i \in C$

The proof of Lemma (7b) is similar to the one of Lemma (7a), we only use Axiom (**PIntrAccept**) instead of Axiom (**NegIntrAccept**). □

Proof. Propositions (2a) and (2b):

(1) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi)$, by (**FixPoint**)

(2) $\bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi$, because Bel_i are normal modal operators

(3) $\bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, by Lemma (7a)

(4) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), (3) by (**ProTau**)

(5) $\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$, by (**PIntrAccept**), for every $i \in C$

(6) $\mathcal{A}_{C:x}\varphi \rightarrow E_C(\mathcal{A}_{C:x}\varphi \wedge \mathcal{A}_{C:x}\varphi)$, from (5), by (**ProTau**) and definition of E_C

(7) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (6) by inference rule (**InductionRule**) (left to right direction of Theorem (2a))

(8) $\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (4) and (7)

The proof of Proposition (2b) is similar to the one of Proposition (2a), we only use Lemma (7b) instead of Lemma (7a) and (**NegIntrAccept**) instead of (**PIntrAccept**). □

Proof of Theorem 8

To prove that these formulas are not valid in $\mathcal{AL}$, we only have to exhibit a model where there is a world where these formulas are false. We give the complete proof only for Theorem (8b), the others are very similar.

Proof. Theorem (8b):
We will build a $\mathcal{AL}$ model $\mathcal{M}$ in which there is a world w is which the formula is false, *i.e.*: $\mathcal{M}, w \vDash (\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Let $ATM = \{\varphi_1, \varphi_2, \varphi_3\}$, $AGT = \{i\}$, $INST = \{x, y, z\}$ and $W = \{w, w_x, w_y, w_z\}$. We build the valuation function $\mathscr{V}$: $\mathscr{V}(\varphi_1) = \{w_y\}$, $\mathscr{V}(\varphi_2) = \{w_z\}$ and $\mathscr{V}(\varphi_3) = \{w_y\}$, and the relation $\mathscr{A}$: $\mathscr{A}_{\{i\}:x}(w) = \{w_x\}$, $\mathscr{A}_{\{i\}:y}(w) = \{w_y\}$ and $\mathscr{A}_{\{i\}:z}(w) = \{w_z\}$.

As we want $\mathcal{M}$ to be a $\mathcal{AL}$ model, we ensure that it satisfies the constraints **S.1**-**S.5**.

- In order to satisfy (**S.1**) and (**S.2**) we impose: $\langle w_y, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_z, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_x, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_z, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_x, w_z\rangle \in \mathscr{A}_{\{i\}:z}$ and $\langle w_y, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;
- as there is only one agent in our model, (**S.3**) and (**S.5**) are satisfied;
- in order to satisfy (**S.4**) we impose that: $\langle w_x, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_y, w_y\rangle \in \mathscr{A}_{\{i\}:y}$ and $\langle w_z, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;

In this model $\mathcal{M}$:

- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash \neg[y](\varphi_1 \rightarrow \varphi_2)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_1 \rightarrow \varphi_2)$
- $\mathcal{M}, w \vDash \neg[z](\varphi_2 \rightarrow \varphi_3)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_3) \wedge [y](\varphi_1 \rightarrow \varphi_3) \wedge [z](\varphi_1 \rightarrow \varphi_3)$, *i.e.* $\mathcal{M}, w \vDash [Univ](\varphi_1 \rightarrow \varphi_3)$

We have built a $\mathcal{AL}$ model which satisfies the formula $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Thus, $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not valid in $\mathcal{AL}$. By Theorem 1, we conclude that $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not a theorem of $\mathcal{AL}$. □

Proof of Theorem 9

Proof. Theorems (9a) and (9b):
Since $[x]$ and $[Univ]$ are normal modal operators, they satisfy the rule of equivalence RE [Chellas, 1980]. Theorems (9a) and (9b) follow straightforwardly from RE. □

Proof. Theorem (9c):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_1 \rightarrow \varphi_3)) \leftrightarrow (\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_1 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, from (1) by Theorem (6b)

(3) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \wedge \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \vee \neg[Univ](\varphi_1 \rightarrow \varphi_3))$, by (**ProTau**)

(4) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \vee \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow \neg[Univ]((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_1 \rightarrow \varphi_3))$, by standard properties of normal modal operator $[Univ]$

(5) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \wedge \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow \neg[Univ](\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, from (3) and (4) by (**ProTau**)

(6) $((\varphi_1 \overset{x}{\triangleright} \varphi_2) \wedge (\varphi_1 \overset{x}{\triangleright} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\triangleright} (\varphi_2 \wedge \varphi_3))$, from (2) and (5) and (**ProTau**)

□

Proof. Theorems (9d) and (9e):
The proofs of Theorems (9d) and (9e) are very similar to the previous one. Both apply (**Nec**), (**K**) and propositional tautologies.

□

Proof of Theorem 10

All these theorems follow from the necessitation rule (**Nec**) and logical tautologies. Proofs also need Axiom (**K**) and theorems (M) and (C)[10] [Chellas, 1980] for the distribution over conjunction. We give the complete proof of Theorem (10b) as an example.

Proof. Theorem (10b):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow (\varphi_1 \rightarrow \varphi_3)$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} [x](((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow (\varphi_1 \rightarrow \varphi_3))$, from (1) by (**Nec**)

(3) $\vdash_{\mathcal{AL}} [x]((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow \varphi_3)$, from (2) by (**K**) and (C).

□

[10]The conjunction of both (M) and (C) give the equivalence: $[x](\varphi_1 \wedge \varphi_2) \leftrightarrow ([x]\varphi_1 \wedge [x]\varphi_2)$.

Proof of Theorem 11

Proof. Theorem (11a):

This theorem comes straightforwardly from Theorems 6a, 6b, 7a, 7b, and the following propositional tautologies:

(1) $\vdash_{\mathcal{AL}} (((\varphi \wedge \neg\psi) \rightarrow \mathit{viol}) \wedge (\neg\varphi \rightarrow \mathit{viol})) \rightarrow (\neg\psi \rightarrow \mathit{viol})$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} (\neg\psi \rightarrow \mathit{viol}) \rightarrow ((\neg\psi \wedge \varphi) \rightarrow \mathit{viol})$, by (**ProTau**)

□

Proof. Theorem (11b):

(1) $\vdash_{\mathcal{AL}} O_x\top \rightarrow \neg\,[\mathit{Univ}]\,(\bot \rightarrow \mathit{viol})$, by the definition of $O_x\top$

(2) $\vdash_{\mathcal{AL}} \neg\,[\mathit{Univ}]\,(\bot \rightarrow \mathit{viol}) \rightarrow \neg\,[\mathit{Univ}]\,\top$, by (**ProTau**)

(3) $\vdash_{\mathcal{AL}} \neg\,[\mathit{Univ}]\,\top \rightarrow \bot$, by standard properties of normal modal operator $[\mathit{Univ}]$

(4) $\vdash_{\mathcal{AL}} \top \rightarrow \neg O_x\top$, from (1), (2) and (3) by (**ProTau**)

□