Forbidding undesirable agreements

Paolo Turrini[a], Davide Grossi[b], Jan Broersen[c], John-Jules Ch. Meyer[c]

[a]*Department of Computing, Imperial College London*
[b]*Department of Computer Science, University of Liverpool*
[c]*Department of Information and Computing Sciences, Universiteit Utrecht*

Abstract

The purpose of this contribution[1] is to set up a language to evaluate the results of concerted action among interdependent agents against predetermined properties that we can recognise as desirable from a deontic point of view. Unlike the standard view of logics to reason about coalitionally rational action, the capacity of a set of agents to take a rational decision will be restricted to what we will call *agreements*, which can be seen as solution concepts to a dependence structure present in a certain game. The language will identify those agreements that act accordingly or disaccordingly with the desirable properties arbitrarily set up in the beginning, and will reveal, by logical reasoning, a variety of structural properties of this type of collective action.

Keywords:
Deontic Logic, Coalitional Game Theory

1. Introduction

In the past decade much research in deontic logic has been aimed at incorporating agent interaction in the semantics of the classical operators of obligations, forbiddance and permission. In philosophy John Horty's *Agency and Deontic Logic* has been a turning point for establishing a semantics of the deontic operators in terms of properties of strategic interaction, while the need of regulation of Multi-Agent Systems in computer science has given rise to deontic extensions of action languages, such as Sergot's $nC+$ [2]. Generally speaking, in the logical account of Multi-Agent interaction, it is often assumed that agents can form coalitions, that is they can join forces to achieve a certain outcome. The most commonly used logics for strategic interaction, such as Coalition Logic (CL) [3], Alternating-time Temporal Logic (ATL) [4], Seeing To It That (STIT) [5], are multi-modal logics where the central modal operator $[C]\varphi$ (or similar symbolism) is read as

[1]The present paper generalizes and significantly extends the conference paper on which it is based [1], presented at the 10th International Conference on Deontic Logic in Computer Science in Fiesole, Italy.

"the set of agents C can cooperate to achieve the property φ"

It goes without saying that if agents are allowed to join their forces their capability of reaching a desirable state as well as an undesirable state increases, and issues concerning its regulation crop up. However, various approaches used in game theory [6] suggest that in strategic settings not all coalitions are equally likely to form, for common interest in collective action may not arise. As observed by social scientists [7, 8] the reason for a collective action is often to be found in the interdependence among the agents taking part in that action. As pointed out by [8, p. 161-162],

> "Sociality obviously presupposes two or more agents in a common, shared world. A "Common World" implies that there is *interference* among the actions and goals of the agents: the effects of the action of one agent are relevant for the goals of another: i.e., they either favour the achievement or maintenance of some goals of the other's (*positive interference*), or threat some of them (*negative interference*)" [...].

In this paper we incorporate the study of dependence relations in the standard logical setting to reason about Multi-Agent interaction, i.e. situations in which agents need other agents to satisfy their goals. Building on the work in [9], we start from the observation that only if agents are endowed with the capacity of negotiating their choices on the grounds of their dependence with other agents, coalitions can be formed. To this purpose we will study the notion of *agreement*, a transformation of the interaction structure that allows agents to exchange favours, which can be seen as *solution concept* to the underlying dependence structure [1]. To say it with a slogan, in our logic the central operator $[C]\varphi$ should rather be read as

"the set of agents C can make a binding agreement to achieve the property φ"

Assuming the perspective of a designer of Multi-Agent Systems, the regulation of the agreements that a coalition can give rise to becomes a crucial matter. Indeed, we can find many examples of agreements violating system properties that we recognize as desirable. Think of cartel formation, where more companies, instead of competing to lower prices, *agree* on establishing a common level of price; the aim of such collusion (also called *the cartel agreement*) is to increase individual members' profits by reducing competition. Here the role of a deontic logic is to reason on these possibilities, and label them as forbidden.

In line with a solid tradition of deontic logic that dates back to Anderson and Kanger (for a broad discussion see [10]), we will label certain outcomes

[1] Our notion of agreement is by no means the same as the one used in the bargaining literature, see for instance [6], where players undergo a negotiation process about the allocation of a certain resource. Rather, we see agreements as one-shot transformations taking place in a strategic normal form game that consist of an *exchange of favours* among the participants.

of an interaction as violations. The newly introduced notion of agreement, confronted with this labelling, will acquire a deontic reading, on top of which we can construct the semantics of the standard modal operators of permission, forbiddance and obligation.

Paper Structure. The paper is structured as follows: in Section 2 we provide an informal introduction to dependence theory and lay down the preliminary definitions we will be using throughout the text. In Section 3 we provide a formal representation of agreements in terms of effectivity functions — an abstract representation of power — and preference relations, to be used in Section 4 to build the syntax and the semantics of a logic of agreements. In Section 5 the classical deontic operators are given a semantics in terms of agreements, and an extension of the logic will allow to reason about desirable and undesirable coalitions. Logical and metalogical properties will be provided, showing an application to game-theoretical scenarios. Finally, Section 6 concludes the paper.

2. Preliminaries

2.1. Dependence Theory and Agreements

The theory of dependence has conceptually been introduced in Multi-Agent Systems in a series of works by Castelfranchi and colleagues [11, 8, 12]. At its core lies the informal notion of dependence relation between two agents:

"i depends on j for achieving goal g"

The idea is that there are situations that an agent would like to be realized, which we refer to as *goals*, for which however the contribution of other agents, which we refer to as *favour*, is needed.

Recently, the results in [9] have shown that this conceptual framework can be fully incorporated in the theory of games and the following example gives an informal introduction to the kind of game-theoretic settings dependence theory is interested in.

Example 1 (Strangers on a Train). *In Patricia Highsmith's novel*[2]*, Strangers on a Train [13], that Alfred Hitchcock turned in 1951 into a movie with the same title, the following story takes place:*

> *Two protagonists wish to get out of an unhappy relationship. Architect Guy Haines wants to get rid of his unfaithful wife, Miriam, in order to marry the woman he loves, Anne Faulkner. Charles Anthony Bruno, a psychopathic playboy, deeply desires his father's death.*

We can illustrate the setting with a two persons' matrix as in Figure 1:

[2]We thank an anonymous reviewer of the tenth International Conference of Autonomous Agents and Multi-Agent Systems (AAMAS 2010)—publication [9]—for having brought this example to our attention.

	N	S	O
N	$2,2$	$2,0$	$9,1$
S	$0,2$	$0,0$	$0,1$
O	$1,9$	$1,0$	$8,8$

Figure 1: Strangers on a Train

	N	S	O
N	$2,2$	$0,2$	$1,9$
S	$2,0$	$0,0$	$1,0$
O	$9,1$	$0,1$	$8,8$

Figure 2: Agreement between the Strangers

Both agents have the same possibilities: either do nothing (N), commit the murder of their own significant other (S), or commit the murder of the other person's significant other (O). For convenience we assign numerical values to the outcomes and we assume that the payoffs are of the form (payoff(Guy), payoff(Bruno)), being Guy the Row agent and Bruno the Column agent. Focusing on the choices of Guy (for Bruno the reasoning is symmetric), N is the best choice he can make, for all possible decisions by Bruno—technically N is a dominant strategy and (N, N) *is a dominant strategy equilibrium [6]—while this does not hold for S and O. To achieve the outcome* (N, N) *in the game, it is fair to say that neither Guy depends on Bruno, nor vice versa. However the intuition tells us that Guy would find it reasonable to kill Bruno's father only if he knew that Bruno would kill his wife, and viceversa. This would be possible if Guy could lend his action of killing in exchange to Bruno's one. If this outcome* (O, O) *were the outcome to be selected, as we might expect, both agents would have to play a dominated strategy which maximizes the opponent's welfare.*

The notion of agreement, seen in [9] as a simultaneous exchange of favours, suggests itself. Along these lines, the story of the strangers takes an interesting twist.

Example 2 (Strangers on a train (cont.)). *On a train to see his wife, Guy meets Bruno, who proposes the idea of exchange murders: Bruno will kill Miriam if Guy kills Bruno's father; neither of them will have a motive, and the police will have no reason to suspect either of them.*

If this agreement could take place then the game would be transformed in the one pictured in Figure 2, the transposition of the matrix in Figure 1 under swap of strategies.

If this game were to be played, both agents would have incentive to stick to their promise, i.e. to bring about the outcome resulting in $(8, 8)$ *which happens to be—in economical terms—the outcome with the highest social welfare [6].*

2.2. Cooperative Game Models

In this paper we depart from the game-theoretical framework proposed in [9] abstractly representing strategic interactions by means of *individual* effectivity functions. Effectivity functions were first adopted in [3], in order to provide

a representation of power of group of agents, otherwise called *coalitions*, in a certain state.

Definition 1 (Individual effectivity functions). *Given a set of agents N and a set of worlds W, an* individual effectivity function *(from now on simply effectivity function) is a function $E : W \to (N \to 2^{2^W})$.*

An effectivity function assigns, at each world, a set of sets of states to every agent. If $X \in E(w)(i)$ then the agent is said to be able to *force* or *determine* that the next state after w will be some member of the set X. Intuitively if an agent has this power, it can thus prevent that any state *not* in X (the complement of a set A will be denoted by the set $\overline{A}$) will be the next state, but it might not be able to determine *which* state in X will be the next state. Possibly, some other agents will have the power to refine the choice of i. We assume effectivity functions to be *outcome monotonic*: i.e., for $X \subseteq Y \subseteq W$, if $X \in E(w)(i)$ then $Y \in E(w)(i)$. Sometimes, to keep the description of an effectivity function manageable, it is useful to use the operation $\mathcal{X}^{sup}$ on a set of sets $\mathcal{X}$, that returns its superset closure.

Let us describe the example of the strangers with individual effectivity functions.

Example 3. *Let w be a situation representing the game in Figure 1. We identify the outcomes with their payoff vector instead of their corresponding strategy profile. Guy's effectivity function $E(w)(G)$ amounts to his choices in the game closed under supersets, that is*

$$E(w)(G) = \{\{(2,2),(2,0),(9,1)\},\{(0,2),(0,0),(0,1)\},\{(1,9),(1,0),(8,8)\}\}^{sup}$$

while Bruno's is

$$E(w)(B) = \{\{(2,2),(0,2),(1,9)\},\{(2,0),(0,0),(1,0)\},\{(9,1),(0,1),(8,8)\}\}^{sup}$$

For simplicity, when no ambiguity arises, we can name sets of outcomes, writing for instance $E(w)(G) = \{N,S,O\}^{sup}$. When instead ambiguity does arise we index choices with agents, for instance we use N_G to indicate that doing nothing is a choice by Guy.

It will be useful to represent explicitly what happens to one's effectivity function if the opponents make a certain decision. To this end we define choice restrictions.

Definition 2 (Choice restriction). *Let $E(w)(i)$ be i's effectivity function at state w and let $X \subseteq W$ be a set. The* choice restriction *of $E(w)(i)$ with X, in symbols $E(w)(i) \sqcap X$, is the set $\{X \cap Y \mid Y \in E(w)(i)\}$.*

In strategic interaction, agents not only have powers, but also preferences. Cooperative Game Frames are the kind of all-encompassing models we will be dealing with in the rest of paper.

Definition 3 (Cooperative Game Frames). *Let N be a set of agents, W a set of states, E an effectivity function on N and W, and $\geq_i$ a preference total preorder for each $i \in N$. We call the tuple $(N, W, E, \geq_i)$ a* Cooperative Game Frame*.*

A Cooperative Game Frame with a valuation function, i.e. a tuple $(N, W, E, \geq_i, V)$, for V be a valuation function over a set of a given set of atomic propositions, will be referred to as a *Cooperative Game Model*.

3. Agreements and Coalitional Rationality

In this section we elaborate a model of agreements in terms of preferences and effectivity functions. In doing so we will follow two paths:

- in the first (Section 3.1) we make use of a new notion of undomination (originally from [14]), namely an undomination *for someone else*, as an analogue of dominant strategy for someone else in dependence games.
- in the second (Section 3.2), we make use of the standard notion of undomination, originally introduced in[14] and a particular case of ours, as an analogue of dominant strategy in strategic games. However, we complement it with an operation on effectivity functions, to model permuted games.

Finally, we investigate the assumptions under which these two representations are equivalent.

3.1. Coalitional rationality for someone else

Agreements [9] encode reciprocity among agents: every one plays in favour of some agent and the favour will eventually be returned to him, not necessarily by the same agent.

To this end we define a *Pareto optimal choice for someone else*, that selects maxima in one's order of choices. But unlike the textbook definition of Pareto optimality [6], the maxima are considered in someone's effectivity function, according to someone else's preference order. We limit ourselves to a *for all - for all* type of preference lifting, meaning that we consider a set of outcomes X preferred to another Y, when all outcomes in the former are preferred to all outcomes in the latter, according to an underlying preference relation $\geq_i$ or its strict counterpart $>_i$ — when this is the case we write $X \geq_i Y$ and $X >_i Y$, respectively.

Definition 4 (Pareto optimal choice for someone else). *Let E be an effectivity function, $i, j \in N$ two agents, $w \in W$ a state and $X \in E(w)(i)$ a set in i's effectivity function at state w. X is* Pareto optimal choice *by i for j (in symbols $PO_{(i \hookrightarrow j)}$) at w if, and only if, for no $Y \in E(w)(i)$, $Y >_j X$.*

The definition says that Pareto optimal choices for someone else are those choices in an individual effectivity function such that no better choice exists *for another given agent*. Despite their name, Pareto optimal choices for someone else become standard Pareto optimal choices, i.e. for *oneself* , in case i and j coincide. Let us have a look at Pareto optimal choices for someone else in the example.

Example 4. *In Figure 1 the choice N and the choice O are Pareto optimal choices by all agents for themselves. As a consequence of outcome monotonicity of Pareto optimality, we have that the only choice that is not individually optimal is S, both for Guy and for Bruno. This simply means that the only choice that the strangers do not like in an absolute sense is to kill their own significant other. Pareto optimality for the other agent is even less informative: all three choices for both agents are Pareto optimal for the other. Once again, Pareto optimality does not represent what agents should rationally do taking the opponents into account, but what they should do in an absolute sense.*

The example reiterates the fact, already noticed in [14], that the mere use of Pareto optimality of choice cannot provide a good characterization of individually rational choice, and even less of rational choice for someone else. Once again the limitations of Pareto optimality can be overcome by undominated choices. Here the intuition is that a choice is *undominated for agent j* if it is Pareto optimal for j no matter what the other agents decide to do. This is the formal definition:

Definition 5 (Undomination for someone else). *Let E be an effectivity function, $i, j \in N$ two agents, $w \in W$ a state and $X \subseteq W$ a set. X is an* undominated choice *by i for j in w (in symbols $X \rhd_{i \hookrightarrow j,w}$) if and only if*

1. $X \in E(w)(i)$
2. *for all $\bigcap_{k \neq j} Y_k$ with $Y_k \in E(w)(k)$, $X \cap \bigcap_{k \neq j} Y_k$ is Pareto Optimal for j in $E(w)(i) \sqcap Y$.*

The definition says that for a choice X in the effectivity function of agent i to be undominated for agent j two conditions need to be satisfied: the first (item 1) that X is really a choice available to agent i and the second that there is no better choice for agent j available to agent i (item 2).

Let us illustrate undominated choices for someone else in our motivating example.

Example 5. *In the effectivity function representing the game in Figure 1 the choice of doing nothing (i.e. N) is an undominated choice by each agent for himself, while it is not in the effectivity function representing the game in Figure 2, where instead the choice of killing the other's significant other (i.e. O) is undominated by each agent for himself. However if we not only want to look at individual rationality, but also at what agents could do for the others, we need to resort to undomination for someone else: the choice O in Figure 1 is an undominated choice by each agent for its opponent and the outcome (O, O), resulting from both agents helping each other can already be seen as a possible agreement which both agents can give rise to.*

The example has made clear how favours, so central for the treatment of agreements, can be naturally incorporated in our framework: i depends on j for a choice X if j's strategy in X is a favour for i or, said formally, is undominated choice by j for i.

Before introducing them let us fix some notation. For a finite set X we denote $PERM_X$ the set of all permutations over X. For a permutation $sw : X \to X$ on X, we denote $P_X(sw)$ the partition induced by permutation sw on X, and $\mathcal{P}_X(sw)$ the nonempty powerset of this partition, closed under finite unions. The fact that a set $Y \subseteq X$ is the union of some members of the partition induced by sw will then be simply denoted with $Y \in \mathcal{P}_X(sw)$. Whenever X is understood the notation $\mathcal{P}(sw)$ will be adopted. Permutations form a group under the operation of function composition and are therefore closed under composition and inverse, i.e. for $sw', sw'' \in PERM_X$, we have that $sw' \circ sw'' \in PERM_X$ and that $sw'^{-1} \in PERM_X$.

Now we are ready to define agreements.

Definition 6 (Agreements and reciprocity). *Let E be an effectivity function on W, $C \subseteq N$ a coalition, $sw : C \to C$ a permutation, $w \in W$ a state defined on a given Coalitional Game Model M. A tuple $(\bigcap(X_i)_{i\in C}, sw)$ with $X_i \in E(w)(i)$ is said to be an* agreement *for coalition C at w if*

- $X_i \rhd_{(i \hookrightarrow sw^{-1}(i), w)}$.

The definition says that an agreement is a set of choices for members of a coalition that are rational for some other member for that coalition.

3.2. Permuting effectivity functions

Another way of seeing agreements is as a reallocation of strategic ability according to a certain pattern of dependence, exactly what happens in our starting example when we perform a matrix permutation. In a Cooperative Game Model, however, we can only use effectivity functions and preferences, which are not enough to talk about permutations of effectivity functions. To define them we need to endow those models with an operation of *choice switch*.

Definition 7 (Choice switch). *Let $E(w)(i)$ be a choice set of agent i at world w and sw a permutation on N. Then $E'(w)(i)$ is the* choice switch *for agent i at w following permutation sw if $E'(w)(i) = E(w)(sw(i))$.*

Basically, the choice switch assigns to an agent a new effectivity function, according to a given permutation. For our purposes it is useful to dispose of a global operation of choice switch, that reallocates effectivity functions according to a certain permutation. We abbreviate with $E^{sw}(w)$ the choice set $E(w)$ constituted by the choice switches for each agent i at world w according to permutation sw.

Example 6. *Let w be a situation representing the game in Figure 1 and let sw be a permutation on the agents such that $sw(G) = B$. Bruno's choice switch following sw at w amounts to Guy's choices in the picture, namely*

$$E(w)(sw(G)) = E(w)(B) = \{\{(2,2),(2,0),(9,1)\},\{(0,2),(0,0),(0,1)\},\{(1,9),(1,0),(8,8)\}\}^{sup}$$

which is the effectivity function of Bruno in Figure 2, representing the game scenario after the agreement is taken. For Guy the result is symmetric:

$$E(w)(sw(B)) = E(w)(G) = \{\{(2,2),(0,2),(1,9)\},\{(2,0),(0,0),(1,0)\},\{(9,1),(0,1),(8,8)\}\}^{sup}$$

A permuted individual effectivity function encodes a sort of *candidate agreement*, i.e. a possible reallocation of agents' strategic ability that does not take preferences into account. To obtain a proper agreement we need to identify the undominated choices for each agent at each permutation, i.e. what the agents find it rational to achieve if they could choose for someone else.

Definition 8 (Agreements and permuted games). *Let E be an effectivity function on W, $C \subseteq N$ a coalition, $sw : C \to C$ a permutation, $w \in W$ a state defined on a given Coalitional Game Model M. A tuple $(\bigcap(X_i)_{i\in C}, sw)$ with $X_i \in E(w)(i)$ is said to be an* agreement *for coalition C at world w if*

- $X_i \rhd_{sw^{-1}(i),w}$ *in* $E^{sw}(w)$.

The definition says that an agreement results from an exchange of strategies of individual agents that are individually rational for the agents receiving them. More specifically, the agreement is made by a set X that is an intersection of sets indexed by the agents, and a permutation on the agents. Each part of this set is an undominated choice of agent i in the effectivity function of the agent j indicated by the permutation.

Let us observe how this works in our example.

Example 7. *We can observe that the choice of doing nothing by Guy and by Bruno are undominated choices in the effectivity function obtained from the game in Figure 1. This is because doing nothing, i.e. the profile (N, N) in the game, is a dominant strategy equilibrium. Once, however, the effectivity functions are permuted, dominant strategy equilibria also change. In the game of Figure 2, the choice to kill the other's significant other (the profile (O, O)) is now a dominant strategy. But given the previous definitions. the choice of doing nothing is undominated for each agent and it is thereby, together with the identity permutation, an agreement.*

Agreements, formulated as undominated choices, inherit several properties typical of undomination. The most representative one is that of monotonicity, and its validity is shown by the following proposition.

Proposition 1. *Let $(\bigcap(X_i)_{i\in C}, sw)$ be an agreement for coalition C at a given state w. Then each (Y, sw) such that $\bigcap(X_i)_{i\in C} \subseteq Y$ is an agreement for coalition C at w.*

Proof. By outcome monotonicity of effectivity functions and by the definition of the set Y, Y is such that $Y = \bigcap(Y_i)_{i \in C}$ for $Y_i \in E(w)(sw(i))$. But, as easy to see, we also have that $Y_i \rhd_{sw(i),w}$ in $E(w)(sw(i))$. This is enough to conclude, following Definition 8, that (Y, sw) is an agreement. Q.E.D.

We have now two definitions of agreement, the one in Definition 8 and the other in Definition 6. The following proposition shows that these two definitions are in fact equivalent.

Proposition 2. *Let E be an effectivity function on W, $C \subseteq N$ a coalition, $sw : C \to C$ a permutation, $X \subseteq W$ a set of outcomes, $w \in W$ a state defined on a given Coalitional Game Model M. The tuple $(\bigcap(X_i)_{i \in C}, sw)$ with $X_i \in E(w)(i)$ is an agreement for C at w in the sense of Definition 8 if and only if it is an agreement for C at w in the sense of Definition 6.*

Proof. It follows from the fact that $X_i \rhd_{(sw^{-1}(i),w)}$ *in* $E^{sw}(w)$ is equivalent to $X_i \rhd_{(i \hookrightarrow sw^{-1}(i),w)}$ in $E(w)$. Q.E.D.

The two ways of formalizing agreement with effectivity functions are now fully disentangled and we can move on to their logical analysis.

4. A Logic for Agreements

In this section we introduce the syntax and the models for a modal language to reason about agreements, providing a semantics to relate them. The language, which we call $\mathcal{L}^{\preceq,[i],\downarrow,sw}$, is an extension of propositional logic, with modalities to talk about preferences (using $\preceq_i$ as reverse relation of $\succeq_i$), single agent choice restriction and permutation of effectivity functions. With a few relatively small extensions, the logical language presented in [14] to reason on undominated choices, turns out to be flexible enough to express dependence relations, and also agreements.

Definition 9 (Syntax). *Let Prop be a countable set of atomic propositions. The formulas of $\mathcal{L}^{\preceq,[i],\downarrow,sw}$ have the following grammar:*

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [i]\varphi \mid A\varphi \mid \Diamond_i^{\preceq}\varphi \mid [i \downarrow \varphi]\psi \mid [sw]\varphi$$

where $p \in Prop$ and sw is a permutation on N. The informal reading of the modalities is "agent i can achieve φ", "φ is globally true", "there is a better world than the current one for agent i that satisfies φ", "after agent i choses φ, ψ holds", "permuting effectivity functions according to sw, leads to φ".

The language is equipped with modalities to formalize both the agreements that involve the permutation of the effectivity function — via the modality $[sw]$, that reasons on the consequences of effectivity functions permutation — and the agreements that involve undomination for someone else — via the modalities $[i]$ and $\Diamond_i^{\preceq}$, that reason respectively about the strategic ability of individual agents and their preferences.

Definition 10. *Semantics*
Let M be a CGM.

$$
\begin{aligned}
M, w \models p \quad & \textit{iff} \quad p \in V(w) \\
M, w \models \neg\varphi \quad & \textit{iff} \quad M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi \quad & \textit{iff} \quad M, w \models \varphi \textit{ and } M, w \models \psi \\
M, w \models [i]\varphi \quad & \textit{iff} \quad \varphi^M = \{v \in W \mid M, v \models \varphi\} \in E(w)(i) \\
M, w \models A\varphi \quad & \textit{iff} \quad M, v \models \varphi, \textit{ for all } v \in W \\
M, w \models \Diamond_i^{\preceq}\varphi \quad & \textit{iff} \quad M, w' \models \varphi, \textit{ for some } w' \textit{ with } w \preceq_i w' \\
M, w \models [i \downarrow \psi]\varphi \quad & \textit{iff} \quad \psi^M \in E(w)(i) \textit{ implies } M\downarrow_{(i,\psi^M,w)}, w \models \varphi \\
M, w \models [sw]\varphi \quad & \textit{iff} \quad M|_{sw;w}, w \models \varphi
\end{aligned}
$$

The interpretation of all the operators, apart from $[i \downarrow \varphi]$ (the subgame operator) and $[sw]$ (the switch operator) which will be discussed next, is standard from coalition and preference logics [15, 3].

4.1. The subgame operator

To model choice restrictions we introduce a modal expression of the form

$$[C \downarrow \psi]\varphi$$

whose informal reading is: "in case coalition C chooses ψ, φ holds", where φ and ψ are formulas of the language $\mathcal{L}^{\preceq,>,\mathcal{E},[C]}$ extended with modalities of the form $[C \downarrow \psi]$. We define the dual $\langle C \downarrow \psi\rangle\varphi$ as an abbreviation of $\neg[C \downarrow \psi]\neg\varphi$. Intuitively what we do is to talk about what holds in case the choice ψ of coalition C is performed. Thanks to this operator formulas of the form

$$[C \downarrow \psi][\overline{C}]\varphi$$

allow us to talk of the *restriction* in the coalitional ability of $\overline{C}$ that is caused by coalition C choosing ψ. This restriction clearly resembles the classical one of *subgame*. For this reason it will be called *the subgame operator*.

Its formal interpretation goes as follows:

$$M, w \models [C \downarrow \psi]\varphi \Leftrightarrow \psi^M \in E(w)(C) \text{ implies } M, w\downarrow_{(C,\psi^M)} \models \varphi$$

The interpretation of the operator has a conditional reading: if a coalition C has a certain choice ψ^M at w, then the world where this choice is actually executed ($w\downarrow_{(C,\psi^M)}$, to be formally defined next) makes a certain proposition φ true. Notice that the capacity of C to choose ψ^M is the precondition for C to actually execute ψ^M.

The *updated world* $w\downarrow_{(C,\psi^M)}$ is so defined:

- It inherits the same valuation function as w
- It updates the effectivity function $E(w \downarrow_{(C,\psi^M)})$.

Definition 11. *Let E be an effectivity function defined on a set of outcomes W and a set of agents N and let $C, C' \subseteq N$, $X \subseteq W$ and $w \in W$. $E(w \downarrow_{(C,X)})$ is defined in the following way:*

$$E(w \downarrow_{(C,X)})(C') \doteq (\{X\})^{sup} \qquad \text{for } C' \cap C \neq \emptyset$$

$$E(w \downarrow_{(C,X)})(C') \doteq (E(w)(C') \sqcap X)^{sup} \qquad \text{for } C' \cap C = \emptyset \text{ and } C' \neq \emptyset$$

$$E(w \downarrow_{(C,X)})(C') \doteq E(w)(C') \qquad \text{for } C' = \emptyset$$

The way the relation is updated deserves some comment. A distinction is made between the strategic ability update of the agents who made a certain choice ψ and all the other agents. After coalition C has made a choice ψ, all the coalitions involving agents belonging to C are given $(\{\psi^M\})^{\mathrm{sup}}$ as a choice set. This view maintains that a coalition comprising agents in the coalition that has already chosen cannot further influence the outcome of the game. This fact implies that the subgame operator is not superadditive, in the sense given in [3], that is, bigger coalitions need not have bigger power. Said in other words, we do not allow agents to make a choice within a certain coalition and then, at the same time, to make a different choice within different coalitions. The models of reference are strategic games, in which strategies are decided in the beginning once and for all [6]. The other (nonempty) coalitions instead *truly update* their choice set having it restricted by the choice of C. Restriction is implemented in this case by intersecting the effectivity function with the move that has been carried out. In case for instance C chooses to force ψ and $\overline{C}$ was able to choose ξ, then given the choice by C, $\overline{C}$ is able to force $\xi \wedge \psi$. The coalitional relation at worlds different from the one where the choice is made remains instead unchanged. This means that the update is local. Again, the references are strategic games, where the sequential structure of strategies is substantially ignored. Notice also that by the last condition the empty coalition never gains power. In sum the strategic ability update is governed by three principles:

- the **irrelevance of hybrid coalitions**, that does not allow the members of the coalition that moved to further influence the interaction,
- the **restriction of opponents' choices**, that truly updates the effectivity function of the coalitions opposing the one that moved,
- the **locality of the update**, that only updates the power of nonempty coalitions at one world.

The update operation is treated as a function that takes a triple world-coalition-set as a value and returns a world. A consequence is that the coalition

frames are *special frames* that contain all instances of their updates. In other words, they are *closed under subgames* .

Definition 12 (Closure under subgames). *Let $F = (W, E)$ be a coalition frame. F is said to be* closed under subgames *if and only if $X \in E(w)(C)$ implies that $w \downarrow_{(C,X)} \in W$.*

This is a frame condition and, as many others that we have seen so far, can be modally characterized.

Proposition 3. *Let $F = (W, E)$ be a coalition frame. The following holds:*
$F \models [C]\xi \leftrightarrow \langle C \downarrow \xi\rangle\top$ if and only if F is closed under subgames.

Proof. From right to left, it is straightforward. From left to right assume $F \models [C]\xi \leftrightarrow \langle C \downarrow \xi\rangle\top$. Consider now a set $X \in E(w)(C)$ and take a valuation function V such that $\xi^M = X$ for some M based on F. By the assumptions we have that $M, w \models \langle C \downarrow \xi\rangle\top$, which means that there is a world $w \downarrow_{(C,\xi^M)} \in W$ such that $M, w \downarrow_{(C,\xi^M)} \models \top$, i.e. F is closed under subgames. Q.E.D.

4.2. The operator $[sw]$*, the* switch *operator*

The operator $[sw]$ accounts for the transformation in a model induced by permuting agents' effectivity functions. In the same way we have done with the subgame operator (Definition 11) its interpretation is nonconstructive. Each world w has an outcoming arrow labelled with a permutation sw on agents that goes to another world w' that is equivalent to w as to valuation function but differs for the agents' effectivity functions, that are reallocated according to sw.

Definition 13 (Switch).

$$M, w \models [sw]\varphi \text{ if and only if } M, (sw, w) \models \varphi$$

The updated world (sw, w) is identical to w in all features apart from the effectivity function, which is interpreted as follows:

Definition 14 (Updated worlds for switches).

$$E((sw, w))(i) \doteq E(w)(j) \quad \text{if} \quad sw(i) = j$$

The clause regulating the update deserves a short comment. It says that updating a world means updating its effectivity function, following the given permutation. In other words, if agent j had choice set $\mathcal{Y}$ at world w, then at world (sw, w) agent i will have $\mathcal{Y}$ whenever $sw(i) = j$. In turn the set $\mathcal{X}$ held by agent i at w will be assigned at (sw, w) to agent $sw^{-1}(i)$.

As for the case of the subgame operator, coalition frames are special frames that are *closed under agents permutations*. The closure can be made precise in the following way.

Definition 15 (Closure under agents permutations). *Let $w \in W$ be a world,(sw, w) its update according to permutation sw, and $F = (W, E)$ be a coalition frame. F is said to be* closed under agents permutations *if and only if $w \in W$ implies that $(sw, w) \in W$.*

As for the closure under subgames, it is a frame condition that can be formally characterized.

Proposition 4. *Let $F = (W, E)$ be a coalition frame. The following holds:*
$F \models \langle sw \rangle \top$ if and only if F is closed under agents permutations.

Proof. From right to left, it is straightforward. From left to right assume $F \models \langle sw \rangle \top$. Consider now a world $w \in W$ and consider any permutation $sw : N \to N$. We must have that $(sw, w) \in W$. Q.E.D.

It is worth noticing that the switches we consider are total, while much attention in the literature has been dedicated to partial agreements, that are instead based on partial permutations [9]. We shall see that, exploiting the features of outcome monotonicity of effectivity function and some other mild assumptions, notions analogous to partial agreements can be defined even when using total permutations.

4.3. Validities

The switch operator shares many structural features with the subgame operator. The most fundamental one is the presence of reduction axioms: also in this case the introduction of the subgame operator does not add expressive power to the language provided the models are closed under agents permutations.

Proposition 5 (Reduction Axioms). *The axioms and the rules displayed in Table 1 are valid in Coalition Models.*

A proof is to be found in the appendix (Section Appendix A).

To see more clearly how the reduction works it can be observed that any formula with the switch operator occurring in it can be eventually rewritten as a formula without the switch operator occurring in it, preserving validity. Similar arguments are used in dynamic epistemic logics [16].

4.4. Characterization results

The coming results essentially concern the characterization power of the language with respect to the notions defined at the structural level. With these characterization results, which generalize and extend the ones in [14] to rational choice for someone else, we can make use of the logical language to express and reason about complex interactions between preferences and choices in interdependence.

To start with, Pareto optimal choices for someone else, introduced in Definition 4, can be characterized within the language provided in Definition 9.

Proposition 6. *φ^M is Pareto optimal choice by i for j in w if and only if $M, w \models [i]\varphi \wedge \langle i \rangle \Diamond_j^{\leq} \varphi$*

	Axioms
A1	$[sw]p \leftrightarrow p$
A2	$[sw]\neg\varphi \leftrightarrow \neg[sw]\varphi$
A3	$[sw](\varphi \wedge \psi) \leftrightarrow ([sw]\varphi \wedge [sw]\psi)$
A4	$[sw][k]\varphi \leftrightarrow [sw^{-1}(k)]\varphi$
A5	$[sw]\Box_i^{\preceq}\varphi \leftrightarrow \Box_i^{\preceq}\varphi$
A6	$[sw'][sw]\varphi \leftrightarrow [sw' \circ sw]\varphi$
	Rules
R1	$\varphi \Rightarrow [sw]\varphi$

Table 1: Axioms and rules for the switch operator

Proof. We show only one direction, the other follows a similar pattern. ($\Rightarrow$) Let us assume that φ^M is Pareto optimal choice by i for j in w , i.e. that φ^M is a Pareto optimal choice for agent j at world w in $E(w)(i)$ according to the $(\forall, \forall)$ preference lifting. This means, by Definition 4, that for no $X \in E(w)(i)$, $X \succ_j^{(\forall,\forall)} \varphi^M$ and that $\varphi^M \in E(w)(i)$. In turn this means that for all $X \in E(w)(i)$ $\exists x \in X, \exists y \in \varphi^M$, such that $x \preceq_j y$. By the definition of effectivity functions, no set $X \in E(w)(i)$ is such that $X \subseteq (\neg\Diamond_j^{\preceq}\varphi)^M$. So we can conclude that $M, w \models [i]\varphi \wedge \langle i\rangle\Diamond_j^{\preceq}\varphi$. Q.E.D.

Proposition 6 shows that saying that a choice φ is Pareto optimal for j boils down to saying that it can be performed by an agent (i.e. $[i]\varphi$) and that the agent cannot avoid ending up in a world that is worse for j than some φ world (i.e. $\langle i\rangle\Diamond_j^{\preceq}\varphi$).

We know from [14] that Pareto optimal choices are particularly weak constructs that can however be refined by taking the opponents into account. [17] has moreover shown that the opponents' possibilities can be made formal by using the subgame operator (Section 4.1). In the present case its use, together with the previous result, makes for the possibility of characterizing the notion of undominated choice for someone else.

In the same fashion as what done with the notion of undominated choice [14] we put forward a variety of characterization results for undominated choice for someone else, where the generalizations apply as sketched for the case of Pareto optimal choices.

Proposition 7. *Let $\mathbb{F}$ be the class of Cooperative Game Frames with individual effec-*

tivity functions closed under subgames and let $F \in \mathbb{F}$ be one of them. Let moreover $E(w)(\bar{i}) = \bigcap E(w)(j)$(for $i \neq j$) be a set of sets obtained by superadding the choice sets of all opponents of agent i. The following holds:

$F \models [i]\varphi \rightarrow [\bar{i} \downarrow \psi]([i](\varphi \wedge \psi) \wedge \langle i \rangle \bigvee_i \Diamond_j^{\leq}(\varphi \wedge \psi))$

if and only if each $X \in E(w)(i)$ is such that X is undominated choice by i for j at w

The proof is a straightforward generalization of the one given in [18] for standard undominated choices, and it allows for similar observations: i) in characterizing undomination as a property of the frames, we do not need any restriction on the choices of coalitions; ii) we can characterize a much finer notion of undomination and Pareto optimality of choice: we can talk about all sets in an effectivity function, and not only those that are the truth set of some proposition.

If instead we would like to characterize undomination for someone else at the model level, we need some more restrictive assumptions, namely finiteness of effectivity functions.

Proposition 8. *Let $PO_{i \hookrightarrow j}\varphi$ abbreviate the formula characterizing the fact that φ is a Pareto optimal choices by i for j and let $\{\psi_1, ..., \psi_n\} = E(w)(\bar{i}) = \bigcap E(w)(j)$(for $i \neq j$) be the effectivity function of i's opponents. The following holds:*

$\varphi^M \triangleright_{i \hookrightarrow j, w} \Leftrightarrow M, w \models \bigwedge_{\psi_i \in \{\psi_1, ..., \psi_n\}} [\bar{i} \downarrow \psi_i] PO_{i \hookrightarrow j}(\varphi \wedge \psi_i)$

The proof is, once again, the generalization of the corresponding one for the rational choice by an agent for himself [18]. In the same line of that proposition it shows that with finite effectivity functions, undomination for someone else can be written as a finite conjunction of formulas that make use of the subgame operator and Pareto optimality for someone else. In other words it says that an undominated choice for someone else is a Pareto optimal choice for someone else in every choice restriction. As the latter ones are finitely many a finite conjunction is sufficient to express the formula in the language.

The coming part will characterize agreements inside the language, using all the machinery that we have introduced so far. It is moreover convenient, to shorten notation, to abbreviate the syntactical correspondents of $\varphi^M \triangleright_{i \hookrightarrow j, w}$ characterized in the previous propositions as $[rational_{(i \hookrightarrow j)}]\varphi$.

4.4.1. Characterizing agreements

As anticipated, the introduction of the switch operator in the framework makes it possible to characterize agreements without explicitly defining modal operators capturing rationality for someone else. We carry out the characterization assuming finiteness of effectivity functions and the following definition will ease the presentation of the result.

Definition 16. $\mathcal{A}_C \bigwedge_{i \in C} \varphi_i := \bigvee_{C \in \mathcal{P}(sw)} [sw] \bigwedge_{i \in C} [rational_{(i \hookrightarrow i)}]\varphi_i$

where $\bigvee_{C \in \mathcal{P}(sw)}$ means that the coalition C is a union of orbits of the cycles induced by the permutation sw on N. This definition draws in a formal language

what a set of agents can agree upon: it says that a coalition can agree on $\bigwedge_{i\in C}\varphi_i$ whenever there is a coalition C that can generate $\bigwedge_{i\in C}\varphi_i$ as a partial agreement. Notice that the coalitional ability is defined in terms of a conjunction of individually rational actions, which in turn quantify over all possible choices of one's opponents.

The syntactical and the model theoretical definition can now be related.

Proposition 9. *Let M be a finite Coalitional Game Model closed under subgames and agents permutations. We have that $M, w \models \mathcal{A}_C \bigwedge_{i\in C}\varphi_i \Leftrightarrow$ there exists a permutation sw on C such that $(\bigcap(\varphi_i^M)_{i\in C}, sw)$ is an agreement for C in w.*

Proof. The result follows from Definition 16, Definition 8, Definition 5 and Proposition 8. Q.E.D.

Using Proposition 2 also the following result is straightforward, providing an alternative characterization of agreements in terms of undominated choices for someone else without the switch operator.

Proposition 10. *Let M be a finite Coalitional Game Model closed under subgames and agents permutations. We have the following validity:*
$\mathcal{A}_C \bigwedge_{i\in C}\varphi_i \leftrightarrow \bigvee_{C\in\mathcal{P}(sw)} \bigwedge_{i\in C}[rational_{(i\hookrightarrow sw(i))}]\varphi_i$

The series of syntactic expressions characterizing agreements has shown that the language is powerful enough to account for transformations of agents' strategic abilities following reciprocity cycles. The next section will label these transformations in a deontic logic fashion, aiming at pointing to the desirable ways of forming coalitions via agreements.

5. Deontic Operators

Our motivating example clearly emphasizes and external-systemic perspective on norms, as it describes a rational agreement going against desirable properties.

Along these lines, outcomes will be labelled in accordance to their deontic status and permutations will be judged against this labelling as follows:

- Permutations are forbidden if leading to undesired outcomes (violations);
- Permutations are permitted if not forbidden;
- Permutations are obliged in case all the other possible permutations are forbidden.

The resemblance of the present definition with the one given in [14] for norms on coalitional choices shows that agreements are treated as one possible coalitional choice, and their regulation is inserted in a more general framework. However there is a notable point of difference: coalitional choices are sets of states, while agreements are sets of states endowed with a permutation on

agents. What is more, the latter may be defined on a subset of the set of agents, giving rise to partial agreements.

To bridge the gap we will exploit outcome monotonicity of effectivity functions. We know that if a set $X \subseteq viol^M$ in some model M belongs to the effectivity function of some agent i at some world w then the set $viol^M$ does as well. In other words if an agent can make a choice that, no matter how the other agents choose, will lead to a state in $X \subseteq viol^M$ then it can also make a choice that, no matter how the other agents choose, will lead to a violation.

Making use of this feature, we can apply the standard deontic operators to permutations.

Definition 17 (Deontic Operators on Agreements). *Let $PERM_N$ be the set of all permutations on N and let $sw \in PERM_N$. The operators $F(sw), P(sw), O(sw)$ indicate forbiddance, permission and obligation as follows.*

$$\begin{aligned} F(sw) &:= [sw] \bigvee_{i \in N} \neg [rational_i] \neg viol \\ P(sw) &:= \neg F(sw) \\ O(sw) &:= \bigwedge_{sw' \in PERM_N \neq sw} F(sw') \end{aligned}$$

Norms are here used to label agents permutations. A permutation sw is forbidden if after the corresponding switch for some agent the set $(\neg viol)^M$ is not a rational choice, it is permitted if it is not forbidden, and it is obligated if all other permutations are forbidden.

The operator $F(sw)$ and $\mathcal{A}_C$ of Definition 16 show a form of duality. The correspondence between the two will turn out to be even stricter when forbiddance is applied to coalitions and not to permutations only. For now we can show some relation between the two. The following proposition states that if some permutation is forbidden then the agents together can cooperate to achieve an undesirable state.

Proposition 11. *Let F be a finite Coalitional Game Model closed under subgames and agents permutations. The following holds: $F \models (\bigvee_{sw \in PERM_N} F(sw)) \rightarrow \mathcal{A}_N viol$*

Proof. Assume $M, w \models F(sw)$ for arbitrary M, w and for some permutation sw on N, that is to say $M, w \models [sw] \bigvee_{i \in N} \neg [rational_i] \neg viol$. By the interpretation of the modal operators, there is an agent $sw^{-1}(k)$ for which $(\neg viol)^M$ is not an undominated choice, i.e. for each of them there is a set $X \in E(w)(sw^{-1}(k))$ for which $X \succ_{sw^{-1}(k)} (\neg viol)^M$. A fortiori $X \subseteq viol^M$ and by outcome monotonicity $viol^M$ is undominated. As by outcome monotonicity $\top^M$ is undominated, too, for all $j \neq sw^{-1}(k)$, $viol^M$ is a possible agreement of N. In other words $M, w \models \mathcal{A}_N viol$. Q.E.D.

The following proposition states that if some permutation is permitted then the agents together can cooperate to achieve a desirable state.

Proposition 12. *Let F be a finite Coalitional Game Frame closed under subgames and agents permutations. The following holds:*
$F \models (\bigvee_{sw \in PERM_N} P(sw)) \rightarrow \mathcal{A}_N \neg viol$

Proof. It follows the same pattern of the previous result. Q.E.D.

In both cases the converse does not hold, as $viol^M$ can be identical with the whole domain or N may not be able to agree upon a desirable property.

The validities in this section have shown that the desirability of a potential agreement — as well as its undesirability — always have some implications in terms of rational action. In particular Proposition 11 states that if some potential agreement is undesirable the grand coalition can rationally choose an undesirable state, while Proposition 12 states that if some potential agreement is permitted the grand coalition can rationally choose a desirable state.

The next section will lift these operators from permutations to coalitions.

5.1. A deontic logic for coalition formation

Speculating on the results of the choices that can be agreed upon by a certain coalition, it is immediate to apply the deontic statements to coalitions themselves. The idea is that coalition C is forbidden to form if and only if all the agreements it can give rise to might not lead to a desirable outcome.

Definition 18 (Deontic Operators on Coalitions).

$$\begin{aligned} F(C) &:= \bigwedge_{C \in \mathcal{P}(sw)} [sw] \bigvee_{i \in C} \neg [rational_i] \neg viol \\ P(C) &:= \neg F(C) \\ O(C) &:= F(\overline{C}) \end{aligned}$$

The operator $F(C)$ says, as anticipated, that a coalition C should not form if all agreements it can give rise to might not lead a desirable outcome; it is permitted when it is not forbidden and it is obligated when the opposite coalition is forbidden.

Notice that the expression $\bigwedge_{C \in \mathcal{P}(sw)} [sw] \bigvee_{i \in C} \neg [rational_i] \neg viol$ due to the assumption of finiteness of choices of coalitions can be described within the language. The following reveals the intimate relation between the newly defined forbiddance operator and the agreement modality:

Proposition 13. *The following is a validity of any finite Coalitional Game Frame:*
$F(C) \leftrightarrow \neg \mathcal{A}_C \neg viol$

Proof. From left to right, take an arbitrary M, w such that $M, w \models F(C)$. By definition of $F(C)$, $M, w \models \bigwedge_{C \in \mathcal{P}(sw)} [sw] \bigvee_{i \in C} \neg [rational_i] \neg viol$. This means that for all permutations sw for which $sw(C) = C$ there is some agent $sw(k) \in C$ for which $(\neg viol)^M$ is not undominated in $E(w)(k)$, which in turn means that there is a set $X \in E(w)(k)$ such that $X \succ_{sw(k)} (\neg viol)^M$. (Notice on the fly that $X \subseteq viol^M$.)

But this means that $M, w \not\models \mathcal{A}_C \neg viol$, i.e. $M, w \models \neg \mathcal{A}_C \neg viol$. From right to left, the proof is similar. Q.E.D.

The previous proposition states that forbidding a coalition is equivalent to stating that that coalition cannot avoid agreeing on an undesirable property. The following section is devoted to applying the full-blown modal apparatus we have introduced to the example of the strangers in the train.

5.2. Colouring the strangers

The deontic operators defined in terms of agreements can be fruitfully used to succinctly reason on the relevant properties of strategic interaction. This section makes use of the characterization results obtained so far to reason on the interaction of Figure 1. The same type of reasoning can be extended to all interactions that can be described using single agent effectivity functions.

Proposition 14. *Let M, w be a representation of the game in Figure 1. Let us assign the atomic proposition viol to hold in the outcome (O, O). For G, B being Guy and Bruno, $x \in \{G, B\}$, O, N, S the respective choices, the following formulas hold in M, w:*

$\neg[rational_x]O$	*agents do not find it rational to kill the other agent's significant other*
$\neg[rational_x]S$	*agents do not find it rational to kill their own significant other*
$[rational_x]N$	*agents do find it rational not to kill anyone*
$[(G,B),(B,G)][rational_x]O$	*agents can agree to kill each other's significant other*
$F((G,B),(B,G))$	*it is forbidden to swap murders*
$O((G,G),(B,B))$	*it is obligatory not to swap murders*
$P((G,G),(B,B))$	*it is permitted not to swap murders*

The deontic operators precisely identify the transformations of the game structure leading to desirable and to undesirable consequences.

6. Conclusion

The contribution of the paper consists in developing a modal logic to express dependence relations as first formalized in [9]. To that we add the machinery of deontic logic, in order to discriminate between agreements that do and agreements that do not reach some desirable properties set up in the beginning.

Unlike the standard logics to reason about coalitionally rational action, such as ATL, STIT or CL, the capacity of a set of agents to take a rational decision

have been restricted to what we have called *agreements*, and formalized as a transformation of the interaction structure that exchanges *favours*, i.e. choices that are rational for someone else, among agents.

Our language is based on the one we have studied in [14], which extends Pauly's Coalition Logic with preferences, to account for undominated choices. We generalize the notion of undominated choice to that of undominated choice for someone else and we consequently generalize all related characterization results. We introduced an explicit operator to talk about effectivity function permutations and showed a reduction result to the language without this operator.

The deontic language has allowed us to identify those agreements that act accordingly or disaccordingly with the desirable properties set up in the beginning, and has revealed, by logical reasoning, a variety of structural properties of this type of collective action.

Acknowledgments

Paolo Turrini acknowledges the support of the IEF Marie Curie fellowship "Norms in Action: Designing and Comparing Regulatory Mechanisms for Multi-Agent Systems" (FP7-PEOPLE-2012-IEF, 327424 "NINA") and of the COFUND Marie Curie fellowship "Trust Games" (FP7-PEOPLE-2011-COFUND, 1196394 "TrustGames"), cofunded by the National Research Fund of Luxembourg.

Davide Grossi acknowledges the support of the Dutch Organization for Scientific Research (NWO) under the VENI grant Nr. 639.21.816.

References

[1] P. Turrini, D. Grossi, J. Broersen, J.-J. C. Meyer, Forbidding undesirable agreements: A dependence-based approach to the regulation of multi-agent systems, in: G. Governatori, G. Sartor (Eds.), DEON, Vol. 6181 of Lecture Notes in Computer Science, Springer, 2010, pp. 306–322.

[2] M. J. Sergot, R. Craven, The deontic component of action language nC+, in: DEON, 2006, pp. 222–237.

[3] M. Pauly, Logic for Social Software, ILLC Dissertation Series, 2001.

[4] R. Alur, T. A. Henzinger, O. Kupferman, Alternating-time temporal logic, in: FOCS '97: Proceedings of the 38th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, USA, 1997, p. 100.

[5] N. Belnap, M. Perloff, M. Xu, Facing The Future: Agents And Choices In Our Indeterminist World, Oxford University Press, Usa, 2001.

[6] M. J. Osborne, A. Rubinstein, A Course in Game Theory, MIT Press, 1994.

[7] J. Coleman, Foundations of Social Theory, Belknap Harvard, 1990.

[8] C. Castelfranchi, Modelling social action for ai agents, Artificial Intelligence 103 (1998) 157–182.

[9] D. Grossi, P. Turrini, Dependence in games and dependence games, Autonomous Agents and Multi-Agent Systems 25 (2) (2012) 284–312. doi:10.1007/s10458-011-9176-3.

[10] J. C. Meyer, R. J. Wieringa, Deontic logic: a concise overview, in: Deontic logic in computer science: normative system specification, John Wiley and Sons Ltd., Chichester, UK, UK, 1993, pp. 3–16.

[11] C. Castelfranchi, A. Cesta, M. Miceli, Dependence relations among autonomous agents, in: D. Y. Werner E. (Ed.), Decentralized A.I.-3, Amsterdam: Elsevier, 1992.

[12] C. Castelfranchi, The micro-macro constitution of power, Protosociology 18-19 (2003) 208–265.

[13] P. Highsmith, Strangers on a Train, Nationwide Book Service, 1950.

[14] J. Broersen, R. Mastop, J.-J. Meyer, P. Turrini, A deontic logic for socially optimal norms, in: eontic Logic in Computer Science, 9th International Conference, DEON 2008, Luxembourg, Luxembourg, July 15-18, 2008. Proceedings, Lecture Notes in Computer Science, 2008, pp. 218–232.

[15] P. Blackburn, M. de Rijke, Y. Venema, Modal Logic, Cambridge Tracts in Theoretical Computer Science, 2001.

[16] H. van Ditmarsch, W. van der Hoek, B. Kooi, Dynamic Epistemic Logic, Synthese Library, 2007.

[17] P. Turrini, J. Broersen, R. Mastop, J.-J. C. Meyer, An update operator for strategic ability, in: X. He, J. F. Horty, E. Pacuit (Eds.), Logic, Rationality, and Interaction, Second International Workshop, LORI 2009, Chongqing, China, October 8-11, 2009. Proceedings, Vol. 5834 of Lecture Notes in Computer Science, Springer, 2009.

[18] P. Turrini, Strategic reasoning in interdependence: logical and game-theoretical investigations, PhD Thesis, SIKS dissertation series, 2011.

Appendix A. The switch operator: validities

$$[sw]p \leftrightarrow p$$

Proof. Take arbitrary M, w. $M, w \models [sw]p \Leftrightarrow M, (sw, w) \models p \Leftrightarrow M, w \models p$. Q.E.D.

$$[sw]\neg\varphi \leftrightarrow \neg[sw]\varphi$$

proof. Take arbitrary M, w. $M, w \models [sw]\neg\varphi \Leftrightarrow M, (sw, w) \models \neg\varphi \Leftrightarrow M, (sw, w) \not\models \varphi \Leftrightarrow M, w \not\models [sw]\varphi \Leftrightarrow M, w \models \neg[sw]\varphi$. Q.E.D.

$$[sw](\varphi \wedge \psi) \leftrightarrow ([sw]\varphi \wedge [sw]\psi)$$

Proof. Take arbitrary M, w. $M, w \models [sw](\varphi \wedge \psi) \Leftrightarrow M, (sw, w) \models \varphi \wedge \psi \Leftrightarrow M, (sw, w) \models \varphi$ and $M, (sw, w) \models \psi \Leftrightarrow M, w \models [sw]\varphi$ and $M, w \models [sw]\psi \Leftrightarrow M, w \models [sw]\varphi \wedge [sw]\psi$. Q.E.D.

$$[sw]A\varphi \leftrightarrow A\varphi$$

Proof. Take arbitrary M, w. $M, w \models [sw]A\varphi \Leftrightarrow M, (sw, w) \models A\varphi \Leftrightarrow \varphi^M = W \Leftrightarrow \varphi^M = W \Leftrightarrow M, w \models A\varphi$. Q.E.D.

$$[sw]\Box_i^{\leq}\varphi \leftrightarrow \Box_i^{\leq}\varphi$$

Proof. Take arbitrary M, w. $M, w \models [sw]\Box_i^{\leq}\varphi \Leftrightarrow M, (sw, w) \models \Box_i^{\leq}\varphi \Leftrightarrow M, sw(v) \models \varphi$ for every v such that $w \leq_i v \Leftrightarrow M, v \models \varphi$ for every v such that $w \leq_i v \Leftrightarrow M, w \models \Box_i^{\leq}\varphi$. Q.E.D.

$$[sw][k]\varphi \leftrightarrow [sw^{-1}(k)]\varphi$$

Proof. Take arbitrary M, w. $M, w \models [sw][k]\varphi \Leftrightarrow M, (sw, w) \models [k]\varphi \Leftrightarrow \varphi^M \in E(sw(w))(k) \Leftrightarrow \varphi^M \in E(w)(j)$, for $sw(k) = j \Leftrightarrow M, w \models [sw^{-1}k]\varphi$. Q.E.D.

$$[sw][k \downarrow \psi]\varphi \leftrightarrow [sw^{-1}(k) \downarrow \psi]\varphi$$

Proof. Take arbitrary M, w. $M, w \models [sw][k \downarrow \psi]\varphi \Leftrightarrow M, (sw, w) \models [k \downarrow \psi]\varphi \Leftrightarrow M, (sw, w) \models [k]\psi$ implies $M \downarrow_{k,\psi,w}, (sw, w) \models \varphi \Leftrightarrow M \models [sw^{-1}(k)]\psi$ implies $M \downarrow_{sw^{-1}(k)(w),\psi^M} \models \varphi \Leftrightarrow M \models [sw^{-1}(k) \downarrow \psi]\varphi$. Q.E.D.