Axiomatising Nash-Consistent Coalition Logic

Helle Hvid Hansen[*] Marc Pauly[†]

April 12, 2002

Abstract

We add a rule for Nash-consistency to Coalition Logic, a modal logic for reasoning about the abilities and rights of groups in multi-agent systems. Rights of agents (constitutions) can be formalised using Coalition Logic, and the additional inference rule of Nash-consistency will guarantee that any multi-agent system implementing these rights will be stable, i.e., for any preferences the agents might have, there will be rights they can exercise such that no individual deviation will be profitable. We apply this logic to obtain a formal analysis of Gibbard's paradox, and we provide meta-theoretic results, in particular a complete axiomatisation.

1 Introduction

Cooperation and coalition formation are issues of central concern in the study of multi-agent systems. The complexity of studying multi-agent systems derives from the difficulty of having to model, e.g., how the abilities and rights of different (groups of) agents interact. While there has been much recent interest in the artificial intelligence community in coalition formation and cooperation [11], there have only been few attempts to capture these notions within a logical formalism [8, 7, 2]. In this latter line of research, coalitional ability is formalised in a modal logic which contains expressions of the form $[C]\varphi$ which express that a group of agents C has the ability to bring about a state which satisfies φ. Here, ability is interpreted as having a joint strategy in a strategic game played by the agents. Past research has focused on the analysis and implementation of *constitutions*: Does a given system of interaction guarantee the agents particular rights? Or given a set of rights, is there a system of interaction which implements these rights? The system of rights is specified as a set of formulas of a logical language, and analysis and implementation problems then will amount to questions of model checking and satisfiability checking.

[*]Universiteit van Amsterdam, E-mail: hhhansen@science.uva.nl
[†]University of Liverpool, E-mail: pauly@csc.liv.ac.uk

In the present paper, we extend this approach to incorporate preferences besides rights. For even if a particular procedure can guarantee all agents their respective rights, the outcome of the procedure may not be their most preferred. An agent might observe, for instance, that he can obtain a more preferred outcome by exercising a different right. This in turn will change the outcome of the procedure, possibly provoking another agent to change his strategy, etc.. Since such behaviour demonstrates a lack of stability, we will in general want to exclude procedures where such a scenario can occur. To this end, we will extend Coalition Logic [8] to Nash-consistent systems of interaction, those which have a Nash-equilibrium no matter what preferences the players have. For Nash-consistent procedures, no agent will want to change his strategy since his choice of strategy is optimal given the strategies chosen by the other agents.

Gibbard's paradox, a concrete example of an implementation problem and the role of Nash-consistency, will be discussed in section 4. We begin by defining the notions needed to obtain a semantics for Nash-consistent Coalition Logic in section 2. The logical language of Coalition Logic is introduced in section 3, where we also provide a complete axiomatisation and some other metatheoretic results.

2 Nash-Consistent Mechanisms

In this section we formally define strategic games and game forms (mechanisms) as is standard in game theory [6]. We then proceed to link game forms to effectivity functions [5, 4, 1, 9] in order to characterise the effectivity functions which correspond to Nash-consistent mechanisms (theorem 2.7). This result is a small adaptation of a result first obtained in [10] to the class of playable effectivity functions (definition 2.1).

2.1 Games and Mechanisms

Let N be the finite nonempty set of agents or players, and let S be the nonempty set of states of the world or alternatives. Agents can prefer one state of the world to another, and we shall model this preference by a complete and transitive relation $\succeq_i \subseteq S \times S$ for each agent $i \in N$. As usual, we define $s \succ_i t$ iff $s \succeq_i t$ and $t \not\succeq_i s$. We let $Prf(S)$ be the set of preference profiles over S, i.e., the set of all $\succeq_N = (\succeq_i)_{i \in N}$ such that each $\succeq_i$ is a preference relation on S. Finally, $L(s, \succeq_i) = \{t \in S | s \succeq_i t\}$ is the lower contour set of s under $\succeq_i$, i.e., all the states not strictly preferred to s by agent i.

A *mechanism* or *strategic game form* $G = (N, \{\Sigma_i | i \in N\}, o, S)$ consists of the set of agents N, a nonempty set of strategies or actions Σ_i for every player $i \in N$, the set of states S and an outcome function $o : \Pi_{i \in N} \Sigma_i \to S$ which associates with every tuple of strategies of the players (strategy profile) an outcome state in S. For notational convenience, let $\sigma_C = (\sigma_i)_{i \in C}$

denote the strategy tuple for coalition $C \subseteq N$ which consists of player i choosing strategy $\sigma_i \in \Sigma_i$. Then given two strategy tuples σ_C and $\sigma_{\overline{C}}$ (where $\overline{C} = N \backslash C$), $o(\sigma_C, \sigma_{\overline{C}})$ denotes the outcome state associated with the strategy profile induced by σ_C and $\sigma_{\overline{C}}$.

Given a mechanism $G = (N, \{\Sigma_i | i \in N\}, o, S)$ and a preference profile $\succeq_N$ we obtain a *strategic game* $G^{\succeq_N}$. A strategy profile σ_N is a *Nash-equilibrium* iff

$$\forall i \in N \; \forall \tau_i : o(\sigma_i, \sigma_{-i}) \succeq_i o(\tau_i, \sigma_{-i}),$$

where σ_{-i} abbreviates $\sigma_{N\backslash\{i\}}$. A mechanism G is called *Nash-consistent* iff for every $\succeq_N$ the strategic game $G^{\succeq_N}$ has a Nash-equilibrium.

2.2 Effectivity Functions

An *effectivity function* is any function $E : \mathcal{P}(N) \rightarrow \mathcal{P}(\mathcal{P}(S))$ which is outcome-monotonic: $\forall C{\subseteq}N, X{\subseteq}Y{\subseteq}S$: $X{\in}E(C) \Rightarrow Y{\in}E(C)$. The function E associates to every group of players the sets of outcomes for which the group is effective. In case the group is a singleton, we write $E(i)$ for $E(\{i\})$. Different kinds of effectivity can be linked to strategic games, depending on what precise interpretation one has in mind. Given a mechanism $G = (N, \{\Sigma_i | i \in N\}, o, S)$ a coalition $C \subseteq N$ will be *α-effective* for a set $X \subseteq S$ iff the coalition has a joint strategy which will result in an outcome in X no matter what strategies the other players choose. More formally, its *α-effectivity function* $E^{\alpha}_{G} : \mathcal{P}(N) \rightarrow \mathcal{P}(\mathcal{P}(S))$ is defined as follows:

$$X \in E^{\alpha}_{G}(C) \text{ iff } \exists \sigma_C \forall \sigma_{\overline{C}} \;\; o(\sigma_C, \sigma_{\overline{C}}) \in X.$$

We say that a mechanism G *represents* an effectivity function $E : \mathcal{P}(N) \rightarrow \mathcal{P}(\mathcal{P}(S))$ iff $E = E^{\alpha}_{G}$. G is a *Nash-consistent representation* of E iff G represents E and is Nash-consistent.

While the preceding discussion shows that every mechanism can be linked to an effectivity function, not every effectivity function will be the α-effectivity function of some mechanism. The properties required to obtain a precise characterisation result are the following.

Definition 2.1 An effectivity function $E : \mathcal{P}(N) \rightarrow \mathcal{P}(\mathcal{P}(S))$ is *playable* if it satisfies the following four conditions:

1. $\forall C \subseteq N : \emptyset \notin E(C)$.
2. $\forall C \subseteq N : S \in E(C)$.
3. E is N-maximal: for all X, if $\overline{X} \notin E(\emptyset)$ then $X \in E(N)$.
4. E is *superadditive*: for all X_1, X_2, C_1, C_2 such that $C_1 \cap C_2 = \emptyset$, $X_1 \in E(C_1)$ and $X_2 \in E(C_2)$ imply that $X_1 \cap X_2 \in E(C_1 \cup C_2)$.

The following lemma states that playable effectivity functions guarantee that opposing coalitions cannot force contradictory outcomes. The result follows easily from playability conditions (1) and (4).

Lemma 2.2 Every playable effectivity function $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ is regular, i.e., for all $X \subseteq S$ and $C \subseteq N$, $X \in E(C)$ implies $\overline{X} \notin E(\overline{C})$.

We define the *polar* $E^* : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ of an effectivity function E as

$$E^*(C) \;=\; \{X \subseteq S | \forall X' \in E(\overline{C}) : X \cap X' \neq \emptyset\}.$$

Note that by monotonicity, $X \in E^*(C)$ iff $\overline{X} \notin E(\overline{C})$. For playable effectivity functions, we can think of the polar of E as a kind of maximally consistent completion of E: We give all coalitions as much power as possible, i.e., without contradicting the powers of anyone else.

When given an effectivity function together with a preference profile, we would like to obtain a prediction regarding the outcomes which are likely to result. A natural solution concept to apply is the core which singles out all outcomes s for which there is no coalition which is effective for a set of outcomes all of which are preferred to s by that coalition. Formally, given an effectivity function $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ and a preference profile $\succeq_N$, we call a state s *dominated* if there is a set of states X and a coalition $C \subseteq N$ such that $X \in E(C)$ and for all $i \in C$ and $x \in X$, $x \succ_i s$. The *core* of E under $\succeq_N$, denoted as $C(E, \succeq_N)$, is the set of undominated states. An effectivity function E is *stable* if for all preference profiles $\succeq_N$, $C(E, \succeq_N) \neq \emptyset$.

2.3 Characterisation Results

The first main result characterises the effectivity functions which have a Nash-consistent representation in terms of lower contour sets. It extends the proof of the following result from [8].

Proposition 2.3 An effectivity function is representable if and only if it is playable.

Proof. We only sketch the difficult half of the proof (see [8] for details) which will be used subsequently, showing that an effectivity function is representable if it is playable. Given a playable effectivity function E, we shall construct a mechanism G such that $E = E^{\alpha}_{G}$. To simplify our definitions, assume that $N = \{1, \dots, n\}$.

For $i \in N$, let $\mathcal{C}_i = \{C \subseteq N | i \in C\}$ be the set of coalitions of which i is a member. Let

$$F_i = \{f_i : \mathcal{C}_i \to \mathcal{P}(S) \mid \forall C : f_i(C) \in E(C)\},$$

so F_i consist of all functions f_i which associate to every coalition C in which i participates a set of outcomes for which C is effective. Note that since for all coalitions C, $S \in E(C)$, F_i will be nonempty for every player i.

Given $f \in \Pi_{i\in N} F_i = F_N$ and a coalition C, let $P(f, C)$ be the coarsest partition $\langle C_1, \ldots, C_m \rangle$ of C such that

$$\forall l \leq m \forall i, j \in C_l : f_i(C) = f_j(C).$$

Then given f, let

$$\begin{aligned} P_0(f) &= \langle N \rangle \\ P_1(f) &= P(f, N) = \langle C_1^1, \ldots, C_{k_1}^1 \rangle \\ P_2(f) &= \langle P(f, C_1^1), \ldots, P(f, C_{k_1}^1) \rangle = \langle C_1^2, \ldots, C_{k_2}^2 \rangle \\ &\vdots \\ P_r(f) &= \langle C_1^r, \ldots, C_{k_r}^r \rangle. \end{aligned}$$

Since there are only finitely many players, this partitioning process will eventually stop at some state r where $P_r(f) = P_{r+1}(f)$, and we let $P_\infty(f) = P_r(f) = \langle C_1, \ldots, C_k \rangle$. Since for all $l \leq k$ and $i, j \in C_l$ we have $f_i(C_l) = f_j(C_l)$ we will simply write $f(C_l)$ for it. Now let

$$G(f) = \bigcap_{l=1}^{k} f(C_l).$$

Since $G(f) \neq \emptyset$, we can define the strategic game form $G = (N, \{\Sigma_i | i \in N\}, o, S)$ as follows: Let $H = \{h : \mathcal{P}(S) \setminus \{\emptyset\} \to S \mid h(X) \in X\}$. Then we define

$$\Sigma_i = F_i \times N \times H \text{ and } o(\sigma_N) = h_{i_0}(G(f)),$$

where $\sigma_N = (f_i, t_i, h_i)_{i\in N}$ is a strategy profile and $i_0 = ((t_1 + \cdots + t_n) \; mod \; n) + 1$ indicates the player who has the power to determine the outcome. One can show that for all $C \subseteq N$, $E(C) = E_G^\alpha(C)$. QED

Theorem 2.4 *An effectivity function $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ has a Nash-consistent representation if and only if E is playable and in addition satisfies the following condition:*

$$\forall {\succeq_N} \in Prf(S) \; \exists s \in S \; \forall i \in N : \; L(s, \succeq_i) \in E(N \setminus \{i\}). \tag{1}$$

Proof. As for the easy direction, any effectivity function E represented by some mechanism G must be playable. To show that E also satisfies condition (1), consider some preference profile $\succeq_N$. By Nash-consistency, the game $G^{\succeq_N}$ must have a Nash-equilibrium σ_N with outcome s_0. Since s_0 is a Nash-equilibrium outcome, we have $L(s_0, \succeq_i) \in E(N\backslash\{i\}) = E^{\alpha}_G(N\backslash\{i\})$ for arbitrary $i \in N$.

For the converse direction, recall the proof of the previous result, showing that any playable effectivity function E is representable by some mechanism G. To prove our claim, we only need to show that in case $E = E^{\alpha}_G$ satisfies condition (1), G is Nash-consistent. So take an arbitrary preference profile $\succeq_N \in Prf(S)$ and the state $s_0 \in S$ provided by condition (1). Now let $\sigma_N = (f_i, t_i, h_i)_{i\in N}$ be a strategy profile satisfying the following conditions:

$$\begin{array}{lll} f_i(N) & = L(s_0, \succeq_1) & \\ f_i(N\backslash\{j\}) & = L(s_0, \succeq_j) & \text{for } i \neq j \\ h_i(X) & = s_0 & \text{if } s_0 \in X \end{array}$$

Note first that there are strategy profiles satisfying these conditions: Since $L(s_0, \succeq_j) \in E(N\backslash\{j\})$ by condition (1) and $i \neq j$, we can define $f_i(N\backslash\{j\}) = L(s_0, \succeq_j)$. In particular, $L(s_0, \succeq_1) \in E(N\backslash\{1\})$, and hence using playability also $L(s_0, \succeq_1) \in E(N)$ which justifies the first equality. The partition induced by the f_i of σ_N on N will simply be $\langle N\rangle$, so the outcome will be $o(\sigma_N) = h_{i_0}(L(s_0, \succeq_1)) = s_0$ since $s_0 \in L(s_0, \succeq_1)$.

We claim that σ_N is a Nash-equilibrium. In case $|N| = 1$, σ_N is a Nash-equilibrium because $L(s_0, \succeq_1) \in E(N\backslash\{1\}) = E(\emptyset)$, i.e., all outcomes of the game are in $L(s_0, \succeq_1)$, so player 1 cannot improve over s_0. In case $|N| > 1$, suppose that some player j plays $\tau_j = (f'_j, t'_j, h'_j) \in \Sigma_j$ rather than σ_j. In case $f'_j(N) = L(s_0, \succeq_1)$, the partition induced will still be $\langle N\rangle$ and the outcome s_0. So assume that $f'_j(N) = Y_N \neq L(s_0, \succeq_1)$ and $f'_j(\{j\}) = Y_j$. Then the partition induced by $(\sigma_{N\backslash\{j\}}, \tau_j)$ will be $\langle\{j\}, N\backslash\{j\}\rangle$ because for all $i \neq j$ we have $f_i(N\backslash\{j\}) = L(s_0, \succeq_j)$. As a consequence, $o(\sigma_{N\backslash\{j\}}, \tau_j) = h_{i_0}(Y_j \cap L(s_0, \succeq_j)) \in Y_j \cap L(s_0, \succeq_j) \subseteq L(s_0, \succeq_j)$. Thus, $o(\sigma_{N\backslash\{j\}}, \tau_j) \in L(s_0, \succeq_j)$, so playing τ_j does not yield an improvement for player j, and hence σ_N is indeed a Nash Equilibrium. QED

The following lemma links Nash-consistency to stability, making use of the auxiliary notions of polarity and residuation. Using the polar of E, we define the *residual* $\hat{E} : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ as follows:

$$\hat{E}(C) = \begin{cases} \emptyset & \text{if } C = \emptyset \\ E^*(i) & \text{if } C = \{i\} \\ \{S\} & \text{if } |C| > 1 \end{cases}$$

Since Nash-equilibria are only concerned with individual deviation, the intuition is that $\hat{E}$ captures the maximal power which can be attributed to individuals, or more precisely, what individuals cannot be prevented from achieving.

Lemma 2.5 An effectivity function E has a Nash-consistent representation if and only if E is playable and $\hat{E}$ is stable.

Proof. Necessity: Assume that $G = (N, \{\Sigma_i | i \in N\}, o, S)$ is the Nash-consistent representation of $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$, and let $\succeq_N$ be any preference profile such that σ_N is a Nash equilibrium with $o(\sigma_N) = s$. Then for every $i \in N$ we have $L(s, \succeq_i) \in E^{\alpha}_G(N \backslash \{i\}) = E(N \backslash \{i\})$. Consequently, $\overline{L(s, \succeq_i)} \notin E^*(i) = \hat{E}(i)$, and hence $s \in C(\hat{E}, \succeq_N)$.

Sufficiency: Assume that $\hat{E}$ is stable, and consider any preference profile $\succeq_N$ such that $s \in C(\hat{E}, \succeq_N)$. This means in particular that for all $i \in N$, $\overline{L(s, \succeq_i)} \notin \hat{E}(i) = E^*(i)$, and hence $L(s, \succeq_i) \in E(N \backslash \{i\})$. So by theorem 2.4, E must have a Nash-consistent representation. QED

Lemma 2.6 Given a playable effectivity function $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ over finite S, $\hat{E}$ is stable if and only if

$$\forall (X_i)_{i \in N}\, \forall i \in N: \; X_i \in E^*(i) \text{ implies } \bigcap_{i \in N} X_i \neq \emptyset. \tag{2}$$

Proof. Necessity: Suppose that $\hat{E}$ is stable, and in order to obtain a contradiction, assume that for each $i \in N$ there is a set $X_i \in E^*(i)$ such that $\bigcap_{i \in N} X_i = \emptyset$. Now for each $i \in N$, define a preference relation $\succeq_i$ by letting $X_i \succ_i \overline{X_i}$, i.e., player i prefers every state in X_i to every state not in X_i, and among the states of X_i, player i does not have any preference, and similarly among states not in X_i. Note that while it is not possible that $X_i = \emptyset$, it is possible that $X_i = S$, in which case we have a completely flat preference relation, every state being valued equally by player i.

By stability, under the preference profile $\succeq_N$ defined in this way $\hat{E}$ will have a nonempty core containing some state s_j. Since by assumption $s_j \notin \bigcap_{i \in N} X_i$, there is some $j \in N$ such that $s_j \notin X_j \in E^*(j) = \hat{E}(j)$. But since every state in X_j is preferred by j to s_j, s_j cannot be in the core after all, a contradiction.

Sufficiency: Suppose $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ is playable and satisfies condition (2). For a given preference profile $\succeq_N$, let

$$X_i = \bigcap_{s \in S} \{\overline{L(s, \succeq_i)} \subseteq S \mid \overline{L(s, \succeq_i)} \in E^*(i)\}.$$

Note that it may be the case that there is no $s \in S$ such that $\overline{L(s, \succeq_i)} \in E^*(i)$, in which case $X_i = \bigcap \emptyset = S$. In that case, since $\emptyset \notin E(N \backslash \{i\})$ we have $S = X_i \in E^*(i)$. Otherwise, note that all the $\overline{L(s, \succeq_i)}$ are ordered by inclusion, and since S is finite, there is a smallest set $\overline{L(s_i, \succeq_i)} \in E^*(i)$ and we have $X_i = \overline{L(s_i, \succeq_i)}$. Hence, by condition (2), there is some $s \in \bigcap_{i \in N} X_i$, and $s \in C(\hat{E}, \succeq_N)$ by the definition of the X_i. QED

As a direct consequence of lemmas 2.5 and 2.6 we obtain our main characterisation result:

Theorem 2.7 *For finite S, an effectivity function $E : \mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S))$ has a Nash-consistent representation if and only if E is playable and satisfies condition (2).*

Remark 2.8 Condition (2) is equivalent to the following:

$$\forall (X_i)_{i \in N} : \bigcup_{i \in N} X_i = S \text{ implies } \exists i \in N : X_i \in E(N \backslash \{i\}) \tag{3}$$

Also note that one direction of theorem 2.7 is not restricted to finite models: If E has a Nash-consistent representation then it is playable and satisfies condition (2). This is due to the fact that only the sufficiency part of lemma 2.6 uses the assumption that the domain is finite. $\dashv$

3 Nash-Consistent Coalition Logic

In this section, we first recall the syntax and semantics of Coalition Logic [8] and define the extension, $\mathtt{CLNC_N}$, of the smallest coalition logic, $\mathtt{CL_N}$, with the NC inference rule which captures condition (3). Since our characterisation of Nash-consistency is for finite models, we need to build a finite model to prove completeness. To this end we define a filtration (definition 3.5) of coalition models, and we show soundness and weak completeness of $\mathtt{CLNC_N}$ with respect to Nash-consistent coalition models (theorem 3.9) by proving that the filtration of the canonical $\mathtt{CLNC_N}$-model satisfies condition (3). The strong finite model property of $\mathtt{CL_N}$ and $\mathtt{CLNC_N}$ follow as easy corollaries (corollary 3.11 and 3.12).

3.1 Syntax, Semantics and Axiomatics

For a finite nonempty set of players $N = \{1, \ldots, n\}$ and a (countable) set of proposition letters Φ_0, we define the syntax of Coalition Logic over the language $\mathcal{L} = \mathcal{L}(N, \Phi_0)$ as follows: $\varphi \in \mathcal{L}$ can have the following syntactic form:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid [C]\varphi \qquad \text{where } p \in \Phi_0, C \subseteq N.$$

$\top, \wedge, \rightarrow$ and $\leftrightarrow$ are defined as the usual abbreviations.

A set of formulas Λ over $\mathcal{L}$ is a *coalition logic for* N, if Λ contains all propositional tautologies, is closed under Modus Ponens (MP) and Monotonicity (RM):

$$\frac{\varphi \quad \varphi \to \psi}{\psi}\ \mathrm{MP} \qquad\qquad \frac{\varphi \to \psi}{[C]\varphi \to [C]\psi}\ \mathrm{RM}$$

and contains all instances of the following axiom schemas:

$$\begin{array}{ll} (\bot) & \neg[C]\bot \\ (\top) & [C]\top \\ (\mathtt{N}) & \neg[\emptyset]\neg\varphi \to [N]\varphi \\ (\mathtt{S}) & ([C_1]\varphi \wedge [C_2]\psi) \to [C_1 \cup C_2](\varphi \wedge \psi)) \\ & \text{where } C_1 \cap C_2 = \emptyset. \end{array}$$

Let $\mathtt{CL_N}$ denote the smallest coalition logic for N and define $\mathtt{CLNC_N}$ as $\mathtt{CL_N}$ together with the following inference rule:

$$\frac{\bigvee_{i \in N} \varphi_i}{\bigvee_{i \in N} [N \backslash \{i\}]\varphi_i}\ \mathrm{NC}$$

Similarly to the Coalition Logic axioms, the NC rule is a straightforward translation of condition (3) into $\mathcal{L}$. The NC rule was first proposed in [7].

A *coalition model* for the language $\mathcal{L}$ is a triple $\mathcal{M} = (S, E, V)$, where S is a nonempty set of states, $V : \Phi_0 \to \mathcal{P}(S)$ is a valuation and

$$E : S \to (\mathcal{P}(N) \to \mathcal{P}(\mathcal{P}(S)))$$

is the playable effectivity structure of $\mathcal{M}$. That is, for all $s \in S$, $E(s)$ is a playable effectivity function. We will use the notation $sE_C X$ for $X \in E(s)(C)$ and when $sE_C X$ is not the case, we will write $\neg sE_C X$.

Truth in a coalition model $\mathcal{M} = (S, E, V)$ at a state $s \in S$, is defined as follows:

$$\begin{array}{lll} \mathcal{M}, s \not\models \bot & & \\ \mathcal{M}, s \models p & \text{iff} & p \in \Phi_0 \text{ and } s \in V(p) \\ \mathcal{M}, s \models \neg\varphi & \text{iff} & \mathcal{M}, s \not\models \varphi \\ \mathcal{M}, s \models \varphi \vee \psi & \text{iff} & \mathcal{M}, s \models \varphi \text{ or } \mathcal{M}, s \models \psi \\ \mathcal{M}, s \models [C]\varphi & \text{iff} & sE_C \varphi^{\mathcal{M}} \end{array}$$

where $\varphi^{\mathcal{M}} = \{s \in S \mid \mathcal{M}, s \models \varphi\}$. Due to the characterisation of proposition 2.3, we can associate a mechanism $G(s)$ with each state $s \in S$ in a coalition model $\mathcal{M}$, which implies that $[C]\varphi$ holds at a state s iff the coalition C is effective for $\varphi^{\mathcal{M}}$ in $G(s)$. The notion of Nash-consistency can also be lifted to coalition models: A *coalition model* $\mathcal{M}$ *is Nash-consistent* if for all $s \in S$, $G(s)$ is Nash-consistent.

We will use the following standard terminology and definitions: If φ is a formula and $\mathcal{M}$ is a model with S as its set of states, then φ is *valid in*

$\mathcal{M}$ (denoted by $\mathcal{M} \models \varphi$) iff $\varphi^{\mathcal{M}} = S$, and φ is *valid in a class of models* K (denoted by $\models_{\mathsf{K}} \varphi$) iff for all $\mathcal{M} \in \mathsf{K} : \mathcal{M} \models \varphi$. The set of formulas that are valid in a class of models K will be denoted by Λ_{K}. If Σ is a set of formulas, then $\mathcal{M}, s \models \Sigma$ denotes that all $\sigma \in \Sigma$ are true at s in $\mathcal{M}$, and $\Sigma \models_{\mathsf{K}} \varphi$ denotes that φ is a local logical consequence of Σ in K, i.e. for all $\mathcal{M} \in \mathsf{K}$ and all $s \in S$, if $\mathcal{M}, s \models \Sigma$ then $\mathcal{M}, s \models \varphi$. If Λ is a coalition logic, we will write $\vdash_\Lambda \varphi$ if φ is a theorem of Λ (i.e. $\varphi \in \Lambda$). For a set of formulas Σ, $\Sigma \vdash_\Lambda \varphi$ denotes that φ is deducible from Σ in Λ, i.e. $\vdash_\Lambda \varphi$ or there are $\sigma_1, \ldots, \sigma_n \in \Sigma$ such that $\vdash_\Lambda (\sigma_1 \wedge \ldots \wedge \sigma_n) \to \varphi$. If $\Sigma \vdash_\Lambda \varphi$ is not the case, we write $\Sigma \nvdash_\Lambda \varphi$. A set of formulas Σ is Λ-*consistent* if $\Sigma \nvdash_\Lambda \bot$, and Λ-*inconsistent* otherwise. A formula φ is Λ-consistent if $\{\varphi\}$ is Λ-consistent, otherwise φ is Λ-inconsistent.

A logic Λ is *strongly complete* with respect to a class of models K if for all $\Sigma \cup \{\varphi\} \subseteq \mathcal{L}$, $\Sigma \models_{\mathsf{K}} \varphi$ implies $\Sigma \vdash_\Lambda \varphi$; Λ is *weakly complete* with respect to K if for all $\varphi \in \mathcal{L}$, $\models_{\mathsf{K}} \varphi$ implies $\vdash_\Lambda \varphi$, and Λ is *sound* with respect to K if for all $\varphi \in \mathcal{L}$, $\vdash_\Lambda \varphi$ implies $\models_{\mathsf{K}} \varphi$. Note that another way of stating that Λ is sound and weakly complete with respect to K, is to write $\Lambda = \Lambda_{\mathsf{K}}$. If K is a class of finite coalition models (i.e. for all $\mathcal{M} \in \mathsf{K}$, $\mathcal{M}$ has a finite universe S), then a coalition logic Λ has the *finite model property with respect to* K, if $\Lambda = \Lambda_{\mathsf{K}}$, and Λ has the *strong finite model property with respect to* K, if $\Lambda = \Lambda_{\mathsf{K}}$ and there is a computable function $f(n)$ such that every Λ-consistent formula φ is satisfiable in a model in K containing at most $f(|\varphi|)$ states. Λ has the *(strong) finite model property* if there is a class K of finite coalition models such that Λ has the (strong) finite model property with respect to K.

Completeness of coalition logics can be obtained via canonical model constructions. If Λ is a coalition logic and Γ is a Λ-consistent set of formulas, we can extend Γ to a maximally Λ-consistent set using the standard argument of Lindenbaum's Lemma for which the usual properties hold: $\forall \varphi, \psi \in \mathcal{L}$: (1) $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$, (2) $\varphi \vee \psi \in \Gamma$ iff $\varphi \in \Gamma$ or $\psi \in \Gamma$, and (3) if $\Gamma \vdash_\Lambda \varphi$ then $\varphi \in \Gamma$. We define the *canonical* Λ-*model* $\mathcal{C}^\Lambda = (S^\Lambda, E^\Lambda, V^\Lambda)$ as follows:

$$
\begin{array}{lcl}
S^\Lambda & = & \text{all maximally } \Lambda\text{-consistent sets,} \\
sE^\Lambda_C X & \text{iff} & \begin{cases} \exists \widehat{\varphi} \subseteq X : [C]\varphi \in s & \text{for } C \neq N \\ \forall \widehat{\varphi} \subseteq \overline{X} : [\emptyset]\varphi \notin s & \text{for } C = N, \end{cases} \\
s \in V^\Lambda(p) & \text{iff} & p \in s,
\end{array}
$$

where $\widehat{\varphi} = \{s \in S^\Lambda \mid \varphi \in s\}$. It can be checked that E^Λ is well-defined and playable, hence $\mathcal{C}^\Lambda$ is indeed a coalition model. See [8] for details. From [8] we also have the following results:

Lemma 3.1 (Truth Lemma) If $\mathcal{C}^\Lambda = (S^\Lambda, E^\Lambda, V^\Lambda)$ is the canonical Λ-model for a coalition logic Λ, then for all $s \in S^\Lambda$ and $\varphi \in \mathcal{L}$:

$$\mathcal{C}^\Lambda, s \models \varphi \quad \text{iff} \quad \varphi \in s.$$

Theorem 3.2 *CL_N is sound and strongly complete with respect to the class of all coalition models.*

3.2 Filtration of Coalition Models

Let $\mathcal{M} = (S, E, V)$ be a coalition model and Σ a subformula closed set of formulas over $\mathcal{L}$. Then $\equiv_\Sigma$ is the equivalence relation induced by Σ on S which is defined as follows for all $s, t \in S$:

$$s \equiv_\Sigma t \text{ iff } \forall \varphi \in \Sigma : \mathcal{M}, s \vDash \varphi \Leftrightarrow \mathcal{M}, t \vDash \varphi\ .$$

Let $S_\Sigma = \{\, |s| \mid s \in S\}$ be the set of equivalence classes induced by Σ on S. For $Y \subseteq S$, denote by $|Y|$ the set $\{|s| \mid s \in Y\}$, and for $X \subseteq S_\Sigma$ let $\nmid X \nmid$ be the set $\{s \in S \mid |s| \in X\}$. Then we have the following identities which will be used without further reference:

$$\begin{array}{ll}
\nmid |\varphi^{\mathcal{M}}| \nmid = \varphi^{\mathcal{M}} & \text{for } \varphi \in \Sigma, \\
|\nmid X \nmid| = X & \text{for } X \subseteq S_\Sigma, \\
\nmid X \cap Y \nmid = \nmid X \nmid \cap \nmid Y \nmid & \text{for } X, Y \subseteq S_\Sigma, \\
\overline{|\varphi^{\mathcal{M}}|} = |(\neg\varphi)^{\mathcal{M}}| & \text{for } \varphi \in \Sigma, \\
\overline{\nmid X \nmid} = \nmid \overline{X} \nmid & \text{for } X \subseteq S_\Sigma.
\end{array}$$

Definition 3.3 Let $\mathcal{M} = (S, E, V)$ be a coalition model and Σ a subformula closed set of formulas over $\mathcal{L}$. Then a coalition model $\mathcal{M}^f_\Sigma = (S^f, E^f, V^f)$ is a *filtration of $\mathcal{M}$ through* Σ if

(i) $S^f = S_\Sigma$,
(ii) $\forall C \subseteq N\ \forall \varphi \in \Sigma : sE_C\varphi^{\mathcal{M}} \ \Rightarrow\ |s|E^f_C|\varphi^{\mathcal{M}}|$,
(iii) $\forall C \subseteq N :$ If $|s|E^f_C X$ then $\forall \varphi \in \Sigma : \nmid X \nmid \subseteq \varphi^{\mathcal{M}} \Rightarrow sE_C\varphi^{\mathcal{M}}$,
(iv) $V^f(p) = |V(p)|$ for all proposition letters $p \in \Sigma$.

$\dashv$

The conditions (ii) and (iii) are designed such that we can prove the following lemma.

Lemma 3.4 (Filtration Lemma) If $\mathcal{M}^f_\Sigma = (S^f, E^f, V^f)$ is a filtration of $\mathcal{M}$ through Σ, then for all formulas $\varphi \in \Sigma$ and all $s \in S$ we have:

$$\mathcal{M}, s \vDash \varphi \text{ iff } \mathcal{M}^f_\Sigma, |s| \vDash \varphi.$$

In other words, for all $\varphi \in \Sigma : |\varphi^{\mathcal{M}}| = \varphi^{\mathcal{M}^f_\Sigma}$.

Proof. Induction on the complexity of φ. The atomic case holds by definition. The boolean cases are straightforward, since the fact that Σ is closed under subformulas allows us to apply the induction hypothesis. So we only show the case where $\varphi = [C]\psi \in \Sigma$.

$$\mathcal{M}, s \vDash [C]\psi \Leftrightarrow sE_C\psi^{\mathcal{M}} \Leftrightarrow |s|E^f_C|\psi^{\mathcal{M}}| \overset{I.H.}{\Leftrightarrow} |s|E^f_C\psi^{\mathcal{M}^f_\Sigma} \Leftrightarrow \mathcal{M}^f_\Sigma, |s| \vDash [C]\psi$$

In the second equivalence, the direction from left to right follows from the closure of Σ under subformulas and definition 3.3(ii). The other direction follows from closure of Σ under subformulas, $\nmid|\psi^{\mathcal{M}}|\nmid = \psi^{\mathcal{M}}$ and definition 3.3(iii). QED

We can now define the announced filtration, which will provide the finite model needed in the completeness proof. Let δ be a formula in the language $\mathcal{L}$, then $sfor(\delta)$ denotes the set of subformulas of δ. Define $\Sigma(\delta)$ as the boolean closure of $sfor(\delta)$, i.e. $sfor(\delta) \subseteq \Sigma(\delta)$ and $\forall \varphi, \psi \in \Sigma(\delta) : \neg\varphi, \varphi \vee \psi \in \Sigma(\delta)$. $\Sigma(\delta)$ is clearly subformula closed and since $sfor(\delta)$ is finite, $\Sigma(\delta)$ is finite up to logical equivalence, which is sufficient to yield a finite model when filtrating through $\Sigma(\delta)$. In fact, $|sfor(\delta)| \leq |\delta|$, the length of δ, hence a filtration through $\Sigma(\delta)$ has at most $2^{2^{|\delta|}}$ states.

Definition 3.5 Let $\mathcal{M} = (S, E, V)$ be a coalition model and δ any formula in $\mathcal{L}$. Then we define $\mathcal{F}^{\mathcal{M}}_\delta := (S^{\mathcal{F}}, E^{\mathcal{F}}, V^{\mathcal{F}})$ where

(1) $S^{\mathcal{F}} = S_{\Sigma(\delta)}$,

(2) $|s|E^{\mathcal{F}}_C X$ iff $\begin{cases} \exists\varphi \in \Sigma(\delta) : \varphi^{\mathcal{M}} \subseteq \nmid X \nmid \text{ and } sE_C\varphi^{\mathcal{M}} & \text{for } C \neq N \\ \text{not } |s|E^{\mathcal{F}}_\emptyset \overline{X} & \text{for } C = N, \end{cases}$

(3) $V^{\mathcal{F}}(p) = |V(p)|$ for all proposition letters $p \in \Sigma(\delta)$.

$\dashv$

Note that for $C \neq N$, $E^{\mathcal{F}}_C$ is the smallest filtration of E_C, i.e. we have only added the X to $E^{\mathcal{F}}_C(|s|)$ which are necessary to satisfy definition 3.3 (ii). However, $E^{\mathcal{F}}_N$ is the largest filtration of E_N, which can easily be seen by N-maximality and regularity of E. We now show that $\mathcal{F}^{\mathcal{M}}_\delta$ is indeed a filtration.

Proposition 3.6 If $\mathcal{M} = (S, E, V)$ is a coalition model and $\delta \in \mathcal{L}$, then $\mathcal{F}^{\mathcal{M}}_\delta = (S^{\mathcal{F}}, E^{\mathcal{F}}, V^{\mathcal{F}})$ is a filtration of $\mathcal{M}$ through $\Sigma(\delta)$.

Proof. We start by showing that the conditions from definition 3.3 hold; (i) and (iv) hold by definition.

For (ii), the case $C \neq N$ is clear since $E^{\mathcal{F}}_C$ was taken as the smallest filtration. So assume $C = N$. By regularity 2.2 of E, we have $sE_N\varphi^{\mathcal{M}} \Rightarrow \neg sE_\emptyset(\neg\varphi)^{\mathcal{M}}$. We must show that $\neg|s|E^{\mathcal{F}}_\emptyset|(\neg\varphi)^{\mathcal{M}}|$. Now suppose for contradiction that $|s|E^{\mathcal{F}}_\emptyset|(\neg\varphi)^{\mathcal{M}}|$ holds, then there is a $\psi \in \Sigma(\delta)$ such that

$\psi^{\mathcal{M}} \subseteq \nmid|(\neg\varphi)^{\mathcal{M}}|\nmid = (\neg\varphi)^{\mathcal{M}}$ and $sE_{\emptyset}\psi^{\mathcal{M}}$ and by monotonicity of $E(s)$, $sE_{\emptyset}(\neg\varphi)^{\mathcal{M}}$, a contradiction.

For (iii), assume that $|s|E_C^{\mathcal{F}}X$, $\varphi \in \Sigma(\delta)$ and $\nmid X\nmid \subseteq \varphi^{\mathcal{M}}$. If $C \neq N$, then $|s|E_C^{\mathcal{F}}X$ implies the existence of a $\psi \in \Sigma(\delta)$ such that $\psi^{\mathcal{M}} \subseteq \nmid X\nmid$ and $sE_C\psi^{\mathcal{M}}$. From monotonicity of $E(s)$ and $\nmid X\nmid \subseteq \varphi^{\mathcal{M}}$ it follows immediately that $sE_C\varphi^{\mathcal{M}}$. If $C = N$, we have $|s|E_N^{\mathcal{F}}X \Leftrightarrow \neg|s|E_{\emptyset}^{\mathcal{F}}\overline{X}$, i.e. $\forall\psi \in \Sigma(\delta)$: $\psi^{\mathcal{M}} \subseteq \nmid\overline{X}\nmid \Rightarrow \neg sE_{\emptyset}\psi^{\mathcal{M}}$. By assumption $\nmid X\nmid \subseteq \varphi^{\mathcal{M}}$, hence $(\neg\varphi)^{\mathcal{M}} \subseteq \overline{\nmid X\nmid} = \nmid\overline{X}\nmid$. As $\Sigma(\delta)$ is closed under negation, $\neg\varphi \in \Sigma(\delta)$ and we have $\neg sE_{\emptyset}(\neg\varphi)^{\mathcal{M}}$. From N-maximality of $E(s)$ we obtain $sE_N\varphi^{\mathcal{M}}$.

It remains to show that $\mathcal{F}_{\delta}^{\mathcal{M}}$ is indeed a coalition model, i.e. for all $|s| \in S^{\mathcal{F}}$, $E^{\mathcal{F}}(|s|)$ satisfies the playability conditions of definition 2.1. Let $|s| \in \Sigma(\delta)$ be arbitrary. It is clear from definition 3.5 that $E^{\mathcal{F}}(|s|)$ is outcome-monotonic and N-maximal, so we only need to check conditions 1, 2 and 4 of 2.1:
Condition 1: Suppose for contradiction that there is a $C \subseteq N$ such that $|s|E_C^{\mathcal{F}}\emptyset$. If $C \neq N$, then by definition of $E^{\mathcal{F}}$ there is a $\psi \in \Sigma(\delta)$ such that $\psi^{\mathcal{M}} \subseteq \emptyset$ and $sE_C\psi^{\mathcal{M}}$, hence by monotonicity $sE_C\emptyset$, which is a contradiction with $E(s)$ being playable. If $C = N$, then $|s|E_N^{\mathcal{F}}\emptyset \Leftrightarrow \neg|s|E_{\emptyset}^{\mathcal{F}}S^{\mathcal{F}}$, i.e. for all $\psi \in \Sigma(\delta)$, if $\psi^{\mathcal{M}} \subseteq \nmid S^{\mathcal{F}}\nmid = S$ then $\neg sE_{\emptyset}\psi^{\mathcal{M}}$. But $\delta \vee \neg\delta \in \Sigma(\delta)$, so this would mean that $\neg sE_{\emptyset}S$, again a contradiction with the playability of $E(s)$.
Condition 2: From playability of $E(s)$, we have sE_CS. For the case $C \neq N$, $|s|E_C^{\mathcal{F}}S^{\mathcal{F}}$ follows from $\delta \vee \neg\delta \in \Sigma(\delta)$; $|s|E_N^{\mathcal{F}}S^{\mathcal{F}}$ follows from Condition 1 for $C = \emptyset$.
Condition 4: To show superadditivity of $E^{\mathcal{F}}(|s|)$, assume $|s|E_{C_1}^{\mathcal{F}}X_1$, $|s|E_{C_2}^{\mathcal{F}}X_2$ and $C_1 \cap C_2 = \emptyset$. First consider the case where $C_1, C_2 \neq N$. By definition of $E^{\mathcal{F}}$ there are $\psi_i \in \Sigma(\delta)$ such that $\psi_i^{\mathcal{M}} \subseteq \nmid X_i\nmid$ and $sE_{C_i}\psi_i^{\mathcal{M}}$, $i = 1, 2$. From superadditivity of $E(s)$, we get $sE_{C_1\cup C_2}(\psi_1 \wedge \psi_2)^{\mathcal{M}}$, and from $\psi_i^{\mathcal{M}} \subseteq \nmid X_i\nmid$, $i = 1, 2$ it follows that $(\psi_1 \wedge \psi_2)^{\mathcal{M}} = \psi_1^{\mathcal{M}} \cap \psi_2^{\mathcal{M}} \subseteq \nmid X_1\nmid \cap \nmid X_2\nmid = \nmid X_1 \cap X_2\nmid$. Furthermore, since $\Sigma(\delta)$ is closed under conjunction, $\psi_1 \wedge \psi_2 \in \Sigma(\delta)$, hence $|s|E_{C_1\cup C_2}^{\mathcal{F}}(X_1 \cap X_2)$.

Now consider the situation where one of the C_i is N. By symmetry, it suffices to show the case $C_1 = N, C_2 = \emptyset$, i.e. assume $|s|E_N^{\mathcal{F}}X_1$ and $|s|E_{\emptyset}^{\mathcal{F}}X_2$, then we need to show that $|s|E_N^{\mathcal{F}}(X_1 \cap X_2)$. Assume ad reductio $\neg|s|E_N^{\mathcal{F}}(X_1 \cap X_2)$, then by definition of $E_N^{\mathcal{F}}$, $|s|E_{\emptyset}^{\mathcal{F}}\overline{(X_1 \cap X_2)}$. Applying superadditivity for the previously shown case to $C_1 = C_2 = \emptyset$, we obtain $sE_{\emptyset}^{\mathcal{F}}(\overline{X_1} \cap X_2)$, and by monotonicity $sE_{\emptyset}^{\mathcal{F}}\overline{X_1}$, which by definition of $E_N^{\mathcal{F}}$ implies $\neg|s|E_N^{\mathcal{F}}X_1$, a contradiction. QED

3.3 Soundness and Weak Completeness of $\mathtt{CLNC_N}$

In order to prove that $\mathtt{CLNC_N}$ is weakly complete with respect to Nash-consistent coalition models, we first show that all subsets of $S^{\mathcal{F}}$ are definable by a formula in $\Sigma(\delta)$.

Lemma 3.7 (Definability) Let $\mathcal{M} = (S, E, V)$ be a coalition model and $\mathcal{F}_\delta^\mathcal{M} = (S^\mathcal{F}, E^\mathcal{F}, V^\mathcal{F})$ as defined in definition 3.5, then for each $X \subseteq S^\mathcal{F}$, there is a formula $\Psi_X \in \Sigma(\delta)$ such that for all $s \in S$:

$$s \in \Psi_X^\mathcal{M} \text{ iff } |s| \in X$$

As a consequence, $|\Psi_X^\mathcal{M}| = X$ and $\Psi_X^\mathcal{M} = \nmid X \nmid$.

Proof. First we note that since $\Sigma(\delta)$ is finite up to logical equivalence, we can find a finite subset $\Sigma_0 \subseteq \Sigma(\delta)$ such that for all $\varphi \in \Sigma(\delta)$ there is a $\varphi_0 \in \Sigma_0$ such that $\mathcal{M} \vDash \varphi \leftrightarrow \varphi_0$.
For each $s \in S$, let Ψ_s be the finite conjunction of $\{\psi \in \Sigma_0 \mid s \in \psi^\mathcal{M}\} \cup \{\neg\psi \mid \psi \in \Sigma_0 \wedge s \notin \psi^\mathcal{M}\}$. We then have

$$t \in \Psi_s^\mathcal{M} \text{ iff } |s| = |t| \tag{4}$$

Since assume $t \in \Psi_s^\mathcal{M}$, i.e. $\mathcal{M}, t \vDash \Psi_s$ and let $\varphi \in \Sigma(\delta)$, then there is a $\varphi_0 \in \Sigma_0$ such that $\varphi^\mathcal{M} = \varphi_0^\mathcal{M}$. If $s \in \varphi^\mathcal{M}$, then $\mathcal{M} \vDash \Psi_s \to \varphi_0$ and hence $\mathcal{M} \vDash \Psi_s \to \varphi$. This implies that $\mathcal{M}, t \vDash \varphi$. If $s \notin \varphi^\mathcal{M}$ then $\mathcal{M} \vDash \Psi_s \to \neg\varphi$ and it follows that $\mathcal{M}, t \vDash \neg\varphi$. For the direction from right to left, assume that $|s| = |t|$, then $\Psi_s = \Psi_t$ and clearly $t \in \Psi_t^\mathcal{M} = \Psi_s^\mathcal{M}$.

Now suppose $X = \{|s_1|, \ldots, |s_k|\}$, then take $\Psi_X := \Psi_{s_1} \vee \ldots \vee \Psi_{s_k}$. Using (4), we have: $s \in \Psi_X^\mathcal{M} \Leftrightarrow s \in \Psi_i^\mathcal{M}$ for some $i \in \{1, \ldots, k\} \Leftrightarrow |s| = |s_i|$ for some $i \in \{1, \ldots, k\} \Leftrightarrow |s| \in X$.

Observe that Ψ_X is a boolean combination of formulas in $\Sigma(\delta)$, and since $\Sigma(\delta)$ is closed under boolean connectives, it follows that $\Psi_X \in \Sigma(\delta)$. QED

With the definability lemma, we can prove the following preservation result:

Proposition 3.8 If $\mathcal{M} = (S, E, V)$ is a coalition model satisfying condition (3) and $\delta \in \mathcal{L}$, then $\mathcal{F}_\delta^\mathcal{M} = (S^\mathcal{F}, E^\mathcal{F}, V^\mathcal{F})$ also satisfies condition (3).

Proof. Suppose $S^\mathcal{F} = \bigcup_{i \in N} B_i$, then we need to show that for all $|s| \in S^\mathcal{F}$,

$$\exists i \in N : |s| E_{N \setminus \{i\}}^\mathcal{F} B_i.$$

We have $S^\mathcal{F} = \bigcup_{i \in N} B_i \Rightarrow S = \bigcup_{i \in N} \nmid B_i \nmid$ and since $\mathcal{M}$ satisfies condition (3), there is an $i \in N$ such that $s E_{N \setminus \{i\}} \nmid B_i \nmid$ for all $s \in S$. By the definability lemma 3.7 there is a formula $\Psi_{B_i} \in \Sigma(\delta)$ such that $\Psi_{B_i}^\mathcal{M} = \nmid B_i \nmid$, hence $|s| E_{N \setminus \{i\}}^\mathcal{F} B_i$. QED

Finally, we are ready to prove our main result of this section:

Theorem 3.9 *$\mathtt{CLNC_N}$ is sound and weakly complete with respect to the class* NC *of Nash-consistent coalition models.*

Proof. The soundness of the playability axioms of $\mathtt{CLNC_N}$ with respect to NC is clear. To see that the NC rule is also sound with respect to NC, let $\mathcal{M} = (S, E, V) \in \mathsf{NC}$ and assume that $\mathcal{M} \vDash \bigvee_{i \in N} \psi_i$. This means that $\bigcup_{i \in N} \psi_i^{\mathcal{M}} = S$. As $\mathcal{M}$ satisfies condition (3) by theorem 2.7 and remark 2.8, it follows that for all $s \in S$ there is an $i \in N$ such that $sE_{[N \setminus \{i\}]} \psi_i^{\mathcal{M}}$, hence $\mathcal{M} \vDash \bigvee_{i \in N} [N \setminus \{i\}] \psi_i$.

To show weak completeness, suppose δ is a $\mathtt{CLNC_N}$-consistent formula. Again by the characterisation of theorem 2.7 and remark 2.8, it suffices to show that δ is satisfiable on a finite coalition model which satisfies condition (3). As δ is $\mathtt{CLNC_N}$-consistent, it can be extended to a maximally $\mathtt{CLNC_N}$-consistent set s. Let $\mathcal{C} = (S, E, V)$ be the $\mathtt{CLNC_N}$-canonical model, then we have from the truth lemma 3.1 that $\mathcal{C}, s \vDash \delta$. Taking the filtration $\mathcal{F}_\delta^{\mathcal{C}} = (S^{\mathcal{F}}, E^{\mathcal{F}}, V^{\mathcal{F}})$ of $\mathcal{C}$ through $\Sigma(\delta)$ we obtain a finite coalition model by proposition 3.6, which, by the filtration lemma 3.4, satisfies δ at state $|s| \in S_{\mathcal{F}}$, so by the characterisation of theorem 2.7 and remark 2.8, it only remains to show that $\mathcal{F}_\delta^{\mathcal{C}}$ satisfies condition (3).

Assume that $B_i \subseteq S^{\mathcal{F}}$, $i \in N$, and $\bigcup_{i \in N} B_i = S^{\mathcal{F}}$. By the definability lemma 3.7 and and the filtration lemma there are $\psi_1, \ldots, \psi_n \in \Sigma(\delta)$ such that $\psi_i^{\mathcal{F}_\delta^{\mathcal{C}}} = |\psi_i^{\mathcal{C}}| = B_i$, hence from the assumption $\mathcal{F}_\delta^{\mathcal{C}} \vDash \bigvee_{i \in N} \psi_i$. Note, since $\Sigma(\delta)$ is closed under boolean connectives, $\bigvee_{i \in N} \psi_i \in \Sigma(\delta)$, so from the filtration lemma it follows that for all $s \in S$, $\mathcal{C}, s \vDash \bigvee_{i \in N} \psi_i$ and from the truth lemma and properties of maximally $\mathtt{CLNC_N}$-consistent sets, $\bigvee_{i \in N} \psi_i \in \mathtt{CLNC_N}$. Applying the NC rule we obtain $\bigvee_{i \in N} [N \setminus \{i\}] \psi_i \in \mathtt{CLNC_N}$ which, again by properties of maximally $\mathtt{CLNC_N}$-consistent sets and the truth lemma, implies that for all $s \in S$ there is an $i \in N$ such that $\mathcal{C}, s \vDash [N \setminus \{i\}] \psi_i$, i.e. $sE_{N \setminus \{i\}} \psi_i^{\mathcal{C}}$. As $\mathcal{F}_\delta^{\mathcal{C}}$ is a filtration, it satisfies (ii) of definition 3.3, hence $|s| E^{\mathcal{F}}_{N \setminus \{i\}} |\psi_i^{\mathcal{C}}|$, and since $|\psi_i^{\mathcal{C}}| = B_i$ and s was arbitrary in S, we have shown that for all $|s| \in S^{\mathcal{F}}$, there is an $i \in N$ such that $|s| E^{\mathcal{F}}_{N \setminus \{i\}} B_i$. QED

Remark 3.10 An easy inspection of the above proof of theorem 3.9 shows that $\mathtt{CLNC_N}$ is also sound and weakly complete with respect to the class FNC of *finite* Nash-consistent coalition models: Soundness of the NC rule follows from theorem 2.7 and remark 2.8, the characterisation of FNC in terms of condition (3), and the weak completeness result of theorem 3.9 is, in fact, proved by showing the stronger result of weak completeness of $\mathtt{CLNC_N}$ with respect to FNC. Hence we have $\mathtt{CLNC_N} = \Lambda_{\mathsf{NC}} = \Lambda_{\mathsf{FNC}}$. $\dashv$

3.4 Strong Finite Model Property

The completeness results for $\mathtt{CL_N}$ and $\mathtt{CLNC_N}$ together with the properties of the filtration of definition 3.5 allow us to show the strong finite model property for both of these coalition logics.

Corollary 3.11 $\mathtt{CL_N}$ has the strong finite model property.

Proof. $\mathtt{CL_N}$ is weakly complete with respect to the class F of finite coalition models, since by theorem 3.2 every $\mathtt{CL_N}$-consistent formula δ is satisfiable on a coalition model $\mathcal{M} = (S, E, V)$ at a state s, and by proposition 3.6, $\mathcal{F}_\delta^{\mathcal{M}}$ is a finite coalition model, which by the filtration lemma 3.4 satisfies δ at $|s|$. $\mathtt{CL_N}$ is also sound with respect to F by theorem 3.2, hence $\mathtt{CL_N}= \Lambda_{\mathsf{F}}$. Furthermore, $\mathcal{F}_\delta^{\mathcal{M}}$ has at most $2^{2^{|\delta|}}$ states, and we have thus proved that $\mathtt{CL_N}$ has the strong finite model property. QED

Corollary 3.12 $\mathtt{CLNC_N}$ has the strong finite model property.

Proof. As stated in remark 3.10, we have $\mathtt{CLNC_N}= \Lambda_{\mathsf{FNC}}$, where FNC is the class of finite Nash-consistent coalition models. The proof of theorem 3.9 also shows that every $\mathtt{CLNC_N}$-consistent formula δ is satisfiable in the filtration $\mathcal{F}_\delta^{\mathcal{C}}$ of the $\mathtt{CLNC_N}$-canonical model $\mathcal{C}$ through $\Sigma(\delta)$, and since $\mathcal{F}_\delta^{\mathcal{C}}$ is in FNC and has at most $2^{2^{|\delta|}}$ states, this proves that $\mathtt{CLNC_N}$ has the strong finite model property. QED

The question arises whether we can improve on the double-exponential bound given by the filtration. In particular, is it possible to filtrate through a $\Sigma(\delta)$ which has size linear in $|\delta|$? This would provide a single-exponential bound. For $\mathtt{CLNC_N}$, there seems to be no improvement possible on the size of $\Sigma(\delta)$, since we need each subset of the filtrated model to be definable by a formula in $\Sigma(\delta)$. But it turns out that the bound for $\mathtt{CL_N}$ can indeed be improved, if we allow $|\Sigma(\delta)|$ to be expressed in terms of the number of agents, as well as $|\delta|$. However an alternative definition of filtrations is needed:

Definition 3.13 (Alternative filtration of coalition models) Let Σ be a subformula closed set of formulas over $\mathcal{L}$ and $\mathcal{M} = (S, E, V)$ a coalition model. Then a coalition model $\mathcal{M}_\Sigma^f = (S^f, E^f, V^f)$ is a *filtration of* $\mathcal{M}$ *through* Σ if

(i) $S^f = S_\Sigma$,
(ii) For all $[C]\varphi \in \Sigma : sE_C\varphi^{\mathcal{M}} \Rightarrow |s|E_C^f|\varphi^{\mathcal{M}}|$,
(iii) If $|s|E_C^f X$ then $\forall [C]\varphi \in \Sigma : \nmid X \nmid \subseteq \varphi^{\mathcal{M}} \Rightarrow sE_C\varphi^{\mathcal{M}}$,
(iv) $V^f(p) = |V(p)|$ for all proposition letters $p \in \Sigma$.

⊣

It can easily be checked that the filtration lemma still follows from definition 3.13. We now adjust $\Sigma(\delta)$ and definition 3.5, such that $\Sigma(\delta)$ is the minimal set of formulas which ensures that we can still prove the filtration to be a coalition model. Given a formula δ, define $\Sigma(\delta)$ as follows:

$$\begin{array}{lll}
\Sigma_0 & := & \mathit{sfor}(\delta) \cup \{[\emptyset]\neg\psi, \neg\psi \mid [N]\psi \in \mathit{sfor}(\delta)\} \cup \{\top\} \cup \{[C]\top \mid C \subsetneq N\} \\
\Sigma_{i+1} & := & \Sigma_i \ \cup \ \{\psi_1 \wedge \psi_2 \mid [C_i]\psi_i \in \Sigma_i, C_1 \cap C_2 = \emptyset, C_1 \cup C_2 \neq N\} \ \cup \\
 & & \{[C_1 \cup C_2](\psi_1 \wedge \psi_2) \mid [C_i]\psi_i \in \Sigma_i, C_1 \cap C_2 = \emptyset, C_1 \cup C_2 \neq N\} \\
\Sigma(\delta) & := & \bigcup_i \Sigma_i
\end{array}$$

Definition 3.14 (Alternative $\mathcal{F}_\delta^\mathcal{M}$) Let $\mathcal{M} = (S, E, V)$ be a coalition model and δ any formula in $\mathcal{L}$. Let $\Sigma_\mathcal{F} := \Sigma(\delta)$, then we define $\mathcal{F}_\delta^\mathcal{M} := (S^\mathcal{F}, E^\mathcal{F}, V^\mathcal{F})$ where

$$\begin{array}{lll}
(1) & S^\mathcal{F} = S_{\Sigma_\mathcal{F}}, & \\
(2) & |s|E_C^\mathcal{F} X \ \text{ iff } & \left\{ \begin{array}{ll} \exists [C]\varphi \in \Sigma_\mathcal{F} : \varphi^\mathcal{M} \subseteq \wr X \wr \text{ and } sE_C\varphi^\mathcal{M} & \text{for } C \neq N \\ \text{not } |s|E_\emptyset^\mathcal{F}\overline{X} & \text{for } C = N, \end{array} \right. \\
(3) & V^\mathcal{F}(p) = |V(p)| \quad \text{for all proposition letters } p \in \Sigma_\mathcal{F}. &
\end{array}$$

$\dashv$

It should be clear again that $\Sigma(\delta)$ is subformula closed and finite. Inspecting the proof of the filtration and playability conditions for this alternative $\mathcal{F}_\delta^\mathcal{M}$, shows that $\{[\emptyset]\neg\psi, \neg\psi \mid [N]\psi \in \mathit{sfor}(\delta)\}$ is needed to prove condition (iii) of definition 3.14 and maintain subformula-closure; $\{[C]\top \mid C \subsetneq N\} \subseteq \Sigma(\delta)$ is needed to prove playability conditions 1 and 2, and finally, the formulas in $\Sigma(\delta) \setminus \Sigma_0$ are the formulas needed to show superadditivity and also maintain subformula-closure. In the worst case, δ contains n different $[C_i]\psi_i$-formulas where all C_i are disjoint. Consider e.g. the formula $\delta = [1][2]\ldots[n]p$. A combinatorial analysis shows that the formulas which must be added to ensure that superadditivity can be proved is at most $\binom{n}{2} + \binom{n}{3} + \ldots + \binom{n}{n-1} = 2^n - (n+2)$. So with this definition of $\Sigma(\delta)$:

$$|\delta| + 2^n \leq |\Sigma(\delta)| \ \leq \ |\delta| + |\delta| + 2^n + 2^n - (n+2) \ = \ 2|\delta| + 2^{n+1} - (n+2)$$

Hence $|\Sigma(\delta)|$ is linear in $|\delta|$, but exponential in $|N| = n$, which means that the number of states in the filtrated model will be of the order $2^{|\delta|} \cdot 2^{2^n}$. If n is much smaller than $|\delta|$, the improvement is clearly substantial. However, if we do not allow n to be part of the expression of $|\Sigma(\delta)|$, then the best upper bound we can give on n is $|\delta| - 1$, consider again $\delta = [1][2]\ldots[n]p$, in which case we are back at the double-exponential bound.

For $\texttt{CLNC}_\texttt{N}$, the alternative definition 3.14 will not work, since we are unable to prove that $\mathcal{F}_\delta^\mathcal{C}$ satisfies condition (3) with this definition. The problem is that we need a formula $[N\backslash\{i\}]\psi_i \in \Sigma(\delta)$, where ψ_i is the formula provided by the definability lemma and the NC rule, to show that

$|s|E^{\mathcal{F}}_{N\backslash\{i\}}|\psi^{\mathcal{C}}_i|$ in the very last step of the proof of theorem 3.9. The definability lemma ensures that ψ_i is in $\Sigma(\delta)$, but it is certainly not clear that $[N\backslash\{i\}]\psi_i$ should also be in $\Sigma(\delta)$. The only way to ensure this, would be to close $\Sigma(\delta)$ under $[C]$-formulas, but since we can make infinitely many different sequences of $[C]$'s, we would no longer have a logically finite set, and hence no finite filtration.

4 Applications

As a simple illustration of the decentralisation of a constitution, consider the following example which is essentially Gibbard's paradox [3]: Abelard and Eloise each have a white and a blue shirt, and they have to decide which shirt to wear. Each person has the right to determine the colour of his/her own shirt. Using Coalition Logic, we can model this situation by letting $N = \{a, e\}$ and using two atomic proposition p_a and p_e, where p_a should be read as "Abelard wears white" and p_e as "Eloise wears white". The rights of Abelard and Eloise can then be captured by the following formula ρ^+ of Coalition Logic:

$$[a]p_a \wedge [a]\neg p_a \wedge [e]p_e \wedge [e]\neg p_e$$

Furthermore, we assume that these are the only rights which we want to give to them, formalised by formula ρ^-:

$$\bigwedge_{i\in\{a,e\}} \neg[i]((p_a \wedge p_e) \vee (\neg p_a \wedge \neg p_e)) \wedge \bigwedge_{i\in\{a,e\}} \neg[i]((p_a \wedge \neg p_e) \vee (\neg p_a \wedge p_e))$$

To ask whether this constitution can be decentralised or implemented by a mechanism means checking whether $\rho = \rho^+ \wedge \rho^-$ is satisfiable by a coalition model. The following coalition model $\mathcal{M} = (S, E, V)$ with $S = \{s_0, (w, w), (w, b), (b, w), (b, b)\}$ satisfies ρ at state s_0: $V(p_a) = \{(w, w), (w, b)\}$, $V(p_e) = \{(w, w), (b, w)\}$, and $E(s_0)$ is the α-effectivity function associated with the following strategic game:

	w	b
w	(w, w)	(w, b)
b	(b, w)	(b, b)

For states other than s_0, E can be defined arbitrarily. At state s_0, each player can choose which shirt to wear, and the resulting state reflects Abelard's choice in the first component and Eloise's choice in the second.

As it turns out, however, this game form is not Nash-consistent. Consider the situation where Abelard is primarily conformist, preferring to wear the same colour as Eloise, and besides that he prefers white to blue. Eloise on the other hand has the same colour preference, but she primarily is a non-conformist, wanting to avoid the same colour. The following game captures these preferences:

	w	b
w	$(4, 2)$	$(2, 3)$
b	$(1, 4)$	$(3, 1)$

As can easily be checked, the game has no Nash equilibrium. Intuitively, the mechanism is not stable under the given preferences. If Abelard and Eloise both wear white, Eloise will want to change to wearing blue, which will in turn induce Abelard to wear blue as well, which will lead to Eloise wearing white again, etc., precisely the kind of instability mentioned earlier. In fact, there is no stable mechanism guaranteeing each individual to choose their own shirt, i.e., it can be shown that the constitution given has no Nash-consistent representation [10]. Using Coalition Logic, this impossibility of a Nash-consistent representation can be formally deduced, for one can show that in a system $\Lambda \supseteq \texttt{CLNC}$, ρ is inconsistent: Since

$$((p_a \wedge p_e) \vee (\neg p_a \wedge \neg p_e)) \vee ((p_a \wedge \neg p_e) \vee (\neg p_a \wedge p_e))$$

is a propositional tautology, we can use the NC inference rule to derive

$$[a]((p_a \wedge p_e) \vee (\neg p_a \wedge \neg p_e)) \vee [e]((p_a \wedge \neg p_e) \vee (\neg p_a \wedge p_e))$$

which contradicts ρ^-.

To sum up, we have seen a very simple example of how a system of rights, represented by an effectivity function, can be decentralised in a strategic game form. Formalising the rights as coalition logic formulas, this decentralisation problem can be turned into a satisfiability problem, and this approach can be extended to cover Nash-consistent decentralisation. The logical reformulation of constitutional decentralisation has the advantage that it can easily handle partial specifications of rights. In the example of the two shirts, we have assumed that the players' rights were completely specified. This forced us to add a conjunct ρ^- which specified what the players were unable to bring about. In order to avoid such an additional conjunct, we would have to make use of some kind of non-monotonic reasoning mechanism which would allow us to conclude that the rights specified are *all* the rights the players are supposed to have. On the other hand, we may also treat ρ^+ as a partial specification, where we want to ensure that the players have the rights specified in ρ^+ but do not care about whether players also obtain additional rights. Checking satisfiability of ρ^+ thus corresponds to decentralising a partially defined effectivity function, a problem which does not seem to have received much attention in the social choice theory literature.

5 Conclusions

As we have shown in the previous section, Nash-consistency can play a role in the formalisation of strategic stability in multi-agent systems. To our knowledge, this paper presents the first attempt to express the notion of Nash-consistency in a formal logical system.

Based on adaptations of the characterisation results of Nash-consistent effectivity functions in [10] to the setting of Coalition Logic [8], we have extended the minimal coalition logic, $\mathtt{CL_N}$, with a new inference rule (NC), and the obtained coalition logic, $\mathtt{CLNC_N}$, is proved to be sound and weakly complete with respect to Nash-consistent coalition models. The main tool used in the completeness proof is a filtration of coalition models, which in turn provides the strong finite model property for $\mathtt{CL_N}$ and $\mathtt{CLNC_N}$.

The principal result of this paper is the axiomatisation of Nash-consistency purely in modal terms, i.e. without introducing preferences into the language. This is, partially, a consequence of the characterisation in [10] of Nash-consistency in terms of the effectivity function alone. But it is noteworthy, that the NC rule yields soundness and (weak) completeness with respect to arbitrary Nash-consistent coalition models, since the rule is a translation of the characterising condition (3) of Nash-consistency for effectivity functions over finite domains. The condition for Nash-consistency of arbitrary effectivity functions is a generalisation of (2) where $E^*(i)$ is replaced by $I(E^*(i))$, the closure of $E^*(i)$ under descending chains (Lemma 3.5 of [10]). This is a condition which is not (immediately) translatable into the language of Coalition Logic.

The main outstanding questions are whether a strong completeness result can also be obtained for $\mathtt{CLNC_N}$, whether we can improve the double-exponential bound given by the filtration, and which implications this has for the complexity of the satisfiability problem.

References

[1] J. Abdou and H. Keiding. *Effectivity Functions in Social Choice.* Kluwer, 1991.

[2] R. Alur, T. Henzinger, and O. Kupferman. Alternating-time temporal logic. In *Compositionality: The Significant Difference*, LNCS 1536, pages 23–60. Springer, 1998.

[3] A. Gibbard. A pareto-consistent libertarian claim. *Journal of Economic Theory*, 7(4):388–410, 1974.

[4] H. Moulin. *The Strategy of Social Choice.* North-Holland, 1983.

[5] H. Moulin and B. Peleg. Cores of effectivity functions and implementation theory. *Journal of Mathematical Economics*, 10(1):115–145, 1982.

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

[7] M. Pauly. *Logic for Social Software*. PhD thesis, University of Amsterdam, 2001.

[8] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[9] B. Peleg. Effectivity functions, game forms, games and rights. *Social Choice and Welfare*, 15:67–80, 1998.

[10] B. Peleg, H. Peters, and T. Storcken. Nash consistent representation of constitutions: A reaction to the Gibbard paradox. *Mathematical Social Sciences*, 43:267–287, 2002.

[11] T. Sandholm. Distributed rational decision making. In G. Weiss, editor, *Multiagent Systems: A Modern Introduction to Distributed Artificial Intelligence*, pages 201–258. MIT Press, 1999.