

Sequentially optimal repeated coalition formation under uncertainty

Georgios Chalkiadakis · Craig Boutilier

© The Author(s) 2010

Abstract Coalition formation is a central problem in multiagent systems research, but most models assume common knowledge of agent *types*. In practice, however, agents are often unsure of the types or *capabilities* of their potential partners, but gain information about these capabilities through repeated interaction. In this paper, we propose a novel Bayesian, model-based reinforcement learning framework for this problem, assuming that coalitions are formed (and tasks undertaken) repeatedly. Our model allows agents to refine their beliefs about the types of others as they interact within a coalition. The model also allows agents to make explicit tradeoffs between exploration (forming “new” coalitions to learn more about the types of new potential partners) and exploitation (relying on partners about which more is known), using value of information to define optimal exploration policies. Our framework effectively integrates decision making during repeated coalition formation under type uncertainty with Bayesian reinforcement learning techniques. Specifically, we present several learning algorithms to approximate the optimal Bayesian solution to the repeated coalition formation and type-learning problem, providing tractable means to ensure good sequential performance. We evaluate our algorithms in a variety of settings, showing that one method in particular exhibits consistently good performance in practice. We also demonstrate the ability of our model to facilitate knowledge transfer across different dynamic tasks.

Keywords Coalition formation · Multiagent learning · Bayesian reinforcement learning

G. Chalkiadakis (✉)

School of Electronics and Computer Science, University of Southampton, Southampton, UK
e-mail: gc2@ecs.soton.ac.uk

C. Boutilier

Department of Computer Science, University of Toronto, Toronto, ON, Canada
e-mail: cebly@cs.toronto.edu

1 Introduction

Coalition formation, widely studied in game theory and economics [3, 31, 35, 38, 45, 50], has attracted much attention in AI as means of dynamically forming partnerships or teams of cooperating agents [32–34, 48, 52]. Most models of coalition formation assume that the values of potential coalitions are known with certainty, implying that agents possess knowledge of the capabilities of their potential partners. However, in many natural settings, rational agents must form coalitions and divide the generated value without knowing a priori what this value may be or how suitable their potential partners are for the task at hand. In other cases, it is assumed that this knowledge can be reached via communication [51, 52]; but without strong mechanisms or contract conditions, agents usually have incentives to lie about (e.g., exaggerate) their capabilities to potential partners.

The presence of uncertainty presents opportunities for agents to learn about the capabilities of their partners if they interact repeatedly. The effects of collective actions provide valuable information about the capabilities of one's coalitional partners. This information should, naturally, impact future coalition formation decisions and choice of coalitional actions, thus refining how coalitions are formed over time. For example, extremely capable agents may find themselves in greater demand as their capabilities become known with greater certainty; and they may, over time, also be able to extract a larger share of the surplus generated by the coalitions in which they participate.

Examples of coalition formation under such uncertainty abound. The case of an enterprise trying to choose subcontractors (e.g., for building projects) while unsure of their capabilities and synergies is one such example. As projects are completed, information gleaned from the outcomes allows for refined assessment of the capabilities and value that specific subcontractors bring to the table, which in turn influences decisions regarding future project participation. The creation and interaction of *virtual organizations* has long been anticipated as a target of agent coalition technologies within e-commerce. While virtual organizations will allow institutions to come together dynamically to share resources and coordinate actions to accomplish common (or partially aligned) objectives, this cannot happen without some means of establishing coalitional terms under type uncertainty. We expect past experiences to influence the formation of future organizations, allowing institutions to select partners dynamically and switch allegiances as appropriate.

Realistic models of coalition formation must be able to deal with both uncertainty regarding the effects of potential coalitional actions and the capabilities of the potential partners. This uncertainty is translated into uncertainty about the values of various coalitions. In addition, learning mechanisms must be assumed to capture the fact that uncertainty is typically reduced as agents gain experience with one another. Finally, these models should reflect the fact that agents must make decisions about which coalitions to form, and which actions to take, knowing that information gained though the course of coalitional interaction will influence future decisions. Research in coalition formation to date has not dealt with the sequential decision making problem facing agents forming coalitions under such type uncertainty. To this end, we develop a model of sequential coalition formation under uncertainty that allows agents to take sequentially rational decisions regarding which coalitions to form and which coalitional actions to take. In our model, the *value* of a coalition is assumed to be a function of the *types* of its participants, where types can be viewed loosely as reflecting any relevant capabilities and qualities of team members.

As an illustrative example, consider a group of contractors, say plumbers, electricians, and carpenters, each possessing trade-specific skills of various degrees corresponding to their *types*; e.g., a carpenter might be highly skilled or moderately incompetent. The contractors

repeatedly come together to collaborate on various construction projects. Any group of contractors that joins together to form a coalition will receive a payoff for the house they build. The payoff for received for a house depends on the type of project undertaken and its resulting quality, which in turn depends on the quality of each team member and potential synergies or conflicts among them. We assume that agents are uncertain about the types of potential partners, but that their *beliefs* are used to determine a distribution over coalitional outcomes and expected coalitional value. It is these beliefs that influence the coalition formation process and the stability of any coalition structure that results. Each coalition must also decide which collective action to take. For instance, a team of contractors may have a choice of what type of housing project to undertake (e.g., a high-rise in Toronto or a townhouse estate in Southampton). The outcome of such *coalitional actions* is *stochastic*, but is influenced by the types of agents in the coalition. This too plays a key role in determination of coalition value. Deliberations about team formation are complicated by the fact that uncertainty about partner types influences coalitional actions decisions (e.g., what type of house to build) and payoff division (e.g., how to split the revenue generated). Our model can incorporate a variety of (say, negotiation or equilibrium-based) mechanisms for determining coalitional stability, value, and agent bargaining/payoff division.

In our model, agents come together repeatedly in *episodes*, during which they can form new coalitions and take coalitional actions. In our construction example, for instance, after one set of housing projects is completed, the agents have an opportunity to regroup, forming new teams. Of course, the outcomes of previous coalitional actions provide each agent with information about the types of its previous partners. In our example, receiving a high price for a house may indicate to a plumber that the electrician and carpenter she partnered with were highly competent. Agents update their beliefs about their partners based on those prior outcomes and use these updated beliefs in their future coalitional deliberations. For example, an agent may decide to abandon its current partners to join a new group that she *believes* may be more profitable.

We propose a *Bayesian reinforcement learning (RL)* model that enables agents to make better decisions regarding coalition formation, coalitional action, and bargaining using experience gained by these repeated interaction with others. The critical *exploration–exploitation tradeoff* in RL is embodied in the tension between forming teams with partners about which types are known with a high degree of certainty (e.g., stay in one’s current coalition) or forming teams with partners about whom much less is known in order to learn more about these new partners’ abilities. Our Bayesian RL model allows this tradeoff to be made optimally by relying on the concept of *value of information*. We develop a partially observable Markov decision process (POMDP) formulation of our Bayesian RL model. The solution of this POMDP determines agent policies that value actions not just for their immediate gains, but also because of the information they provide about the types of others and the values of potential coalitions.

Since the solution of POMDPs is computationally intractable, we develop several computational approximations to allow for the more effective construction of sequential policies. We investigate these approximations experimentally and show that our framework enables agents to make informed, rational decisions about coalition formation and coalitional actions that are rewarding in both the short and long term. This is true even if agents do not converge to “stable” coalitions in the end of a series of coalition formation episodes. We also demonstrate that our model allows for effective *transfer of knowledge* between tasks: agents that learn about their partners’s abilities are able to re-use this knowledge when encountering those partners in different circumstances.

This paper focuses on the online behaviour of agents that learn by observing the results of coalitional actions, that is, actions that are agreed upon during coalition formation episodes and executed upon such an episode's completion. We do not focus here on the negotiation processes that determine the coalitions formed in each episode, nor on the strategic considerations of agents during bargaining. However, our Bayesian RL model is fully general, and it allows for the incorporation of any potential bargaining process that might be used to determine coalitional structure in each episode. For this reason, we do not model or analyze agent behaviour during the repeated interaction as a game. Indeed, our repeated coalition formation problem under uncertainty could formally be modeled as an infinite-horizon *Bayesian extensive form game (BEFG)*, in which the coalitional negotiations among agents at each episode are explicitly modelled. An appropriate solution concept for such a game is a *perfect Bayesian equilibrium (PBE)* [36], in which agents adopt behavioural strategies that are optimal at all subgames with respect to their own beliefs and the strategies adopted by opponents, and beliefs are determined by Bayesian updates with respect to these behavioural strategies. Unfortunately, due to the size of the beliefs/strategies space, obtaining a PBE solution is a practically infeasible task [13]. The POMDP approximation methods we propose in this work can be viewed as heuristic approximations of the solution of the corresponding BEFG (though this comes without any bounds or guarantees regarding the PBE-optimality of an agent's policy). We elaborate on these issues later in the paper.

The paper is structured as follows. We begin with a brief review of coalition formation and Bayesian reinforcement learning in Sect. 2. In Sect. 3 we describe a generic Bayesian coalition formation model, detail our Bayesian RL framework for optimal repeated coalition formation under uncertainty, and describe its POMDP formulation. We also motivate the use of this approximation to the strategic behaviour that arise in a Bayesian extensive form game formulation. In Sect. 4 we present several Bayesian RL algorithms that approximate the solution of the POMDP, and in Sect. 5 explain how our RL algorithms can be combined with different negotiation processes for coalition formation. We evaluate our algorithms experimentally in Sect. 6 and compare our approach with related work in Sect. 7. We conclude in Sect. 8 with a summary and discussion of future directions. Earlier versions of some aspects of this research were presented in [12, 14].

2 Background

We begin with background on coalition formation and Bayesian reinforcement learning. A deeper discussion of related work is found in Sect. 7.

2.1 Coalition formation

Cooperative game theory deals with situations where players act together in a cooperative equilibrium selection process involving some form of bargaining, negotiation, or arbitration [38]. The problem of *coalition formation* is one of the fundamental areas of study within cooperative game theory.

Let $N = \{1, \dots, n\}$, $n > 2$, be a set of players (or “agents”). A subset $C \subseteq N$ is called a *coalition*, and we assume that agents participating in a coalition will coordinate their activities for mutual benefit.¹ A *coalition structure (CS)* is a partition of the set of agents containing exhaustive and disjoint coalitions. Coalition formation is the process by which individual

¹ Seeking “mutual benefit” does not imply that the agents are not *individually rational*—i.e., seeking to maximize their own individual payoffs by participating in coalitions. This will become more evident shortly.

agents form such coalitions, generally to solve a problem by coordinating their efforts. The *coalition formation problem* can be seen as being composed of the following activities [48]: (a) the search for an optimal coalition structure; (b) the solution of a joint problem facing members of each coalition; and (c) division of the value of the generated solution among the coalition members.

While seemingly complex, coalition formation can be abstracted into a fairly simple model under the assumption of *transferable utility*, which assumes the existence of a (divisible) commodity (such as “money”) that players can freely transfer among themselves. Thus, it is easy to describe the possible *allocations* of utility among the members of each coalition, as it is sufficient to specify a single number denoting its *worth* (i.e., the total payoff available for division among its members).

This is the role of the *characteristic function* of a *coalitional game with transferable utility* (*TU-game*): A characteristic function $v : 2^N \Rightarrow \Re$ defines the *value* $v(C)$ of each coalition C [60]. Intuitively, $v(C)$ represents the maximal payoff the members of C can jointly receive by cooperating effectively. An *allocation* is a vector of payoffs (or “demands”) $\mathbf{d} = (d_1, \dots, d_n)$ assigning some payoff to each $i \in N$. An allocation is *feasible* with respect to coalition structure CS if $\sum_{i \in C} d_i \leq v(C)$ for each $C \in CS$, and is *efficient* if this holds with equality. The *reservation value* rv_i of an agent i is the amount it can attain by acting alone (in a *singleton* coalition): $rv_i = v(\{i\})$.

One important concept regarding characteristic functions is the concept of superadditivity. A characteristic function is called *superadditive* if any pair (C, T) of disjoint coalitions C and T is better off by merging into one coalition: $v(C \cup T) \geq v(C) + v(T)$. Since superadditivity is unrealistic in many real-world applications, we do not assume it in our work. When transferable utility is not assumed, we lie in the realm of *non-transferable utility* (*NTU*) games [38]. We do not deal with NTU games in this work.

When rational agents seek to maximize their individual payoffs, the *stability* of the underlying coalition structure becomes critical. Intuitively, a coalition structure is stable if the outcomes attained by the coalitions and agreed-upon payoffs are such that both individual and group rationality are satisfied. Research in coalition formation has developed several notions of stability, among the strongest being the *core* [28, 31, 35, 48, 24].

Definition 1 The *core* of a characteristic function game is the set of coalition structures and payoff configuration pairs:

$$\left\{ \langle CS, \mathbf{d} \rangle \mid \forall C \subseteq N, \sum_{i \in C} d_i \geq v(C) \text{ and } \sum_{i \in N} d_i = \sum_{C \in CS} v(C) \right\}$$

A core allocation $\langle CS, \mathbf{d} \rangle$ is both feasible and efficient, and no subgroup of players can guarantee all of its members a higher payoff. As such, no coalition would ever “block” the proposal for a core allocation. Unfortunately, in many cases the core is empty, as there exist games for which it is impossible to divide utility to ensure the coalition structure is stable (i.e., there might always be alternative coalitions that could gain value if they were given the opportunity to negotiate). Moreover, computing the core or even deciding its non-emptiness is, in general, intractable [17, 23, 44, 47].

Other cooperative solution concepts include the *kernel* [18], a stability concept that combines individual rationality with group rationality by offering stability within a *given* coalition structure (and under a given payoff allocation). The kernel is a payoff configuration space in which each payoff configuration $\langle CS, \mathbf{d} \rangle$ is stable in the sense that any pair of agents i, j belonging to the same coalition $C \in CS$ are *in equilibrium with one another*, given payoff

vector \mathbf{d} . Agents i and j are said to be in equilibrium if they cannot outweigh one another within their common coalition—in other words, neither of them can successfully claim a part of the other's payoff under configuration (CS, \mathbf{d}) . The kernel is always non-empty. In particular, for every CS for which there exists at least one allocation \mathbf{y} such that all agents receive at least their reservation value in \mathbf{y} , there also exists an allocation \mathbf{d} such that the resulting configuration is in the kernel (we say that it is *kernel-stable*). Blankenburg et al. [9] have recently introduced the *fuzzy kernel stability* concept for use in fuzzy cooperative games under coalitional value uncertainty.

In recent years, extensive research has covered many aspects of the coalition formation problem. Dynamic coalition formation research in particular is interested in the question of establishing endogenous processes by which agents form coalitions that reach stable structures, such as the core. Dieckmann and Schwalbe [24] recognize the need to deal with dynamic coalition formation processes, combining questions of stability with the explicit monitoring of the process by which coalitions form. They describe a *dynamic process* of coalition formation, in which agents are given, at random, the opportunity to abandon or join existing coalitions and demand a certain payoff. At each stage of the process, a given coalition configuration (CS, \mathbf{d}) prevails. With some specified small probability γ , any player may independently decide which of the existing coalitions to join, and states a (possibly different) payoff demand for himself. A player will join a coalition iff it is in her best interest to do so. These decisions are determined by a non-cooperative best-reply rule, given the coalition structure and allocation prevailing at the beginning of the period: a player switches coalitions if her expected payoff in the new coalition exceeds her current payoff; and she demands the most she can get subject to feasibility. The players observe the coalitional structure and the demands of the other agents in the beginning of the period, and expect the current coalition structure and demand to prevail in the next period—which is not unrealistic if γ is small. It is assumed that coalitions so formed persist (i.e., continue to participate in the process until the end of all bargaining rounds). The process allows for *experimentation*: agents can explore suboptimal coalition formation actions as well.

The process in which all players adopt the best-reply rule induces a finite Markov chain with at least one absorbing state. If the players *explore* with myopically suboptimal actions, Dieckmann and Schwalbe prove that if the core is non-empty, each core allocation corresponds to an absorbing state of the resulting *best reply with experimentation (BRE)* process, and each absorbing state of this process can be associated with a core allocation. Moreover, the process converges to a core allocation with probability 1 (if the core is non-empty). However, Dieckmann and Schwalbe's model does not explicitly allow for the agents to suggest and agree on coalitional actions to perform. Moreover, it adopts the usual assumption of full information regarding coalitional values. Their work is influenced by the work of Agastya [2], which is, unlike [24], confined to superadditive environments.

Suijs et al. [56, 57] introduce *stochastic cooperative games (SCGs)*, comprising a set of agents, a set of coalitional actions, and a function assigning to each action a random variable with finite expectation, representing the payoff to the coalition when this action is taken. Thus uncertainty in coalitional value is present. To accommodate stochastic payoffs, they use *relative shares* for the allocation of the residual of the stochastic coalitional values, and make the—in some cases unrealistic—assumption that agents have *common expectations* regarding expected coalitional values; thus, the degree of partial information permitted in this model is quite limited. This work provides strong theoretical foundations for games with this restricted form of uncertainty, and describes classes of games for which the core of an SCG is non-empty. No explicit coalition formation process is assumed. Also, no assumption of incomplete information about partners's types is made, and thus there is no direct

translation of type uncertainty into coalition value uncertainty. However, [57] discusses the effect that different risk behaviour on the part of agents might have on the existence of a core allocation within a specific class of SCG games.

Chalkiadakis et al. [10, 12, 15] define the concept of the *Bayesian core (BC)* to describe stability under *type uncertainty* in cooperative games with stochastic coalitional actions. Specifically, they examine properties of three variants of the Bayesian core concept: the *strong*, the *weak* and the *strict* BC. Intuitively, the BC is the set of coalition-structure, demand-vector, agent-belief triples that are stable under the agents's private probabilistic beliefs regarding the types (capabilities) of their potential partners. They also extend Dieckmann and Schwalbe's BRE process to uncertain environments, guaranteeing convergence to the (strong) BC if it is non-empty. Unlike [24], this process allows the agents to explicitly propose and agree to actions to be performed by the coalitions that are formed. We refer to [10, 12, 15] for further details regarding the BRE process and the Bayesian core. Since the underlying Bayesian coalition formation problem introduced there is the one we adopt, we briefly define the BC in the next section when we present our repeated coalition formation model. Our algorithms are, however, orthogonal to the specific means by which coalitions are formed and the stability concept used (if indeed, one is required at all). Nevertheless, in some of the experiments in this paper, we use this BRE process as the negotiation process to illustrate the performance of our Bayesian RL framework for repeated coalition formation.

2.2 Bayesian reinforcement learning

Consider an agent learning to control a stochastic environment modeled as a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, R, D \rangle$, with finite state and action sets \mathcal{S}, \mathcal{A} , reward function R , and transition dynamics D . D refers to a family of transition distributions $\Pr(s, a, \cdot)$, and $\Pr(s, a, s')$ is the probability of reaching state s' after taking action a at s . The probability with which reward r is obtained when state s is reached after executing a , is denoted $R(s, a, r)$. The agent has to construct an optimal Markovian policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ maximizing the expected sum of future discounted rewards over an infinite horizon. This policy, and its value, $V^*(s)$ at each $s \in \mathcal{S}$, can be computed using standard algorithms, such as value or and policy iteration [43, 58].

In the *reinforcement learning* setting, an agent does not have direct access to D and/or R , so it must learn a policy based on its interactions with the environment. While striving to do so, it has to face the well-known *exploration–exploitation tradeoff*: should one *exploit* what is already known by following a policy that currently appears best, or should one *explore*, that is, try different actions in order to gain further information about rewards R and dynamics D , and thus potentially revise its view of the optimality of available actions? If the underlying uncertainty is not properly accounted for, agents risk exploring very unrewarding regions of policy space.

When *model-based RL* is used, the agent maintains an estimated MDP $\langle \mathcal{S}, \mathcal{A}, \hat{R}, \hat{D} \rangle$, based on the set of experiences $\langle s, a, r, t \rangle$ obtained so far; an experience tuple $\langle s, a, r, t \rangle$ describes the reward r and transition to state t experienced by the agent when taking an action a while at state s . At each stage (or at suitable intervals) this MDP can be solved (or approximated). Single-agent Bayesian methods [20, 21, 25, 42, 49] assume some prior density P over all possible dynamics models D and reward functions R , which is updated with past experiences. By acting optimally (in a sequential sense), Bayesian RL methods allow agents to make the exploration–exploitation tradeoff appropriately, providing a framework for *optimal learning*—acting so as to maximize performance while learning.

More specifically, assume a prior density P over D, R reflecting an agent's *belief state* regarding the underlying model. Letting H denote the (current) state-action history of the observer, we can use the posterior $P(D, R|H)$ to determine an appropriate action choice at each stage. The formulation of [20] renders this update by Bayes rule tractable by assuming a convenient prior. Specifically, the following assumptions are made: (a) the density P is factored over R and D , with $P(D, R)$ being the product of independent local densities $P(D^{s,a})$ and $P(R^{s,a})$ for each transition and each reward distribution; and (b) each density $P(D^{s,a})$ and $P(R^{s,a})$ is Dirichlet [22]. The choice of Dirichlet is appropriate assuming discrete multinomial transition and rewards models, for which Dirichlet priors are conjugate. As a consequence, the posterior can be represented compactly: after each observed experience tuple, the posterior is also a Dirichlet. In this way, the posterior $P(D|H)$ over transition models required by the Bayesian approach can be factored into posteriors over local families, each of the form:

$$P(D^{s,a}|H^{s,a}) = z \Pr(H^{s,a}|D^{s,a})P(D^{s,a})$$

where $H^{s,a}$ is the history of s, a -transitions—captured by updates of the Dirichlet parameters—and z is a normalizing constant. Similarly,

$$P(R^{s,a}|H^{s,a}) = z \Pr(H^{s,a}|R^{s,a})P(R^{s,a}).$$

To model $P(D^{s,a})$, a Dirichlet parameter vector $\mathbf{n}^{s,a}$ is used, with entries $n^{s,a,s'}$ for each possible successor state s' ; similarly, to model $P(R^{s,a})$ a parameter vector $\mathbf{k}^{s,a}$ is used, with entries $k^{s,a,r}$ for each possible reward r . The expectation of $\Pr(s, a, s')$ with respect to P is given by $n^{s,a,s'}/\sum_i n^{s,a,s_i}$. Updating a Dirichlet is straightforward: given prior $P(D^{s,a}; \mathbf{n}^{s,a})$ and data vector $\mathbf{c}^{s,a}$ (where c^{s,a,s_i} is the number of observed transitions from s to s_i under a), the posterior is given by parameter vector $\mathbf{n}^{s,a} + \mathbf{c}^{s,a}$. Thus, the Bayesian approach allows for the natural incorporation of prior knowledge in the form of a prior probability distribution over all possible MDPs, and admits easy update. In a similar fashion, multi-agent Bayesian RL agents [11] update prior distributions over the space of possible strategies of others in addition to the space of possible MDP models.

In a fully observable MDP, the value of an action is a function of both the immediate reward it provides and the expected state transition, which dictates the opportunity to accrue *future* value. In a belief state MDP, the “state transition” includes both a transition in the underlying state space as well as an update to the belief state. Thus the value of performing an action at a belief state can implicitly be divided into two components: the expected value given the current belief state and the value of the action's impact on the current belief state. The second component captures the *expected value of information (EVOI)* of an action. Each action gives rise to some immediate response by the environment changing the agent's beliefs, and subsequent action choice and expected reward is influenced by this change. EVOI need not be computed directly, but can be combined with “object-level” expected value via Bellman equations. This can be viewed as the solution of a partially observable MDP (POMDP) or equivalently, the belief state MDP. A number of prior studies [11, 20, 21, 41] have demonstrated the practical value of the Bayesian approach, and the effectiveness of related approximation algorithms, in allowing exploration costs to be weighed against their expected benefits. This leads to informed, intelligent exploration, and better online performance while learning than offered by other RL exploration models.

3 A Bayesian RL framework for repeated coalition formation under uncertainty

Our goal is to develop a framework for modeling situations in which sets of agents come together repeatedly to form coalitions. Agents participating in coalition formation activities will generally face two forms of uncertainty: (a) type uncertainty, i.e., uncertainty regarding the *types* (or capabilities) of potential partners; and (b) uncertainty regarding the results of coalitional actions. Unlike the case in one-shot coalitional settings, the potential for repeated interaction provides agents with an opportunity to *learn* about both the abilities of their partners and the nature of coalitional actions over time. This opens up the possibility that rational agents might explicitly take actions that reduce specific type or action uncertainty rather than try to optimize the myopic value of the “next” coalition they join. This is, of course, nothing more than the *exploration–exploitation* tradeoff faced by any reinforcement learning agent.

To this end, our model for *optimal repeated coalition formation* brings together coalition formation under uncertainty (specifically the Bayesian coalitional model proposed in [12, 15]) with Bayesian reinforcement learning to properly capture the both aspects of this learning process and the interactions that arise between them.

To capture the “stage games” within which agents form coalitions at each stage of the RL process, in Sect. 3.1 we review the model of Bayesian coalition formation introduced in [12, 15]. We then describe the full reinforcement learning model in Sect. 3.2. While we briefly discuss certain stability concepts for Bayesian coalition formation, we note that our Bayesian RL framework is largely independent of the means by which coalitions are formed, relying only on the Bayesian formulation of the coalition problem and certain assumptions about the form of coalitional agreements.

3.1 A Bayesian model for cooperative games under uncertainty

The need to address type uncertainty, one agent’s uncertainty about the abilities of its potential partners, is critical to the modeling of realistic coalition formation problems. For instance, if a carpenter wants to find a plumber and electrician with whom to build a house, her decision to propose (or join) such a partnership, to engage in a specific type of project, and to accept a specific share of the surplus generated should all depend on her (probabilistic) assessment of their abilities. To capture this, we start by introducing the problem of Bayesian coalition formation under type uncertainty. We then show how this type uncertainty can be translated into coalitional value uncertainty.

We adopt the model proposed in [12, 15]. A *Bayesian coalition formation problem* under type uncertainty is a cooperative game defined as follows:

Definition 2 (*Bayesian coalition formation problem* [12, 15]) A Bayesian coalition formation problem (BCFP) is a coalition formation problem that is characterized by a set of agents, N ; a set of types T_i for each agent $i \in N$; a set A_C of coalitional actions for each coalition $C \subseteq N$; a set \mathcal{O} of stochastic outcomes (or states); with transition dynamics $Pr(s|\alpha_C, t_C)$ denoting the probability of an outcome $s \in \mathcal{O}$ given that coalition C whose members have type vector t_C takes coalitional action α_C ; a reward function $R : \mathcal{O} \rightarrow \mathfrak{R}$; and agent beliefs B_i for each agent $i \in N$ comprising a joint distribution over types T_{-i} of potential partners.

We now describe each of the BCFP components in turn: we assume a set of agents $N = \{1, \dots, n\}$, and for each agent i a finite set of possible *types* T_i . Each agent i has a specific type $t \in T_i$, which intuitively captures i ’s “abilities”. An agent’s type is private information.

We let $T = \times_{i \in N} T_i$ denote the set of type profiles. For any coalition $C \subseteq N$, $T_C = \times_{i \in C} T_i$, and for any $i \in N$, $T_{-i} = \times_{j \neq i} T_j$. Each i knows its own type t_i , but not those of other agents. Agent i 's beliefs B_i comprise a joint distribution over T_{-i} , where $B_i(t_{-i})$ is the probability i assigns to other agents having type profile t_{-i} . We use $B_i(t_C)$ to denote the marginal of B_i over any subset C of agents, and for ease of notation, we let $B_i(t_i)$ refer to i "beliefs" about its own type (assigning probability 1 to its actual type and 0 to all others). We may assume that the prior B_i is derived from a common knowledge B over T conditioned agent i 's true type, but this is not critical in this paper (but see [15]).

A coalition C has available to it a finite set of *coalitional actions* A_C . We can think of A_C as the set of decisions available to C on how to deal with the underlying task at hand—or even a decision on what task to deal with. When an action is taken, it results in some outcome or *state* $s \in \mathcal{O}$. The odds with which an outcome is realized depends on the types of the coalition members (e.g., the outcome of building a house will depend on the capabilities of the team members). We let $\Pr(s|\alpha, t_C)$ denote the probability of outcome s given that coalition C takes action $\alpha \in A_C$ and member types are given by $t_C \in T_C$.² This probability is assumed to be known by all agents. Our model can be generalized to allow uncertainty over the action dynamics: for example, agents may have Dirichlet priors over the probabilities of each outcome, which could be updated in the standard Bayesian fashion given observed outcomes (and influenced by estimated types of its partners). This would make our model more like standard single-agent RL models. However, we ignore such action uncertainty in order to simplify the presentation and focus purely on the impact of type learning on coalition formation. Finally, we assume that each state s results in some *reward* $R(s)$. If s results from a coalitional action, the members are assigned $R(s)$, which is assumed to be divisible/transferable among them.

We illustrate the basic formulation with a simple, partial example. Consider a three-agent scenario with one carpenter (agent 1) and two electricians (agents 2 and 3). The electricians and carpenters have three types—good (g), medium (m) and bad (b)—and each agent has beliefs about the others. For simplicity, we focus on the beliefs B_1 of the carpenter, who believes agent 3 is somewhat more competent than agent 2:

Type	Agent 2			Agent 3		
	g	m	b	g	m	b
B_1	0.3	0.4	0.3	0.5	0.3	0.2

The true types of each agent will be private information. Let's assume the true type of the carpenter is $t_1 = g$, and suppose that a coalition consisting of a carpenter and an electrician can undertake an ambitious or a moderate housing project. The probabilities of success or failure are dictated by the type-vector of the pair, for example (listing only the outcome probabilities for carpenter type $t_c = g$ for brevity):

² The model can be extended to allow action effects to depend on not just the types of the coalition members, but on other factors as well. Dependence on the actions taken by other coalitions, for example, would induce a stage game of incomplete information between coalitions. Dependence on the state of the environment (e.g., as dictated by the outcomes of prior actions) would require modeling the environment as a Markov decision process. We do not consider such extensions here in order to isolate the problem of learning in repeated coalition formation.

Type vector (t_c, t_e)	(g, g)		(g, m)		(g, b)	
	P(succ)	P(fail)	P(succ)	P(fail)	P(succ)	P(fail)
Ambitious	0.8	0.2	0.4	0.6	0.1	0.9
Moderate	0.9	0.1	0.8	0.2	0.7	0.3

Finally assume a successful ambitious project has coalitional reward 1000, a successful moderate project has reward 500, and any failed project has reward 0.

Now we turn to the problem of showing how type (and action) uncertainty in a BCFP can be translated into coalitional value uncertainty. In a BCFP, the (immediate) *value* of coalition C with members of type t_C is:

$$V(C|t_C) = \max_{\alpha \in A_C} \sum_s \Pr(s|\alpha, t_C)R(s) = \max_{\alpha \in A_C} Q(C, \alpha|t_C) \tag{1}$$

where, intuitively, $Q(C, \alpha|t_C)$ represents the value (or quality) of coalitional action α to coalition C that is made up of members with types t_C . $V(C|t_C)$ therefore represents the (maximal) payoff that coalition C can obtain by choosing the best coalitional action. Unfortunately, this coalition value cannot be used in the coalition formation process if the agents are uncertain about the types of their potential partners (since any potential partners may have one of several types, any agent in any C would be uncertain about the type profile t_C of its members, and thus about the value $V(C)$). However, each agent i has beliefs about the (immediate, or myopic) value of any coalition based on its expectation of this value with respect to other agents’s types:

$$V_i(C) = \max_{\alpha \in A_C} \sum_{t_C \in T_C} B_i(t_C)Q(C, \alpha|t_C) = \max_{\alpha \in A_C} Q_i(C, \alpha) \tag{2}$$

where, intuitively, $Q_i(C, \alpha)$ represents the expected value (or, expected quality) of α to coalition C , according to i ’s beliefs. Note that $V_i(C)$ is not simply the expectation of $V(C)$ with respect to i ’s belief about types. The expectation Q_i of action values (i.e., Q -values) cannot be moved outside the max operator: a single action must be chosen which is useful *given* i ’s uncertainty. Of course, i ’s estimate of the value of a coalition, or any coalitional action, may not conform with those of other agents (e.g., i may believe that k is competent, while j may believe that k is incompetent; thus, i will believe that coalition $\langle i, j, k \rangle$ has a much higher value than j does). However, i is certain of its *reservation value*, the amount it can attain by acting alone: $rv_i = V_i(\{i\}) = \max_{\alpha \in A_{\{i\}}} \sum_s \Pr(s|\alpha, t_i)R(s)$.

In our example above, agent 1’s beliefs about 3’s type inform her beliefs about the immediate, or one-shot, values of partnering with agent 3. Specifically, the expected values of ambitious and moderate projects with agent 3 can be computed as follows:

$$Q_1(\{1, 3\}, \textit{Ambitious}) = (0.8 \cdot 0.5 + 0.4 \cdot 0.3 + 0.1 \cdot 0.2)1000 = 540 \tag{3}$$

$$Q_1(\{1, 3\}, \textit{Moderate}) = (0.9 \cdot 0.5 + 0.8 \cdot 0.3 + 0.7 \cdot 0.2)500 = 415 \tag{4}$$

Hence, given these quality values, agent 1 would probably want to engage in an ambitious project with 3 in a one-shot problem. (Notice that these must dominate the expected values of partnering with 2 given 1’s beliefs.) However, the final evaluation of such a partnership decision would have to take into account the agreed allocation within a coalition, as will now become apparent.

Because of the stochastic nature of payoffs in BCFPs, we assume that agents join a coalition with certain *relative payoff demands* [56, 57]. Intuitively, since action uncertainty means

agents cannot predict coalition payoff (and, consequently, the payoff shares to coalition members) with certainty, it is natural to place relative demands on the fractional share of the realized payoff. This directly accounts for the allocation of unexpected gains or losses. Formally, let \mathbf{d} represent the *payoff demand vector* $\langle d_1, \dots, d_n \rangle$, and \mathbf{d}_C the subset of these demands corresponding to agents in coalition C , and assume that these demands are observable by all agents. For any $i \in C$ we define the *relative demand* of agent i to be $r_i = \frac{d_i}{\sum_{j \in C} d_j}$. If reward R is received by coalition C as a result of its choice of action, each i receives payoff $r_i R$. This means that the gains or losses deriving from the fact that the reward function is stochastic will be allocated to the agents in proportion to their agreed upon demands. As such, each agent has beliefs about any other agent's expected payoff given a coalition structure and demand vector. Specifically, i 's beliefs about the (maximum) *expected stochastic payoff* of some agent $j \in C$ is denoted $\bar{p}_j^i = r_j V_j(C)$. Similarly, if $i \in C$, i believes its *own* (maximum) expected payoff to be $\bar{p}_i^i = r_i V_i(C)$.

If all agents in coalition C had the same beliefs about the expected value of the available actions A_C , then all would agree to execute the action with maximum expected value (assuming risk neutrality). However, agents enter coalitions with potentially different beliefs about the types of their partners. Since the expected value of action α predicted by agent $i \in C$ depends critically on i 's beliefs B_i , each agent may have different estimates of the expected value of any coalitional action.³ Hence, the *choice* of action must also be part of the negotiated agreement. Given a coalition structure CS , an *action vector* is a tuple consisting of one action $\alpha \in A_C$ for each $C \in CS$. To this end, we define a *coalition agreement vector* to be a triple $\langle CS, \mathbf{d}, \boldsymbol{\alpha} \rangle$ where CS is a coalition structure, \mathbf{d} is a demand vector, and $\boldsymbol{\alpha}$ is an action vector, with C_i denoting the $C \in CS$ of which i is a member (and let \mathbf{r} be the relative demand vector corresponding to \mathbf{d}).

The stability of a vector of coalitional agreements can be defined in several different ways. While the specific stability concept used is not critical to the repeated coalition formal model we develop below, we briefly two forms of the *Bayesian core*, a stability concept introduced for BCFPs in [12, 15].

Definition 3 (*weak Bayesian core* [15]) Let $\langle CS, \mathbf{d}, \boldsymbol{\alpha} \rangle$ be a coalition agreement vector, with C_i denoting the $C \in CS$ of which i is a member. $\langle CS, \mathbf{d}, \boldsymbol{\alpha} \rangle$ (or equivalently $\langle CS, \mathbf{r}, \boldsymbol{\alpha} \rangle$) is in the *weak Bayesian core* of a BCFP iff there is no coalition $S \subseteq N$, demand vector \mathbf{d}_S and action $\beta \in A_S$ s.t. $\bar{p}_i^i(S, \mathbf{d}_S, \beta) > \bar{p}_i^i(C_i, \mathbf{d}_{C_i}, \alpha_{C_i}), \forall i \in S$, where $\mathbf{d}_{C_i}, \alpha_{C_i}$ is the restriction of $\mathbf{d}, \boldsymbol{\alpha}$ to the C_i coalition.

In words, there is no coalition such that all of its members believe that they would be strictly better off in it (in terms of expected payoffs, given some choice of action) than they are in CS . The agents's beliefs, in every $C \in CS$, "coincide" in the weak sense that there is a payoff allocation \mathbf{d}_C and some coalitional action α_C that is commonly believed to ensure a better payoff. This doesn't mean that \mathbf{d}_C and α_C is what each agent believes to be best. But an agreement on \mathbf{d}_C and α_C is enough to keep any other coalition S from forming. Even if one agent proposed its formation, others would disagree because they would not expect to become strictly better off themselves.

A stronger notion can be defined as well:

³ To focus on learning of agent types over time, we assume that the distribution of outcomes given action α and the reward function are known. In model-based RL these too are uncertain. The definition of a BCFP is easily generalized to allow for unknown action models, which would add a further source of discrepancy among the beliefs of agents.

Definition 4 (*strong Bayesian core* [15]) Let $\langle CS, \mathbf{d}, \boldsymbol{\alpha} \rangle$ be a coalition agreement vector, with C_i denoting the $C \in CS$ of which i is a member. $\langle CS, \mathbf{d}, \boldsymbol{\alpha} \rangle$ (or equivalently $\langle CS, \mathbf{r}, \boldsymbol{\alpha} \rangle$) is in the *strong Bayesian core* iff there is no coalition $S \subseteq N$, demand vector \mathbf{d}_S and action $\beta \in A_S$ s.t. for some $i \in S$

$$\bar{p}_i^i(S, \mathbf{d}_S, \beta) > \bar{p}_i^i(C_i, \mathbf{d}_{C_i}, \alpha_{C_i})$$

and

$$\bar{p}_j^j(S, \mathbf{d}_S, \beta) \geq \bar{p}_j^j(C_j, \mathbf{d}_{C_j}, \alpha_{C_j})$$

$\forall j \in S, j \neq i$.

The strong BC is more tightly linked to an agent's subjective view of the potential acceptability of their proposals and is thus more "endogenous" in nature.

3.2 Optimal repeated coalition formation under uncertainty

We now turn our attention to the problem of acting and learning in repeated coalition formation settings.

3.2.1 A model for repeated coalition formation

The learning process proceeds in stages. Intuitively, at each stage agents come together with only partial knowledge of the types of their counterparts. They engage in some coalition formation process (e.g., negotiation or bargaining), forming various teams, each governed by a coalition agreement (i.e., agreement on the action to be taken by the team as well as the relative payoff share for each team member). Once the coalitions have completed their actions and observed the outcomes, each agent gains some information about the members of its team. Specifically, the action outcome provides (noisy) evidence about the coalition type vector to each member of that coalition. Agent's update their beliefs, and then enter the next stage of coalition formations.

More formally, the process can be described as follows: we assume an infinite horizon model in which a set of agents N faces a Bayesian coalition formation problem at each stage $0 \leq t < \infty$. The BCFP at each stage is identical except that, at stage t , each agent i may enter the coalition formation process with *updated* beliefs B_i^t that reflect its past interactions with previous partners.⁴ Each agent i enters the coalition formation process with beliefs B_i^t about the types of all agents (including the certain knowledge of its own type). Coalitions are formed, resulting in a coalition agreement vector $\langle CS^t, \mathbf{d}^t, \boldsymbol{\alpha}^t \rangle$, with coalition structure CS^t , demand vector \mathbf{d}^t (and induced relative demand vector \mathbf{r}^t), and action vector $\boldsymbol{\alpha}^t$. Let C_i denote the $C \in CS^t$ of which i is a member. Each $C \in CS^t$ takes its agreed upon action α_C^t and observes the stochastic outcome s that is realized. The reward $R(s)$ is obtained, with each i in C obtaining its relative share $r_i R(s)$ (where $r_i = d_i / \sum_{j \in C} d_j$). Such outcomes are "local," that is, depending only upon the action α_C taken by C and the type vector \mathbf{t}_C , with $\Pr(s|\alpha, \mathbf{t}_C)$ dictating outcome dynamics. We assume limited observability: agent i observes only the outcome the action of its own coalition C_i , not those of other coalitions.

Once coalitional agreements are reached, actions are executed, and outcomes observed at stage t , the process moves to stage $t + 1$ and repeats. We assume a discount factor γ (with

⁴ We discuss below—and experiment with—settings where the action and reward model vary from stage to stage; but we always assume that the collection of agents, the set of possible types T_i , and the private type of each agent is fixed across all stages.

$0 \leq \gamma < 1$), and model assume agents wish to maximize the expected discounted sum of future rewards. Specifically, let R_i^t be a random variable denoting agent i 's realized reward share at stage t of the process (i.e., in the outcome of coalitional agreements and actions at stage t). Then i 's goal is to maximize

$$\sum_{t=0}^{\infty} \gamma^t R_i^t.$$

3.2.2 A POMDP formulation

To model the behaviour of agents in the repeated coalition formation, the most appropriate formal model would be that of a Bayesian extensive form game. However, we make certain simplifying assumptions, and instead model this as a partially observable Markov decision process. We contrast the POMDP with a full game-theoretic treatment after describing the POMDP formulation.

Our primary goal is to capture how the repeated coalitional setting should influence how agents should approach their coalitional negotiations at each stage game. Each member of coalition C can thus update its beliefs about the types of the members of its coalitions:

$$B_i^{t+1}(t_C) = z \Pr(s|\alpha, t_C) B_i^t(t_C) \quad (5)$$

where z is a normalizing constant. When time is clear from context, we denote this updated belief state $B_i^{s,\alpha}$.

We make one key simplifying assumption, namely, that agent's beliefs are not updated during the coalition formation process itself. In general, any negotiation process that is used to determine coalitions and coalitional agreements can reveal considerable information about one's partners. For example, strategic models of coalition formation as Bayesian extensive form games have been developed for standard coalition formation problems [16, 39] and for BCFPs [15], during which agents update their beliefs during a negotiation process. If this process is well-defined, an intermediate stage of belief update can be factored into our model (and updates based on observing action outcomes would be made against this intermediate model). However, our aim is to abstract away from the specifics of the coalition formation process adopted.

We illustrate the computation of updated beliefs in our earlier example. Suppose agents 1 and 3 partnered and attempted an ambitious project which failed. Agent 1 would update its beliefs about 3's type by conditioning its prior on the observed failure, obtaining posterior beliefs B'_1 :

Type	Agent 2			Agent 3		
	g	m	b	g	m	b
B'_1	0.3	0.4	0.3	0.22	0.39	0.39

The failure of the ambitious project strongly suggests that 3's type is not g . Note that its beliefs about agent 2 are unchanged since we assume 1 can make no observation of the actions or outcomes of coalitions other than its own.

Given the sequential nature of the process, the decisions made by agents regarding which specific coalitions to join—and which coalitional agreements to adopt—should be informed

not only by the immediate expected value of those coalitions; such decisions should also be influenced by the impact they may have on *future decisions*. For example, one agent i may be so uncertain about the type of a potential partner j that the risk of joining a coalition with j has lower expected immediate value than a joining a “safer” team about which it has more certain beliefs. However, should i join with j and discover that, in fact, j ’s capabilities complement its own, this knowledge would prove very valuable, allowing i to partner with j over a long period of time (multiple stages) to reap the dividends of this information. A coalition formation process that does not account for the *value of information* would never allow i to partner with j , hence never discover this hidden value.

To account for just such considerations, we approach optimal repeated coalition formation using a sequential model that allow for just such exploration–exploitation tradeoffs to be made explicitly. By using *Bayesian exploration*, agents optimize long-term sequential value, balancing exploration with exploitation to optimize expected accumulated reward given their current beliefs. We cast the problem as a *partially observable Markov decision process (POMDP)*, or equivalently, a *belief-state MDP*. We assume an infinite horizon problem, with discount factor γ (with $0 \leq \gamma < 1$), and model agents who wish to maximize the expected discounted sum of future rewards. It is reasonably straightforward to formulate the optimality equations for this POMDP; however, certain subtleties will arise because of an agent’s lack of knowledge of other agent beliefs.

Let agent i have beliefs B_i about the types of other agents. Let $Q_i(C, \alpha, \mathbf{d}_C, B_i)$ denote the *long-term value* i places on being a member of coalition C with agreed-upon action α and stated agent demands $\mathbf{d}_C = \langle d_i; i \in C \rangle$ (denoting the agents’s requested payoffs from the formation process), realizing that after this action is taken the coalition formation process will repeat. This is accounted for using Bellman equations [6] as follows:

$$\begin{aligned}
 Q_i(C, \alpha, \mathbf{d}_C, B_i) &= \sum_s \Pr(s|C, \alpha, B_i)[r_i R(s) + \gamma V_i(B_i^{s,\alpha})] \\
 &= \sum_{t_C} B_i(t_C) \sum_s \Pr(s|\alpha, t_C)[r_i R(s) + \gamma V_i(B_i^{s,\alpha})] \tag{6}
 \end{aligned}$$

$$V_i(B_i) = \sum_{C|i \in C, \mathbf{d}_C} \Pr(C, \alpha, \mathbf{d}_C|B_i) Q_i(C, \alpha, \mathbf{d}_C, B_i) \tag{7}$$

(Recall that r_i is i ’s relative demand of the payoff received by C_i , hence $r_i R(s)$ describes i ’s reward.) $V_i(B_i)$ reflects the value of belief state B_i to i , deriving from the fact that given beliefs B_i , agent i may find itself participating in any of a number of possible coalitional agreements, each of which has some Q-value (we elaborate below).

Of note is the fact that agent i considers the (expected, discounted) value of being in its updated belief state $B_i^{s,\alpha}$ —obtained after joining coalition C_i , demanding d_i , and executing coalitional action α_{C_i} —when computing the value of any coalitional agreement. Specifically, the Q-value and value functions for the belief state MDP, described by Eqs. 6 and 7, explicitly incorporate both immediate and future coalitional value in a way that embraces the expected *value of information*: coalitional decisions at future stages may exploit information gleaned from the current interaction.

Unlike typical Bellman equations, the value function V_i cannot be defined by maximizing Q-values. This is because the choice that dictates reward, namely, the coalition that is formed, is not in complete control of agent i . Instead, i must predict, based on it beliefs, the probability $\Pr(C, \alpha, \mathbf{d}_C|B_i)$ with which a specific coalition C (to which it belongs) and a corresponding action-demands pair $\langle \alpha_C, \mathbf{d}_C \rangle$ will arise as a result of negotiation. However,

with this in hand, the value equations provide the means to determine the long-term value of any coalitional agreement.

3.2.3 A Bayesian extensive form game model

A fully general formulation of the repeated coalition setting would require a *Bayesian extensive form game (BEFG)* in which the bargaining or negotiation actions at each BCFP stage game are explicitly modeled. Once coalitional actions are determined at stage t , there are no further action choices at that stage t , since the agreements themselves determine actions taken by the agents; but the information revealed about one's partners by the execution of coalitional actions would cause additional belief update in the game process. Such a model would require explicit representation of the beliefs of all agents at each stage of the negotiation process and after execution of coalitional actions. The appropriate solution concept would then be a *perfect Bayesian equilibrium (PBE)*: agents possess a system of beliefs (beliefs about opponent types at each decision point), and adopt behavioural strategies (mappings from beliefs to actions at each decision point); and these beliefs and strategies must be in equilibrium in the sense that each agent's strategy must be optimal (at all subgames) given its beliefs and the strategies of its opponents, and the beliefs must be determined by Bayesian update w.r.t. the behavioural strategies.

There are two key reasons we do not adopt such a model. First, our model is intended to work with *different forms* of coalitional negotiation. The formulation of the BEFG requires a commitment to a specific negotiation protocol, whereas our model assumes only that some coalition formation process is adopted and that the probability of specific coalitional agreements can be estimated. While a BEFG formulation that abstracts away the details of the negotiation processes during the BCFP stage games would be desirable, this is not feasible since the only strategic behaviour available to the agents is embedded within the stage games themselves. Outside of the stage games in which coalitional negotiation takes place, agents have no decisions to make: they simply implement their agreed upon coalitional actions and observe the results. One might consider adopting an *empirical games* approach [61], using detailed simulation to obtain a high-level description or characterization of the underlying negotiation process and use these results to inform PBE computation. However such an approach is beyond the scope of this work.

The second reason is the analytic and computational intractability of BEFGs. Analyzing equilibria of BEFGs is notoriously difficult, especially given the complexity of the repeated coalition formation problems we propose. Even studying the BEFG formulation of the stage games themselves (see [15] for such an analysis) often requires simplifying assumptions. The computation of such equilibrium is also unlikely to prove practical for anything but the simplest settings.

Our POMDP model can thus be viewed as a heuristic approximation to the modeling of the behaviour of strategically-aware agents. The model abstracts much of the detail about strategic interactions into beliefs about agent types, and makes the simplifying assumption that updates of these beliefs is based only on the "objective" evidence supplied by action outcomes (i.e., evidence that is not manipulable by the strategic choices of other agents). This does prevent us from capturing sophisticated strategies that explicitly link behaviour across stages (e.g., in which the outcomes of future negotiations can be influenced by past strategic behaviour, for example, via "threats"). However, while this an important avenue for future research, we believe that, by taking such an approach, practical analytic and computational results would be possible only for the simplest of settings.

3.3 Estimating agreement probabilities

One aspect of the POMDP model that is not directly part of the input are the “agreement” probabilities $\Pr(C, \alpha, \mathbf{d}_C | B_i)$ that an agent i needs to predict the coalitional agreements that might arise. These probabilities obviously depend on the precise coalition formation mechanism adopted, and can be estimated or approximated in a variety of ways.

If a discounted coalitional bargaining model is assumed [15, 16, 39], an agent i can compute these probabilities by simulating the process of solving the game tree (possibly using a heuristic algorithm). If the equilibrium solution is unique then the resulting agreement vector $\langle CS, \mathbf{d}, \alpha \rangle$ will occur with certainty. If multiple solutions arise, each can be assigned uniform probability of occurrence, or some other biases could be incorporated.

In contrast, a simpler form of negotiation could be assumed, such as the best response with experimentation (BRE) process [12, 15, 24] discussed in Sect. 2, a simple dynamic negotiation process in which agents make proposals (in some predetermined or random order) and groups of agents accept or reject them using myopic best responses w.r.t. the existing set of coalitional agreements. The process can be modeled as Markov chain [24] whose state at time t is simply the prevailing agreement vector $\langle CS^t, \mathbf{d}^t, \alpha^t \rangle$. The steady-state distribution P^* of this chain then dictates the probability of ending up with any specific set of coalitional agreements.

However, such an approach is inherently problematic in our setting. The transition matrix of the Markov chain, hence the steady-state distribution, requires knowledge of all parameters affecting the state transitions of the BRE process (or any other heuristic negotiation process). Unfortunately, a specific agent i does not have complete knowledge of these parameters, since it has only probabilistic information of the types of other agents, and is unaware of their beliefs. Expectations w.r.t. agent types can be used of course. The use of *common prior* can help approximate the beliefs of other agents as well, though this can be somewhat problematic. Suppose for example, that at each RL stage, there is a common prior, shared by all agents, specifying the probability with which the agent type profiles are drawn; and that agents use this common prior to estimate the probabilities (beliefs) that other agents of a specific type assign to type profiles (i.e., each agent uses the common prior to represent the beliefs of their opponents). Using the prior in this static fashion to account for the beliefs of others is unrealistic, since agents update their beliefs at the conclusion of each stage of the RL process. Furthermore, even with a common prior, it is not possible for agents to accurately monitor the belief dynamics of other agents, since they are unable to observe the outcomes of the actions of coalitions other than their own.

For these reasons (and others discussed in the next section), we do not try to compute the Markov chain or its steady-state distribution. Rather, we approximate the quantities $\Pr(C, \alpha, \mathbf{d}_C | B_i)$ determined by BRE (or other dynamic) processes, in other ways (which we elaborate in the next section). In addition, should it be necessary for an agent to estimate the beliefs of others, we avoid the simple static common prior assumption, and instead use a heuristic approach.

4 Computational approximations

Computing the exact solution Bayesian optimal solution to the reinforcement learning problem for repeated coalition formation, as specified by Eqs. 6 and 7, is generally infeasible for two reasons. First, as discussed above, the difficulty of estimating relevant quantities, agreement probabilities in particular, will often require some approximation. Second, because

Table 1 Approximating the optimal solution to the problem of repeated coalition formation under uncertainty

1. Each agent i with belief state B_i computes the Q-value of potential agreements $\langle C, \alpha, \mathbf{d}_C \rangle$ in which it participates by approximating the solution to Eqs. 6 and 7, using one of the following algorithms (described below): *OSLA*, *VPI*, *VPI-over-OSLA*, *Myopic*, *MAP*
2. The agents engage in some coalition formation process, with each agent using the Q-values computed above in order to represent the *long term value* of coalitional agreements. The process results in a coalition structure CS , a payoff allocation vector \mathbf{d} , and a vector of coalitional actions α , one for each coalition $C \in CS$
3. After observing the result of the coalitional action α_{C_i} of the coalition C_i to which it belongs, each agent i updates it beliefs about its partners $j \in C_i, j \neq i$ using Eq. 5
4. The RL process repeats

solving POMDPs is generally intractable, especially given the size of the state space (the cross-product of agent types), action space (the set of coalitional agreements), solving the induced POMDP will also require some form of approximation to be practical.

In this section, we describe several algorithms that approximate the POMDP solution. Each of the Bayesian RL algorithms below can be combined with any underlying negotiation process: we require only that the process result in coalitional agreements of the type $\langle C, \alpha, \mathbf{d}_C \rangle$ assumed above, and that agents can estimate the probability of specific agreements given beliefs about agent types. A skeletal algorithm is shown in Table 1. We now describe four instantiations of this framework, of varying levels of sophistication.

4.1 The one-step lookahead algorithm

The first approximate RL algorithm is the *one-step lookahead algorithm (OSLA)*. While dynamic programming techniques are often used for solving POMDPs [30, 55], (tree) search techniques are often more effective for MDPs (and POMDPs) [5, 19] over when the initial state (or belief state) is known. In OSLA, we use a very restricted form of tree search, looking ahead only a single step at the value of coalitional agreements at the current state B_i^t at time t and possible states at the next stage B_i^{t+1} given possible outcomes of coalitional actions. This maintains tractability, exploiting the fact that only a (relatively) small number of belief states can be reached after the execution of a single coalitional action.

More precisely, we compute the Q-values of successor states in Eq. 6 in the OSLA method myopically. Specifically, we define the *one-step lookahead* Q-value of agreement $\langle C, \alpha, \mathbf{d}_C \rangle$ for agent i , under belief state B_i , to be:

$$\begin{aligned} Q_i^1(C, \alpha, \mathbf{d}_C, B_i) &= \sum_s \Pr(s|C, \alpha, B_i)[r_i R(s) + \gamma V_i^0(B_i^{s,\alpha})] \\ &= \sum_{\mathbf{t}_C} B_i(\mathbf{t}_C) \sum_s \Pr(s|\alpha, \mathbf{t}_C)[r_i R(s) + \gamma V_i^0(B_i^{s,\alpha})] \end{aligned} \quad (8)$$

$$V_i^0(B_i) = \sum_{C, \beta \in A(C), \mathbf{d}_C | i \in C} \Pr(C, \beta, \mathbf{d}_C | B_i) Q_i^0(C, \beta, \mathbf{d}_C, B_i) \quad (9)$$

$$Q_i^0(C, \beta, \mathbf{d}_C, B_i) = r_i \sum_{\mathbf{t}_C \in \mathcal{T}_C} B_i(\mathbf{t}_C) \sum_s \Pr(s|\beta, \mathbf{t}_C) R(s) \quad (10)$$

In Eq. 8, $V_i^0(B_i^{s,\alpha})$ represents the myopic, or immediate value of the successor belief state to B_i , which is defined (Eq. 9) to be the expected *immediate*, rather than long-term, value of the agreement that will emerge.

As discussed above, computing the probability of agreement terms $\Pr(C, \beta, \mathbf{d}_C | B_i)$ can be very difficult. As a result, we approximate these probabilities by assuming a specific negotiation process, and again adopt a simple one-step look ahead perspective on coalitional negotiations. Specifically, we assume that agents engage in a the best-response with experimentation (BRE) process for BCFPs discussed above. (We refer to [10, 12] for a detailed description of the BRE process.) While this process induces a Markov chain over coalition structures and agreements, we assume that the process will terminate after a *single step* of this process. We then estimate the probability of an agreement to be the probability at reaching that agreement at the first stage of the induced Markov chain. We note that agents need only compute Q-values, both $Q_i^1(C, \alpha, \mathbf{d}_C, B_i)$ and $Q_i^0(C, \alpha, \mathbf{d}_C, B_i)$, for agreements that can be reached at those stages. Since only a small number of agreements could be reached by a restricted negotiation process, this further enhances computational tractability.

We chose to use a negotiation tree-depth bound of one when computing agreement probabilities using the BRE process in our experiments, largely for reasons of computational efficiency. However, this lookahead bound could take any value of $l \geq 1$, depending on the specific setting's requirements. In addition, the continuous nature of agent demands and the combinatorial number of type vectors can also cause computational difficulty in evaluation the search tree, even to depth one. Nevertheless, the use of sampling and appropriate discretization of demands can alleviate these problems. We discuss this in Sect. 6 when describing our experiments.

Another issue that we address is the evolution of beliefs. A static common prior is inappropriate, since each agent will update its beliefs at each stage of the RL process. But as discussed above, agents have little evidence upon which to update their knowledge of the belief states of *other* agents due to the limited observability in our model. We account heuristically for this within OSLA by having agents initially adopt a *static assumption* about the beliefs of all other agents: specifically, an agent i will assume that all agents $j \neq i$ maintain their prior for the first k stages of the RL process. After k stages, agents will adopt a *convergent assumption*: specifically, agent i will assume that the beliefs of all agents j coincide with its own.⁵ This is a very crude heuristic that attempts to account for the dynamic nature of opponent beliefs, exploiting the fact that the beliefs of all agents (with sufficient exploration of structures and agreements) will converge toward the true vector, despite the very limited observability of opponent belief dynamics.

We note that many of these approximations provide only crude estimates of long-term value. We will see this reflected in the performance of OSLA below.

4.2 The VPI exploration method

The *value of perfect information (VPI) exploration method* is an RL technique that approximates optimal Bayesian exploration using the (myopic) expected value of perfect information inherent in an agent's actions. VPI exploration was initially developed in [20, 21] for single-agent environments. We extend the VPI formulation to the multiagent, repeated coalition setting by defining how to estimate the expected (myopic) value of perfect information of coalitional agreements given an agent's current beliefs. The sequential value of any coalitional action, accounting for its value of information, is then used in the formation process.

Let us consider what can be gained by learning the true value of a coalitional agreement $\sigma = \langle C, \alpha, \mathbf{d}_C \rangle$. Suppose σ is adopted, its action α executed, and assume that it provides *exact evidence* regarding the types of the agents in C . Thus, we assume that the real type

⁵ Our experiments in Sect. 6 use static beliefs for the first 50 RL steps and convergent beliefs after that.

vector \mathbf{t}_C^* is revealed following σ . If agent types are revealed, then the *true value* of σ is also revealed: it is simply agent i 's share of expected coalition value given α , which we denote by $q_\sigma^* = q_{(C,\alpha,\mathbf{d}_C)}^* = Q_i(C, \alpha, \mathbf{d}_C | \mathbf{t}_C^*)$, where

$$Q_i(C, \alpha, \mathbf{d}_C | \mathbf{t}_C^*) = r_i \sum_s \Pr(s | \alpha, \mathbf{t}_C^*) R(s). \tag{11}$$

This is a ‘‘myopic’’ calculation of the specific (future) coalitional agreement value, assuming the adoption of σ and subsequent revelation of actual agent types.

This new knowledge is of value to agent i only if it leads to a change of its policy. This can happen in two cases: (a) when the information shows that a coalitional action that was previously regarded as inferior to the best action is now revealed to be the best choice; or (b) when the information indicates that the action previously regarded as best is actually worse than the second best action.

For case (a), suppose that given current belief state B_i the value of i 's current best action $\sigma_1 = \langle C_1, \alpha_1, \mathbf{d}_{C_1} \rangle$ is $q_1 = Q_i(C_1, \alpha_1, \mathbf{d}_{C_1} | B_i) = E_{B_i}[q_{(C_1,\alpha_1,\mathbf{d}_{C_1})}]$. Moreover, suppose that the new knowledge indicates that σ is a better action; that is, $q_\sigma^* > q_1$. Thus, we expect i to gain $q_\sigma^* - q_1$ by virtue of performing σ instead of σ_1 .

For case (b), suppose that the value of the second best action $\sigma_2 = \langle C_2, \alpha_2, \mathbf{d}_{C_2} \rangle$ is $q_2 = Q_i(C_2, \alpha_2, \mathbf{d}_{C_2} | B_i) = E_{B_i}[q_{(C_2,\alpha_2,\mathbf{d}_{C_2})}]$. If action σ coincides with the action considered best, σ_1 , and the new knowledge indicates that the real value $q_{\sigma_1}^* = q_\sigma^*$ is less than the value of the previously considered second-best action—that is, if $q_{\sigma_1}^* < q_2$ —then the agent should perform σ_2 instead of σ_1 and we expect it to gain $q_2 - q_{\sigma_1}^*$.

Thus, the *gain* from learning the true value q_σ^* of the σ agreement is:

$$gain_\sigma(q_\sigma^* | \mathbf{t}_C^*) = \begin{cases} q_2 - q_\sigma^*, & \text{if } \sigma = \sigma_1 \text{ and } q_\sigma^* < q_2 \\ q_\sigma^* - q_1, & \text{if } \sigma \neq \sigma_1 \text{ and } q_\sigma^* > q_1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

However, agent i does not *know* what types (and, consequently, which Q-value) will be revealed for σ ; therefore, we need to take into account the expected gain given its prior beliefs. Hence, we compute the *expected* value of perfect information of σ :

$$VPI(\sigma | B_i) = \sum_{\mathbf{t}_C^*} gain_\sigma(q_\sigma^* | \mathbf{t}_C^*) B_i(\mathbf{t}_C^*) \tag{13}$$

Expected VPI gives an upper bound on the myopic value of information for exploring coalitional action σ . The expected *cost* of this exploration is the difference between the (expected) value of σ and the value of the action currently considered best, i.e., $q_1 - E_{B_i}[q_\sigma]$, where $E_{B_i}[q_\sigma] = E_{B_i}[q_{(C,\alpha,\mathbf{d}_C)}]$ is given by $E_{B_i}[q_\sigma] = r_i \sum_{\mathbf{t}_C \in T_C} B_i(\mathbf{t}_C) \sum_s \Pr(s | \alpha, \mathbf{t}_C) R(s)$. Consequently, an agent should choose the action that maximizes

$$VPI(\sigma | B_i) - (q_1 - E_{B_i}[q_\sigma]). \tag{14}$$

This strategy is equivalent to choosing the proposal that maximizes:

$$QV_i(\sigma | B_i) = E_{B_i}[q_\sigma] + VPI(\sigma | B_i) \tag{15}$$

Agents then use these *QV* values instead of using the usual Q-values in their decision making for forming coalitions. The computation of expected values and VPI above can be done in a straightforward manner if the number of possible type configurations is small. If, however, this number is large, sampling of type vectors can be employed.

In summary, the VPI algorithm proceeds as follows:

1. The “true” Q-values of any potential agreement σ , with respect to each realization of the relevant type vector, are myopically calculated via Eq. 11.
2. The gain from reaching agreement σ is calculated via Eq. 12.
3. The VPI for σ is calculated via Eq. 13.
4. The Q-values QV_i for (any) σ are calculated through Eq. 15 (and are subsequently used in the coalition formation process).

VPI exploration is in a sense non-myopic, since it implicitly reasons about the value of future belief states through its use of value of perfect information in estimating the value of future coalitional agreements. However, VPI uses myopic calculations when determining the value of agreements. Even though this is an approximation, it enables the method to exploit the value of (perfect) information regarding agent types, however myopic the estimation of this value may be, instead of estimating the specific value of anticipated coalitional actions (in contrast to lookahead methods). Unlike lookahead methods, VPI does not have to explicitly incorporate the common prior hypothesis when estimating the Q-values used during coalition formation, nor does it need to account for the probability of agreement when transitioning to future belief states. The VPI exploration method is thus less tightly coupled to the details of the underlying coalition formation process. As we demonstrate below, myopic VPI estimation works very well in a variety of experimental settings.

Nevertheless, we also develop a method which combines VPI with OSLA. *VPI-over-OSLA* uses the application of VPI over Q-values estimated using the OSLA method. When this method is used, the values of currently expected best action, second best action and exploratory action σ are estimated using one-step lookahead (and, thus, there is a need to approximate the probabilities of future agreements). In brief, VPI-over-OSLA proceeds as follows:

1. The “true” q-values of any potential agreement σ are calculated, assuming one-step lookahead and calculation of the V_i^0 and Q_i^0 values of the successor belief state (following the revelation of the true t_C^*) through Eqs. 9 and 10.
2. The gain from reaching agreement σ is calculated via Eq. 12, where the values q_1 and q_2 of the best and second-best actions are calculated through Eqs. 8, 9 and 10.
3. The VPI for σ is calculated via Eq. 13.
4. The Q-values QV_i for (any) σ are calculated through Eq. 15 (and are subsequently used in the coalition formation process).

4.3 Myopic Bayesian RL algorithm

A very simple RL algorithm is the *myopic Bayesian RL algorithm*. It is purely myopic, with agents reasoning only about the immediate value of coalitional agreements and ignoring the value of any future coalitions or the evolution their belief states. An agent i using myopic Bayesian RL computes the value of an agreement given belief state B_i , as follows:

$$Q_i(C, \alpha, d_C, B_i) = r_i \sum_{t_C \in T_C} B_i(t_C) \sum_s \Pr(s|\alpha, t_C) R(s).$$

4.4 Maximum a posteriori type assignment RL algorithm

Another relatively simple, but sometimes effective algorithm is the *maximum a posteriori type assignment (MAP)* algorithm. This algorithm effectively reduces the problem of estimating the Q-values of agreements given an agent’s beliefs about opponent types to the problem

of estimating Q-values about agreements given a specific opponent type vector, namely, the most probable type vector given its beliefs.

More precisely, given belief state B_i , agent i assumes that the actual type t_j^i of opponent j is the most probable type: $t_j^i = \operatorname{argmax}_{t_j} B_i(t_j)$. The vector of types t_C assumed by i for any coalition C is defined in this way, giving the following estimate of the value of any agreement:

$$Q_i(C, \alpha, d_C | t_C) = r_i \sum_s \Pr(s | \alpha, t_C) R(s)$$

Notice that this calculation is myopic, not accounting for the sequential value of an agreement. However, the sequential value of agreement under the MAP type assumption is simply the discounted sum of the (best) myopic agreement value: with type uncertainty is assumed away, there is no need for belief update or negotiation under type uncertainty.

5 Combining RL algorithms with the coalition formation process

It is easy to define variants of our Bayesian RL algorithms, in order to accommodate different environment requirements. What is more, we can partition the space of the possible variants of RL algorithms by examining their combination with various coalition formation processes. For example, we can consider the following four classes of reinforcement learners, combining Q-value estimation with dynamic formation (or negotiation) processes.

The first are *non-myopic/full negotiation (NM-FN)* agents. Agents in this class employ *full negotiation* when forming coalitions, attempting to find a stable (e.g., Bayesian core) structure and allocation before engaging in their actions. For instance, they might use the dynamic process described above to determine suitable coalitions given their current beliefs. Furthermore, they employ sequential reasoning (using the OSLA or the VPI RL method, for example), in their attempt to solve the POMDP described by Eqs. 6 and 7.

Myopic/full negotiation (M-FN) agents use full negotiation to determine coalitions at each stage. However, they do not reason about future (belief) states when assessing the value of coalitional moves. Essentially, M-FN agents engage in repeated application of a coalition formation process (such as BRE), myopically choose actions, and repeat.

Myopic/one-step proposers (M-OSP) are agents that are myopic regarding the use of their beliefs when estimating coalition values (like M-FN), but do not employ full negotiation to form coalitions. Rather, at each stage of the RL process, one random proposer is assumed to be chosen, and once a proposal has been made and accepted or rejected, no further negotiations are assumed to take place: the coalitional action is assumed to be executed after a *single* proposal. Finally, *non-myopic/one-step proposers (NM-OSP)* are, naturally, the obvious combination of NM-FN and M-OSP agents. Notice that the fact that OSP agents assume (from an RL perspective) that the negotiation process has only one round, does not necessarily mean that the actual negotiations will last for just one round. Specifically, an agent may deliberate about the value of various agreements by supposing one-step negotiation, to simplify its reasoning. This is possible even if the actual negotiation uses multiple rounds. Nevertheless, in our experiments, all OSP agents simulations involve actual negotiations that last for one round.

FN approaches have the advantage that at the end of each RL stage, before actions are executed, the coalition structure is in a stable state (depending on the nature of the coalition formation process). Another advantage of FN is that agents have the opportunity to update

their beliefs regarding other agents's types during the negotiation itself.⁶ However, FN-methods generally do not permit agents to fully explore the full space of coalitions or actions: at the end of each stage, the agents will indeed have strong information on a sub-space of the coalition structure space, specifically the subspace that contains the stable coalition structure the agents have led themselves into; but the agents may not have the opportunity to explore coalition structures that are unreachable given their beliefs (since if they reach a stable structure, they have little interest in further exploration). This is in contrast to OSP approaches, which may potentially provide the agents with more flexibility to investigate the whole space of structures.

6 Experimental evaluation

We conduct three sets of experiments to evaluate the different RL methods and computational approximations proposed above. Evaluation includes varying the nature of the negotiation process by examining distinctions between the full negotiation and one-step proposals for coalition formation during the stage games. In all cases, we measure the total discounted reward accumulated by all agents to measure the quality of the coalitions formed, and examine their evolution over time. We also examine “convergence” behaviour in some cases, to test whether stable coalitions are formed, and to look at quality of performance after learning has stabilized. For reasons discussed above, solving POMDPs for problems of this scope is impractical, as is the formulation of a BEFG for our model. Hence we are unable to compare our approximations to an “exact” solution.

The first set of experiments examines the performance of our learning methods with agents that face the *same* coalition formation problem—specifically with the same action dynamics and reward model—at each stage (though with different beliefs). The second considers *dynamic tasks*, in which the actions and rewards vary at each stage; this will demonstrate the ability of our RL framework and methods to support knowledge transfer across tasks through its focus on type learning. The third set of experiments compares our methods to an adaptation of a successful algorithm for coalition formation under uncertainty proposed in the literature [34].

The process used during the coalition formation stages is the BRE dynamic process [12] for BCFPs. In estimating Q-values, we sample type vectors rather than exhaustively enumerating them all as follows. Let $|T|$ be the number of types and $|C|$ the size of coalition in question. If $|T|^{|C|} \leq 1000$, no sampling is used; otherwise, 100 type vectors are sampled according to the belief distribution.

6.1 Static tasks

We first test our approach in two related static settings, i.e., in which the action dynamics and rewards are identical at each stage. This can be viewed as agents facing the same choice of tasks throughout the RL process. In both settings, coalitions have three actions available, each with three possible stochastic outcomes depending on the coalition's type vector. In the first setting, five agents and five types are used; in the second, ten agents and ten types.

⁶ We do not explore this possibility in our experiments, in order to focus on the RL aspects of the repeated coalition formation problem, rather than those of bargaining. We note, however, that one can devise methods that make use of the update of beliefs while observing the opponents's responses to proposals. See, for example, [13].

Intuitively, the agents form companies to bid for software development projects. There are three agent *roles*, corresponding to project roles, each having three or four *quality levels*, with the combination of role and quality determining an agent's type: *interface designer* = $\langle \text{bad}, \text{average}, \text{expert} \rangle$, *programmer* = $\langle \text{bad}, \text{average}, \text{good}, \text{expert} \rangle$ and *systems engineer* = $\langle \text{bad}, \text{average}, \text{expert} \rangle$. The quality levels correspond to quality "points" (0 points for bad, and increasing by 1 for each quality increment), and the overall *quality* of a coalition is the sum of these points. Agents know the role of their opponents, but not their quality levels. Companies (coalitions) can bid for a large, medium or small projects (actions), and they can make large, average or small profits (outcomes), depending on the bid and the members's types. The outcome (and subsequent coalitional reward) of an action depends on the quality of the coalition. A coalition can (with high probability) receive large profits by bidding on a large project only if its total quality is high and there is sufficient diversity of roles amongst its members. A coalition with two (respectively, more) members is "punished" if it does not have two (resp., at least 3) members with different roles, by receiving only a fraction of the reward it is entitled to given the quality of its members. Tables 6, 9, and 10 in Appendix A illustrate the dynamics of the problem and rewards. Reward shares that the members of size two coalitions can expect to receive are equal to their reward for acting alone, but less if the two-member coalition is made up of members with the same role. Thus agents using a Myopic method will find it hard to form size two coalitions (starting from a configuration structure of singletons), even if in fact these coalitions can serve as the "building blocks" for more promising ones.

Tables 7 and 8 in Appendix A shows the types of the agents in both the five and ten-agent experiments. The five-agent environment is such that the (classic deterministic) core is non-empty, while the 10-agent environment has an empty core. Within each environment, we consider two distinct settings, one in which agents have a *uniform* prior over (quality) types of opponents, and one with a *misinformed* prior—in this case agents believe with probability 0.7 that each of its opponents has a type different than its actual.

Agents are homogeneous in the sense that each employs the same learning algorithm in any given setting. Each experiment consists of 30 runs, each with 500 RL steps. A discount factor of 0.985 is used in all experiments. Whenever a full negotiation (FN) approach is used, formation negotiations last for 50 rounds (per RL stage). Agents observe only the results of the action taken by their coalition, not those of any other coalition, and only update beliefs about their partners at any specific stage.

Figures 1, 2, 3 and 4 plot the discounted reward accumulated by the coalitions in each homogeneous environment consisting of Myopic, OSLA, VPI, MAP, or VPI-over-OSLA agents (averaged over 30 runs). In addition, Tables 3 and 4 report the average "per step" reward accumulated during the final 50 RL steps, once agents's beliefs and behaviour are expected to have stabilized. For comparison against an optimal (nonlearning) benchmark, we also tested the behaviour of agents who knew the exact types of each agent (i.e., were fully informed): total discounted average reward (over 30 runs) after 500 iterations is shown in Table 2 (not plotted for legibility reasons). In the 5-agent case, the structure (and actions) agreed upon by the fully informed agents is optimal (i.e., with maximum expected collective payoff) and core-stable.

In the five-agent experiments (Figs. 1, 2), we see that VPI tends to be the best performing algorithm (with the exception of uniform priors, full negotiation, Fig. 1b), though the advantage over MAP is not statistically significant in the one-step proposal settings. Myopic consistently performs worst. Interestingly, MAP performs reasonably well, beating all other methods with uniform priors and full negotiation. MAP effectively employs a crude form of exploration, with agents behaving in an overly optimistic or pessimistic manner w.r.t. the

Table 2 Discounted average (over 30 runs) total accumulated payoffs after 500 RL steps for *fully informed* agents employing either FN or OSP formation algorithms

Fully informed agents	5 agents	10 agents
Full negotiation	183713	258726
One-step proposals	139965	226490

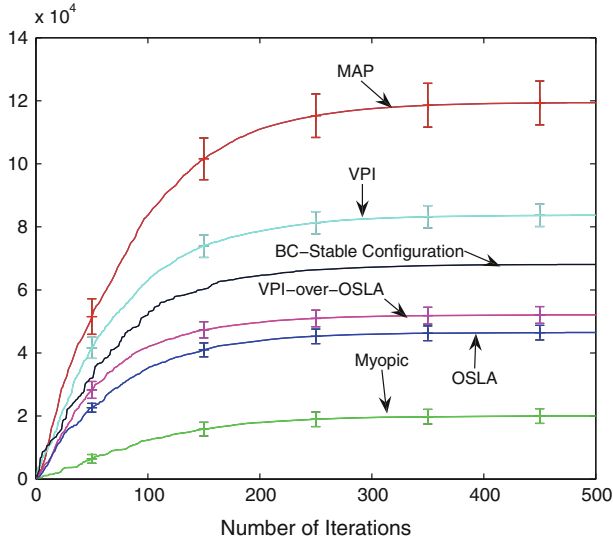
value of information they receive: a slight modification of their beliefs may “point” to a completely different partner. This turns out to be helpful in the five-agent setting, since roles are known, and type uncertainty limited to only 3 or 4 unknown quality levels; furthermore, the reward signal is quite clear regarding the quality of coalitions. Hence, MAP agents are able to quickly determine which partners have good quality types and stick to their choices. In fact, MAP agents manage to achieve high reward without ever reducing type uncertainty regarding most of their potential partners. Define $D(x, \tau_y) = 1 - B_x(t_y = \tau_y)$ to be the distance between x 's beliefs about the type τ_y of agent y and y 's true type. We observe distances of approximately 0.75 or 0.66667 regularly at the end of RL stage 500 with MAP agents, which coincides with the initial distances prior to stage 1.

In the five agent-full negotiation experiments we also plot the “BC-Stable Configuration” curve corresponding to reward accumulated by a group of agents that are placed in a strong Bayesian core configuration (w.r.t. their initial beliefs) at each stage (i.e., with no renegotiation or learning involved). This BC-stable configuration is quite rewarding (though not optimal). With both uniform and misinformed priors (Fig. 1a, b), we see that VPI agents performing substantially better agents statically placed in the Bayesian core. (The plot is identical in the one-step proposer cases and is not shown.)

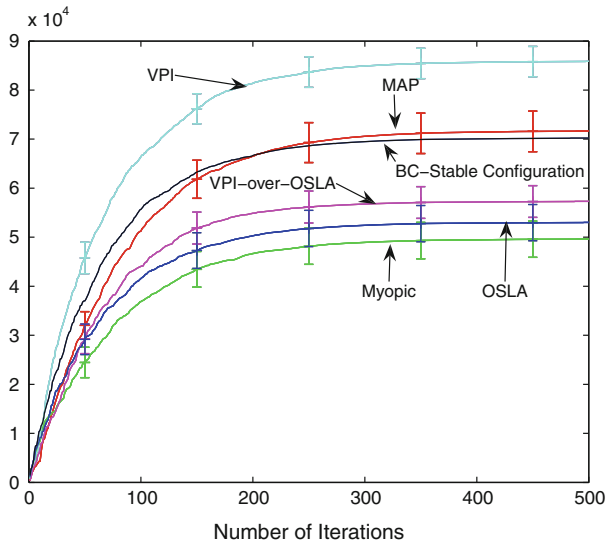
Our results indicate that agents using full negotiation are more successful than those using one-step proposals by a wide margin (from 2–4 times better performance in most cases). The fact that FN agents engage in lengthy dynamic formation processes at each RL stage enables them to reach more stable configurations (as these are to a greater extent the product of “collective consensus,” and likely closer to more rewarding states). The more “exploratory” nature of OSP agents is also evident when observing the error bars in these figures. Nevertheless, Tables 3 and 4 shows that OSP agents do eventually match the performance of FN agents in the stage games. In fact, in the five-agent settings (Table 3), they consistently (with the exception of MAP-Uni) achieve per-stage reward which is higher than that gathered by their FN counterparts. Thus, the more exploratory nature of OSP agents pays benefits in the long run—but at the expense of performance while learning.

In the ten-agent experiments, with a larger number of agents and a more complicated environment, VPI establishes itself as the most successful of our methods, both in terms of discounted accumulated reward (Figs. 3, 4) and per-stage reward once agent beliefs have “converged” (Table 4). VPI accumulates 76.9% of the average discounted reward earned by fully informed agents in the misinformed priors-full negotiation case (and 74.5% in the uniform priors-full negotiation case). One important observation is that VPI agents achieve good performance without, in most cases, significantly reducing their type uncertainty. For example, in the experiments shown in Fig. 3b, in most cases $D(x, \tau_y)$ ranges from 0.5 to 0.96 at the end of stage 500. (To be more exact, only 8 out of the 90 possible runs—since we have 10 agents with beliefs about 9 possible partners—had $D(x, \tau_y)$ less than 0.5.) This illustrates the point that it is not always necessary for agents to explicitly reduce total uncertainty, but only relevant uncertainty, to achieve good performance.

Our results, especially those involving ten agents, show that OSLA and VPI-over-OSLA perform poorly w.r.t. discounted accumulated reward. We attribute this to the strong assump-



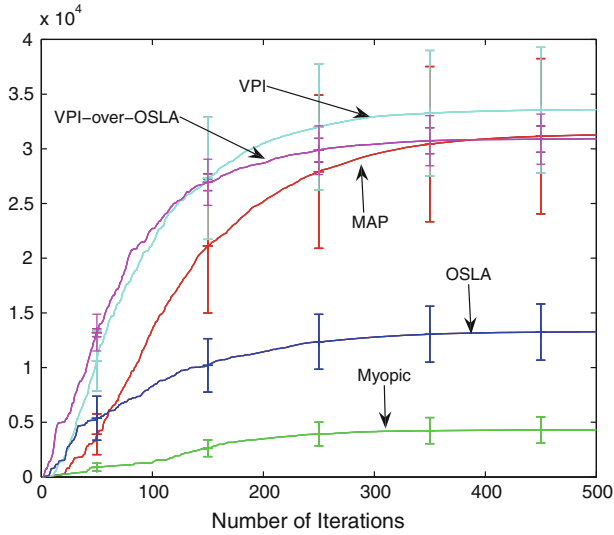
(a) Uniform priors



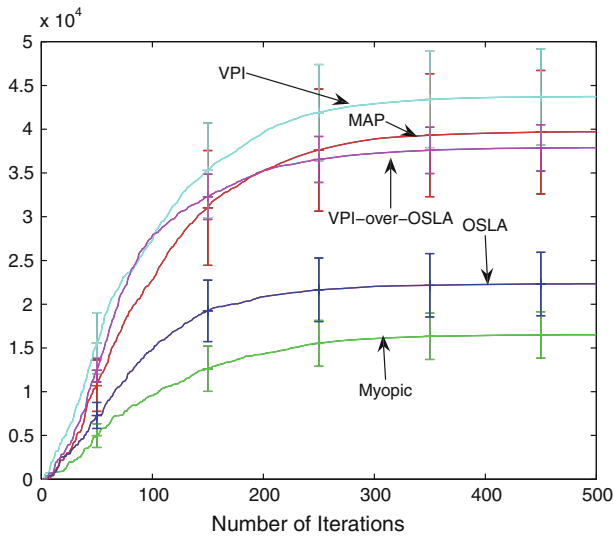
(b) Misinformed priors

Fig. 1 Experiments with five agents, full negotiation. Discounted average total payoff accumulated by coalitions (in 30 runs). *Error bars* are 95% confidence intervals. The “BC-Stable configuration” is a non-optimal one, and involves no learning. The discounted average accumulated payoff for an optimal core-stable configuration at step 500 is as shown in Table 2 (i.e., 183,713)

tions and minimal lookahead of OSLA, though it is notable that VPI-over-OSLA consistently achieves better performance than OSLA. Interestingly, the performance of OSLA and VPI-over-OSLA in the final RL stages (Tables 3, 4) is often comparable (and in some cases superior) to that of methods that fare better w.r.t. discounted accumulated reward. Unsur-



(a) Uniform priors

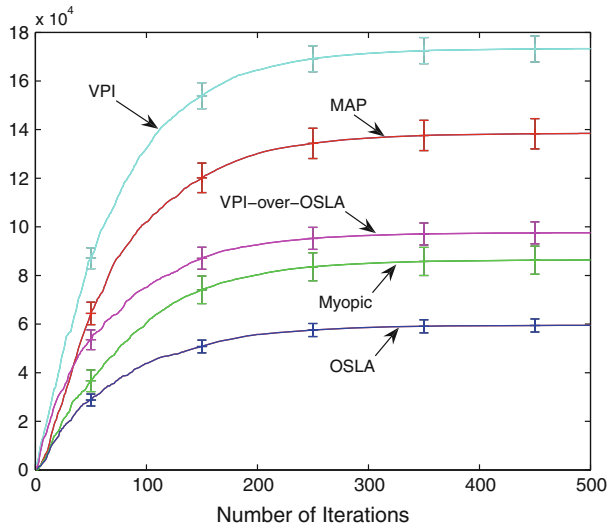


(b) Misinformed priors

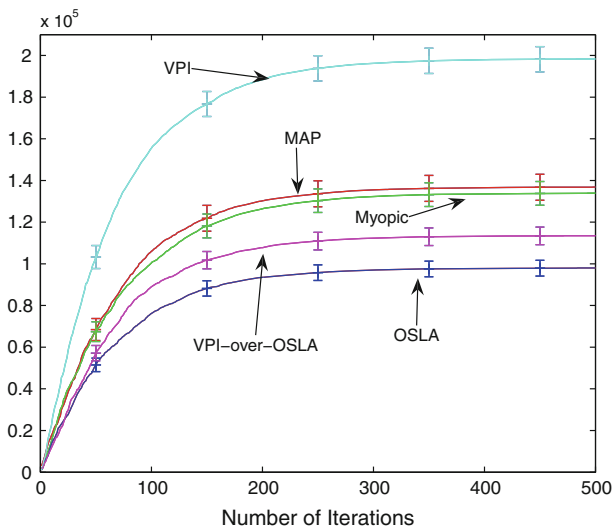
Fig. 2 Experiments with five agents, one-step proposals. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals. The discounted average accumulated payoff for an optimal core-stable configuration at step 500 is as shown in Table 2 (i.e., 139,965)

prisingly, Myopic usually exhibits poor performance—it is far too cautious, and unable to progressively building profitable coalitions.

With regard to the stability of the coalitions formed in the 5-agent setting, VPI, OSLA, VPI-over-OSLA, and Myopic agents frequently find themselves in a BC configuration while learning (i.e., at the end of formation stages, before executing coalitional actions), even if



(a) Uniform priors



(b) Misinformed priors

Fig. 3 Experiments with ten agents, full negotiation. Discounted average total payoff accumulated by coalitions (in 30 runs). *Error bars* are 95% confidence intervals

they do not “converge” to one. Convergence results are shown in Table 5.⁷ Convergence is assumed if a least 50 consecutive RL trials before final stage resulted in a BC configuration. MAP agents managed to converge to the rewarding stable configurations quite often (and

⁷ When we tried some runs for 10000 RL steps, the methods did seem to be able to converge to BC allocations more often.

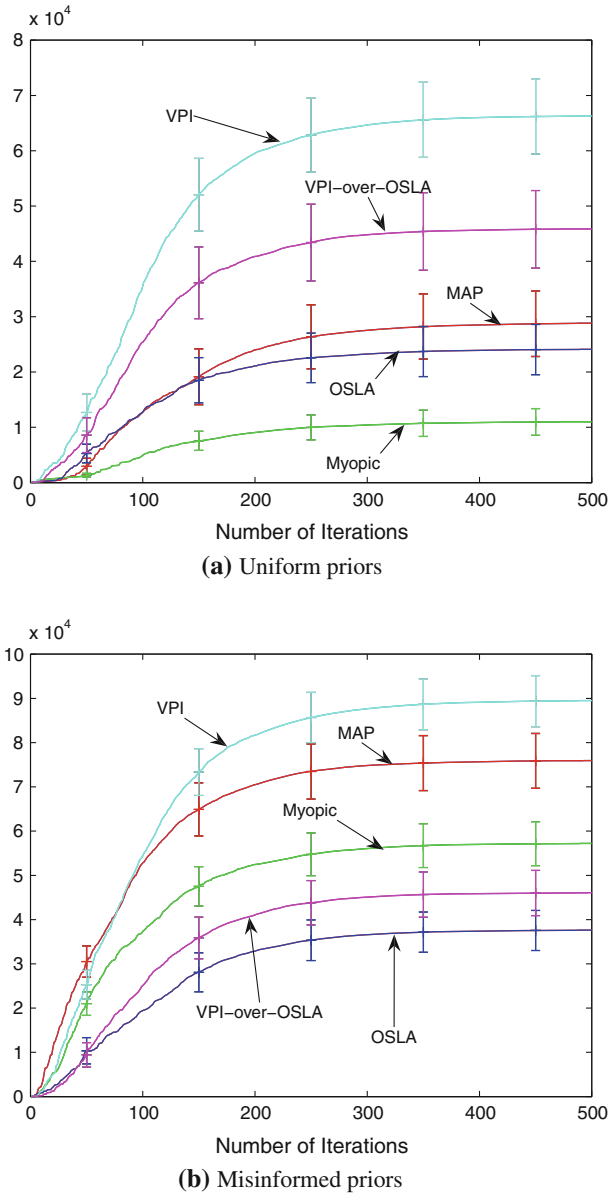


Fig. 4 Experiments with ten agents, one-step proposals. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals

this contributed to their good performance in the uniform priors-full negotiations case.) With ten agents, the core is empty so convergence is not measurable.

While MAP exhibits reasonable performance, these result suggest that VPI is the superior method for reinforcement learning in repeated coalition formation. Further, VPI is much more computationally effective than the other methods.

Table 3 Experiments with 5 agents. Average “per step” reward accumulated within the final 50 RL steps of a run. *Uni*, uniform, *Mis*, misinformed prior

Method	Reward
(a) Full negotiations	
Fully informed agents	2992.82
VPI-Uni	1503.08 (50.23%)
VPI-Mis	1387.28 (46.35%)
VPI-over-OSLA-Uni	873.74 (29.19%)
VPI-over-OSLA-Mis	783.44 (26.18%)
OSLA-Uni	860.72 (28.76%)
OSLA-Mis	807.18 (26.97%)
MAP-Uni	2745.6 (91.73%)
MAP-Mis	1218.24 (40.7%)
Myopic-Uni	824.96 (27.56%)
Myopic-Mis	1046.64 (34.97%)
(b) One-step proposals	
Fully informed agents	2392.54
VPI-Uni	1611.4 (67.35%)
VPI-Mis	1562 (65.29%)
VPI-over-OSLA-Uni	1063.14 (44.44%)
VPI-over-OSLA-Mis	973.92 (40.71%)
OSLA-Uni	1562.86 (65.32%)
OSLA-Mis	1253.8 (52.4%)
MAP-Uni	1588.6 (66.4%)
MAP-Mis	1459.4 (61%)
Myopic-Uni	674.44 (28.2%)
Myopic-Mis	723.76 (30.25%)

6.2 Dynamic tasks

We now consider a setting in which agents face dynamic tasks, in other words, where the possible coalitional actions, their dynamics, or their rewards can change from stage to stage of the repeated interaction. Such dynamic tasks [32, 37, 51, 54] are an important aspect of coalition and team formation. The ability of our RL framework to allow agents to learn about the *types* of other agents facilitates the *transfer of knowledge* between tasks. Indeed, this is one of the major benefits of a model that assumes type uncertainty and uses learning to tackle it: once agents learn about the abilities of partners, they can re-use this knowledge when encountering those partners under different circumstances.

We test this ability in a setting with five agents (again, homogeneous in the RL method used), which form coalitions over 500 stages. Coalitions are formed using the BRE method (50 full negotiation steps). There are five types, each agent having a *distinct* type (different from the other four) but not aware of this constraint. Agents share a uniform prior regarding the types of their opponents (but know their own types). Coalitions have three actions at their disposal with three possible outcomes each. However, the outcome probabilities vary from one RL stage to another (see below) reflecting different tasks or environments. We assume

Table 4 Experiments with 10 agents. Average “per step” reward accumulated within the final 50 RL steps of a run. *Uni* uniform, *Mis* misinformed prior

Method	Reward
(a) Full negotiations	
Fully informed agents	3884.77
VPI-Uni	2987.6 (76.9%)
VPI-Mis	2893.6 (74.48%)
VPI-over-OSLA-Uni	1622.5 (41.76%)
VPI-over-OSLA-Mis	1768.86 (45.53%)
OSLA-Uni	1564.6 (40.27%)
OSLA-Mis	1669.4 (42.97%)
MAP-Uni	2736 (70.42%)
MAP-Mis	2144.34 (55.2%)
Myopic-Uni	2419.4 (62.28%)
Myopic-Mis	2235.2 (57.54%)
(b) One-step proposals	
Fully informed agents	3881.7
VPI-Uni	2764.2 (71.21%)
VPI-Mis	2736.4 (70.49%)
VPI-over-OSLA-Uni	1642.88 (42.32%)
VPI-over-OSLA-Mis	1710.96 (44.08%)
OSLA-Uni	1542.4 (39.74%)
OSLA-Mis	1541.8 (39.72%)
MAP-Uni	2657.58 (68.46%)
MAP-Mis	1660.7 (42.78%)
Myopic-Uni	1078.68 (27.8%)
Myopic-Mis	1462.8 (37.68%)

Table 5 The convergence to BC results (converged/30 runs) for the algorithms (for 5 agents). “Convergence” is assumed if at least 50 consecutive RL trials before a run’s termination result in a BC configuration

	FN Unif.	FN Misinf.	OSP Unif.	OSP Misinf.
MAP	27/30	0/30	14/30	0/30
Myopic	0/30	0/30	1/30	2/30
VPI	0/30	0/30	1/30	3/30
OSLA	0/30	0/30	2/30	2/30
VPI-over-OSLA	0/30	0/30	0/30	0/30

agents know which task they will face at which stage so that the dynamics of the underlying POMDP is known and can be factored into the Bellman equations.⁸

⁸ The tasks at each stage need not be known with certainty; a distribution over tasks is sufficient to apply our model. If the task distribution is not known, the POMDP model will break down. However, we can still apply our methods as we discuss later in this section.

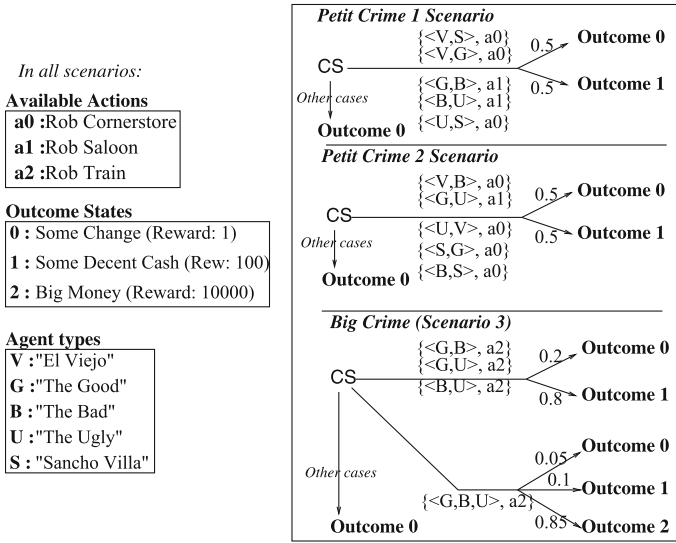


Fig. 5 The good, the bad and the ugly

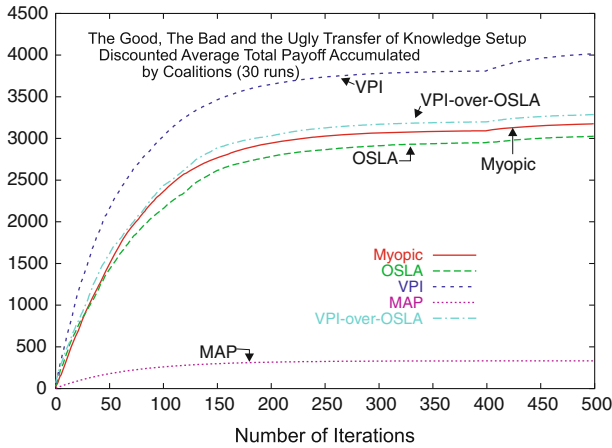


Fig. 6 Transfer of knowledge setup: discounted accumulated reward results

The precise set up is as follows: the agents are five bandits in the Wild West trying to form a successful gang. The ultimate goal of the three agents of types “Good”, “Bad” and “Ugly” is to discover each other and come together to “Rob the Train” (coalitional action), so as to get the “Big Money” (outcome). Before doing so, they will go through an experience-gathering phase, during which it is possible to coalesce with other villains (“El Viejo” and “Sancho Villa”), performing “petit crime” actions of lesser significance (such as “Rob Cornerstore” or “Rob Saloon”) which may result in “Some Change” or “Some Decent Cash” (outcomes) states—given the coalition qualities and underlying stochasticity. The setup is summarized in Fig. 5.

During the experience-gathering phase, i.e., the first 400 RL stages, the bandits are faced with problems 1 and 2 in an alternating fashion, with each problem having its own, distinct

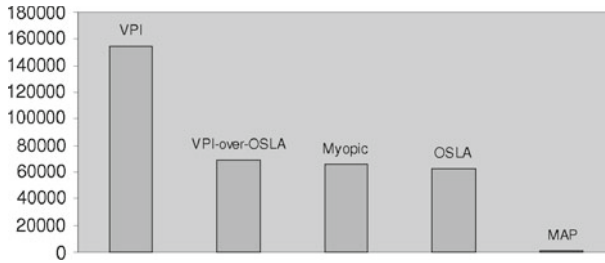


Fig. 7 Transfer of knowledge setup: rewards gathered during the “Big Crime” phase (averaged over 30 runs)

outcome transition model. They face problem 3 during the last 100 RL stages: this is the “Big Crime” phase of the experiment). By the time stage 401 is reached, they should have gained enough experience to form good coalitions to tackle problem 3 (through identifying each other correctly) and fare well in their “Big Crime;” if not, they will only make “Some Change” during this phase. Specifically, if all of them form a coalition and decide to rob the train, they have 85% probability of making Big Money; if only two of them form a coalition, they can expect, with 80% probability, to make Some Decent Cash by taking that same action. Problems 1 and 2, by contrast, suggest that agents should form two-agent coalitions, so that they get information regarding their partners’s types.

Results are presented in Figs. 6 and 7. VPI dominates the other methods both in terms of discounted accumulated rewards (i.e., behaviour during the “experience-gathering” phase), and also in terms of accumulated rewards during the final phase of the experiment. Performance of the lookahead methods and Myopic are not unreasonable, but clearly worse than VPI (and none get the same bump in performance during the big crime phase experienced by VPI). MAP agents appear to be utterly confused by the setup. These results illustrate that our framework, and the VPI algorithm specifically, supports the transfer of learned knowledge in coalition formation across different tasks/settings.

We repeated the experiment with the following change: agents are now unaware of the order in which they would be presented with tasks (in other words, they cannot predict what action transition model will be in place at any stage except the current one). In order to facilitate the computation required by lookahead methods, each agent assumes that the current transition model will be encountered. In this model, OSLA and VPI-over-OSLA agents are unable to accurately evaluate 1-step Q-values, since they have incorrect beliefs regarding the coalition formation problem to be faced at their successor belief states.

Results are presented in Figs. 8 and 9. Again VPI dominates the other methods in all respects. Unsurprisingly, OSLA and VPI-over-OSLA fare much more poorly given their inability to accurately evaluate Q-values. Nevertheless, the OSLA and VPI-over-OSLA agents do manage to collect, in the last phase of the experiment, approximately 10 and 6 times more reward, respectively, than the MAP agents; hence they still exhibit some knowledge transfer.

6.3 Comparison to kernel-based coalition formation

While no existing work prior to ours combines dynamic coalition formation with learning under type uncertainty, Kraus et al. [34] have dealt with coalition formation under uncertainty over coalitional values in a specific domain.⁹ Though their method is better tailored to set-

⁹ We explain the domain details later in Sect. 7.

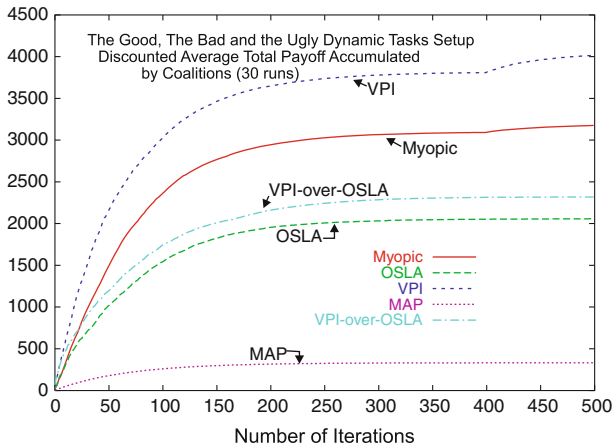


Fig. 8 The good, the bad and the ugly: discounted accumulated reward results

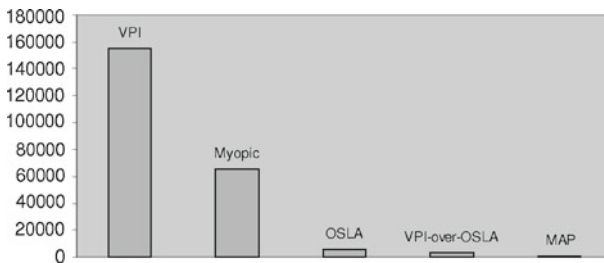


Fig. 9 The good, the bad and the ugly: rewards gathered during the “Big Crime” phase (averaged over 30 runs)

tings focusing on social welfare maximization, it is a rare example of a discounted coalitional bargaining method under a restricted form of uncertainty, which combines heuristics with principled game theoretic techniques.

We compare our Bayesian RL methods with an algorithm inspired by the work Kraus et al. [34]. We adapt their method to our repeated coalition setting with type uncertainty, referring to the modification as the *KST* algorithm. The method essentially computes an approximation of a kernel-stable allocation for coalitions that are formed during the negotiation phase of the RL process, with agents intentionally compromising part of their payoff to successfully form coalitions. The level of compromise is determined by a “compromise factor,” and following [34], our *KST* algorithm uses a compromise factor of 0.8. We assume no central authority, and have only one agent proposing per round, with coalition values estimated given type uncertainty.

A direct comparison of our techniques with [34] would not be appropriate, since it uses no learning and is not designed to handle type uncertainty and other aspects of our setting (e.g., that work makes certain heuristic assumptions which are inappropriate here, such as computing the kernel for the coalition with the greatest coalitional value, even though this might not at all be the coalition ensuring the highest payoff to the agent). Nevertheless, after some adaptations, it can serve as a useful benchmark, exhibiting the benefits of learning versus non-learning approaches to repeated coalition formation. We also combined *KST* with our Myopic RL algorithm, treating *KST* as its dynamic coalition formation component, in

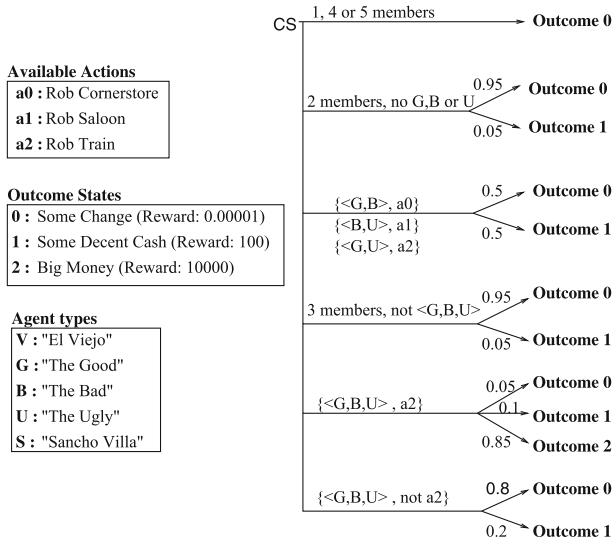


Fig. 10 Setup for the fourth set of experiments (comparison to the KST method)

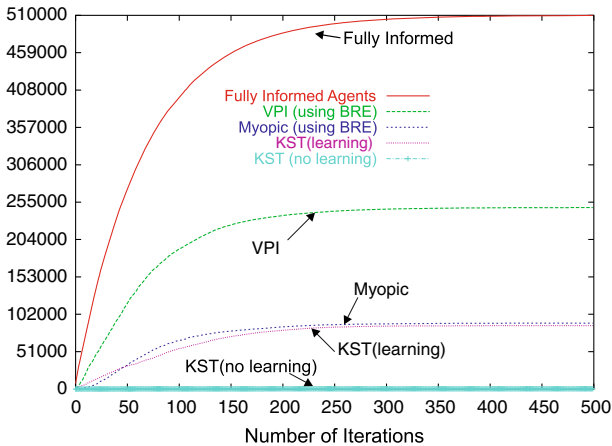


Fig. 11 Comparison with the (adapted) KST coalition formation approach. “KST(learning)” is Myopic RL having KST as its coalition formation component. The y axis shows discounted average accumulated reward gathered in 30 runs

an attempt to assess whether there are any clear distinctions between core-based or kernel-based coalition formation in our RL setting. In this respect, we can view KST as a myopic RL method, using kernel-stable payoff allocations and compromise [34].

We compared these methods on a setting with five agents, five types, and each agent having a different type. The setup is shown in Fig. 10. Agents have uniform prior beliefs. We compare KST (with and without learning) to our VPI method (since this is the method that performed best in the preceding experiments) and to Myopic (since KST is essentially a myopic method). As above, 50 negotiation steps are used to form coalitions, and we run the process for 1000 RL stages (30 runs). Results are shown in Fig. 11. Clearly, VPI (using BRE) is the best method in this domain, achieving total discounted reward close to 50% of the

maximum achievable by fully informed agents.. When agents use no learning, performance is very poor: KST (no learning) achieves negligible reward. Myopic RL seems to work equally well whether a core-based (BRE) or a kernel-based coalition formation process is used (the core-based approach does marginally, but not significantly, better).

7 Related work

In this section we review some related work from both the AI and game theory communities and, where appropriate highlight, the differences from our work.

Shehory and Kraus [51, 52] develop coalition formation algorithms which account for the capabilities of agents. However, agents rely on information communicated to them by their potential partners in order to form initial estimates of the capabilities of others. Pay-off allocation is addressed in [52], where they present two algorithms for self-interested agents in non-superadditive environments. The algorithms deal directly with expected payoff allocation, and the coalition formation mechanisms used are based on the kernel stability concept. Shehory et al. [53] also assume communication to inform agents of the capabilities of partners, and develop coalition formation algorithms to achieve agent collaboration in the RETSINA framework, so that tasks of common interest are executed successfully. This work focuses on serving the needs of the team (i.e., social welfare) and does not deal with payoff allocation issues.

Kraus et al. [33, 34] propose heuristic methods for coalition formation in a “Request for Proposal” domain, modeling a restricted form of coalitional value uncertainty (as opposed to agent type uncertainty). In the RFP domain, agents come together to perform tasks comprised of subtasks, each to be performed by a different agent. The agents may not know the value of a subtask to another agent or the cost of performing it, but they know the overall payoff associated with performing a task and the capabilities of the other agents. A kernel-based approach is combined with the use of “compromises” for the payoff allocation. Their focus is again social-welfare maximization rather than individual rationality. Furthermore, no learning is involved and repeated coalition formation is not addressed.

Campos and Willmott [37] address iterative coalition formation. They define “iterative coalition games” in which different agents may possess different abilities that will collectively enable coalitions to complete a task that does not change over time. Agents are initially assigned to coalitions randomly. They do not concern themselves with the payoff allocation problem; instead, they use several pre-defined strategies for choosing coalition formation moves, based essentially on whether their current coalition is “winning” over several rounds of play. Those limitations make the approach basically static, and there is no attempt to employ learning to facilitate coalition formation. By contrast, Abdallah and Lesser [1] utilize reinforcement learning in their approach to “organization-based coalition formation.” They assume an underlying organization guides the coalition formation process, and Q-learning is used to optimize the decisions of coalition managers, who assess communication or action-processing costs. However, agents are assumed to be cooperative, and there is no attempt to solve the payoff allocation problem. Furthermore, managers are assumed to possess full knowledge of their “child” agents’s capabilities.

Banerjee and Sen [4] address uncertainty regarding payoffs to members entering a coalition. The authors do not concern themselves with the process of coalition formation itself or payoff allocation, but rather address only with the problem of “coalition selection”: an agent has imperfect summary information of the anticipated payoff for joining a coalition, and has to choose one coalition over another after a fixed number of allowed interactions

with them. This “summary information” is provided by a payoff-structure encoding in the form of a multinomial distribution over possible payoffs for joining the coalition. The proposed mechanism for choosing a coalition makes use of this distribution, and also employs an arbitration mechanism from voting theory to resolve ties. In the case of limited allowed interactions, the proposed mechanism notably outperforms a *maximization of expected utility* mechanism in terms of selecting the most beneficial coalition. If the interactions allowed are infinite, however, the former mechanism reduces to the latter.

Klusch and Gerber [32] design and implement a simulation-based dynamic coalition formation scheme which can be instantiated using different computational methods and negotiation protocols. Their framework can be employed for the development, implementation and experimental evaluation of different coalition formation and payoff allocation algorithms (even fuzzy or stochastic coalition formation environments). Several different algorithms for learning of optimal coalition coalitions using Bayesian and RL methods are explored as well [27]. As discussed in Sect. 2, Blankenburg et al. [9] introduce the concept of the fuzzy kernel to cope with uncertain coalition values in sequential, bilateral coalition negotiations. More recently, Blankenburg and Klusch [8] propose a coalition formation algorithm which enables agents to negotiate bilateral Shapley value-stable coalitions in uncertain environments. Both approaches, however, are restricted to static, non-adaptive coalition forming in uncertain environments. By contrast, Blankenburg et al. [7] have implemented a coalition formation process that allows the agents to progressively update *trust* values regarding others, by communicating their private estimates regarding task costs and coalition valuations. They use encryption-based techniques and develop a payment protocol that ensures agents have the incentive to truthfully report their valuations. However, the proposed mechanism involves extensive inter-agent communication, and its effectiveness relies on computing optimal coalition structures and kernel stable solutions (both intensive computationally). Our approach could incorporate theirs as the internal coalition formation “stage” of the larger RL process (though some extension would be necessary to allow for the overlapping coalitions their model admits).

In contrast to some of the methods described above, agents in our framework have the ability not only to dynamically choose the tasks they wish to tackle, but also to choose the proper way (action) to deal with them. The incorporation of task execution in our model can be realized by simply viewing the tasks as requiring the use of specific action sets. Tasks can be thought of as defining a relevant action set at a specific stage of the RL process, specifically, triggering the existence of those actions that could be used to accomplish the task in question. Therefore, we can abstract away tasks, folding them into the specification of action sets. Finally, we note that our approach can be used to enable agents to form the most suitable coalitions for a new problem “online” in the sense that knowledge acquired during execution of one task can be readily “transferred” to another as shown in the previous section. Thus, the agents do not require experience with a new problem before deciding on ways to attack it.

8 Conclusions

We have proposed a Bayesian multiagent reinforcement learning framework for (repeated) coalition formation under type uncertainty. The framework enables the agents to improve their ability to form useful coalitions through the experience gained by repeated interaction with others and observation of the effects of coalitional actions. Agents in our model maintain and update beliefs about the types of others, and make sequentially rational decisions that reflect their interests, accounting for both potential coalition formation activities, and the

potential choice of actions by the coalitions they join. We developed a POMDP formulation that enables agents to assess the long-term value of coalition formation decisions, accounting for: the value of potential collective actions, uncertainty regarding both the types of others and the outcomes of coalitional actions, and the need to choose actions and coalitions not only for their immediate value, but also for their value of information.

Our RL framework is generic, allowing the agents to dynamically form coalitions, serve tasks and transfer knowledge between them. Critically, our framework enables the agents to weigh their need to explore the abilities of their potential partners against their need to exploit knowledge acquired so far. Specifically, coalition participants are able to make informed, sequentially rational decisions (regarding both the bargaining and the coalitional actions to take, and taking into account the value of information of the various actions), balancing exploration of actions with exploitation of knowledge in repeated coalition formation scenarios. Our framework can in principle accommodate any underlying negotiation process, and is not tightly bound to any specific cooperative solution or equilibrium concept.

We developed and evaluated several RL algorithms, each based on different computational approximations or assumptions. Our experiments demonstrate the effectiveness of our Bayesian approach, and of the concrete RL algorithms, under a variety of assumptions. Our Bayesian VPI technique, in particular, proved to be very robust, outperforming other methods consistently, and exhibiting very good computational performance. It works well with both full negotiation and one-step proposals, under conditions of high stochasticity, and when the initial beliefs held by agents is poor or misleading. Furthermore, it supports the effective transfer of knowledge among diverse tasks.

In future work, we intend to test our algorithms in open distributed environments. More specifically, experimenting in such environments with a greater variety of problem sizes, and different degrees of accuracy of agents's prior beliefs is of interest, as is experimenting with multi-step lookahead methods.

We believe our framework and algorithms are well-suited to realistic, complex task allocation environments, such as environments requiring the formation of coalitions to provide services in the computational grid [40] under time constraints that will not admit intensive computation [51]. In these settings, we expect VPI to be the preferred technique for several reasons. Our lookahead approaches have intense computational requirements and are not well-suited to such settings. Furthermore, in a variety of settings, priors may be uninformative and transition models may not be particularly discriminative (i.e., may not strongly distinguish the types of team members). Our experiments showed Myopic or MAP agents to perform poorly in such situations. In contrast, VPI appears to be far more robust, balancing value of information with immediate payoff. Consequently, we expect VPI to prove to be superior in realistic, repeated coalition formation environments. Indeed, these ideas were recently recast in a large computational trust setting by Teacy et al. [59], who report very encouraging results when applying our method to the exploration–exploitation problem faced by Bayesian agents that have to choose trusted information providers over time in a sequentially optimal manner. A variant of our VPI algorithm, assuming a continuous type space, was shown there to dramatically outperform all finalists in the 2006 and 2007 International Agent Reputation and Trust Competition [26]. Their experiments, involving up to 60 agents, showed VPI to operate in near-linear time.

To handle large-scale problems, another interesting direction is the extension of our model to include reasoning about the cost of computation [29, 46] as part of the inferential process. An agent using an approximate method for inference might specify the cost of improving that approximation by using more computation. The computation thus would have a value, arising as the expected gains or losses incurred by its use.

Theoretical directions include extending the model to include simultaneous learning of types and transition/reward models. Despite its intractability in general, we are also interested in developing a full Bayesian extensive form game formulation of the repeated coalition problem for specific forms of coalitional bargaining, analyzing its equilibria—at least in certain simple cases—and developing practical computational approximations.

Acknowledgements We thank the anonymous referees for their very helpful suggestions on improving the presentation and positioning of this work. The first author acknowledges the support of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Networks) project. ALADDIN is funded by a BAE Systems and EPSRC strategic partnership (EP/C548051/1). The second author acknowledges support of the Natural Sciences and Engineering Research Council (NSERC).

Appendix A: Experimental parameters from Sect. 6.1

See Tables 6, 7, 8, 9, and 10

Table 6 Symbols used in tables describing transition functions (for the first experimental setting in Sect. 6)

a: “action”; *s*: “state”; *q*: quality points; *: any;
N: number of coalition members
penalty = $N * 0.1$: penalty to discourage employing “cheap” workers
N_{MT}: number of different “major” types present in coalition
SP: small profit state;
AP: average profit state;
LP: large profit state
BFS: bid for small project action;
BFA: bid for average project action;
BFL: bid for large project action

Table 7 Participants in the five-agent experiments

Agent	Type	Quality points
0	Expert interface designer	2
1	Good programmer	2
2	Expert systems engineer	2
3	Bad programmer	0
4	Bad systems engineer	0

Table 8 Participants in the ten-agent experiments

Agent	Type	Quality points
0	Expert interface designer	2
1	Good programmer	2
2	Expert systems engineer	2
3	Bad programmer	0
4	Bad systems engineer	0
5	Bad interface designer	0
6	Average interface designer	1
7	Average programmer	1
8	Average systems engineer	1
9	Bad programmer	0

Table 9 Outcome transition function for 5-agent environments (for the first experimental setting in Sect. 6)

1-member coal.		$\Pr(LP a = *, q) = 0$ $\Pr(AP a = BFS, q) = q * 0.02$ $\Pr(AP a = BFA, q) = q * 0.01$ $\Pr(AP a = BFL, q) = 0$ $\Pr(SP a, q) = 1 - \Pr(AP a, q)$
2-member coal.		if $N_{MT} < 2$ then $q = q/2$ $\Pr(LP a = *, q) = 0$ $\Pr(AP a = BFS, q) = q * 0.04$ $\Pr(AP a = BFA, q) = q * 0.02$ $\Pr(AP a = BFL, q) = 0$ $\Pr(SP a, q) = 1 - \Pr(AP a, q)$
3-member coal.	if $N_{MT} < 3$ then:	if $N_{MT} = 1$ then $q = q/3$ if $N_{MT} = 2$ then $q = q/2$ $\Pr(LP a = *, q) = 0$ $\Pr(AP a = BFS, q) = q * 0.06$ $\Pr(AP a = BFA, q) = q * 0.02$ $\Pr(AP a = BFL, q) = q * 0.01$ $\Pr(SP a, q) = 1 - \Pr(AP a, q)$
	if $N_{MT} = 3$ then:	$\Pr(LP a = BFS, q) = q * 0.01$ $\Pr(LP a = BFA, q) = q * 0.04$ $\Pr(LP a = BFL, q) = q * 0.05$ $\Pr(SP a, q) = (1 - \Pr(LP a, q))/(q + 1)$ $\Pr(AP a, q) = 1 - \Pr(LP a, q) - \Pr(SP a, q)$
4 or 5-member coal.	if $N_{MT} < 3$ then:	if $N_{MT} = 1$ then $q = q/3$ if $N_{MT} = 2$ then $q = q/2$ $\Pr(LP a = *, q) = 0$ $\Pr(AP a = BFS, q) = q * 0.03$ $\Pr(AP a = BFA, q) = q * 0.05$ $\Pr(AP a = BFL, q) = q * 0.03$ $\Pr(SP a, q) = 1 - \Pr(AP a, q)$
	if $N_{MT} = 3$ then:	$\Pr(LP a = BFS, q) = q * 0.01$ $\Pr(LP a = BFA, q) = q * 0.04$ $\Pr(LP a = BFL, q) = q * 0.05$ $\Pr(SP a, q) = (1 - \Pr(LP a, q))/(q + 1)$ $\Pr(AP a, q) = 1 - \Pr(LP a, q) - \Pr(SP a, q)$

In all cases, $\Pr(SP|a, q)$, $\Pr(AP|a, q)$ and $\Pr(LP|a, q)$ are eventually normalized in order to sum to one

Table 10 Outcome transition function for 10-agent environments (for the first experimental setting in Sect. 6)

1 or 2-member coal.		As in a 5-agent environment
3-member coal.	if $N_{MT} < 3$ then:	if $N_{MT} = 1$ then $q = q/3$ if $N_{MT} = 2$ then $q = q/2$ $\Pr(LP a = *, q) = 0$ $\Pr(AP a = BFS, q) = q * 0.06$

Table 10 continued

1 or 2-member coal.	As in a 5-agent environment
	$\Pr(AP a = BFA, q) = q * 0.02$
	$\Pr(AP a = BFL, q) = q * 0.01$
	$\Pr(SP a, q) = 1 - \Pr(AP a, q)$
if $N_{MT} = 3$ then:	$\Pr(LP a = BFS, q) = q * 0.01$
	$\Pr(LP a = BFA, q) = q * 0.04$
	$\Pr(LP a = BFL, q) = q * 0.05$
	$\Pr(SP a, q) = (1 - \Pr(LP a, q))/(q + 1) + penalty$
	$\Pr(AP a, q) = 1 - \Pr(LP a, q) - \Pr(SP a, q)$
4,5,6 or 7-member coal.	if $N_{MT} < 3$ then:
	if $N_{MT} = 1$ then $q = q/3$
	if $N_{MT} = 2$ then $q = q/2$
	$\Pr(LP a = *, q) = 0$
	$\Pr(AP a = BFS, q) = q * 0.03$
	$\Pr(AP a = BFA, q) = q * 0.05$
	$\Pr(AP a = BFL, q) = q * 0.03$
	$\Pr(SP a, q) = 1 - \Pr(AP a, q)$
if $N_{MT} = 3$ then:	$\Pr(LP a = BFS, q) = q * 0.01$
	$\Pr(LP a = BFA, q) = q * 0.04$
	$\Pr(LP a = BFL, q) = q * 0.05$
	$\Pr(SP a, q) = (1 - \Pr(LP a, q))/(q + 1) + penalty$
	$\Pr(AP a, q) = 1 - \Pr(LP a, q) - \Pr(SP a, q)$
8,9 or 10-member coal.	if $N_{MT} < 3$ then:
	if $N_{MT} = 1$ then $q = q/3$
	if $N_{MT} = 2$ then $q = q/2$
	$\Pr(LP a = *, q) = 0$
	$\Pr(AP a = BFS, q) = q * 0.035$
	$\Pr(AP a = BFA, q) = q * 0.05$
	$\Pr(AP a = BFL, q) = q * 0.04$
	$\Pr(SP a, q) = 1 - \Pr(AP a, q)$
if $N_{MT} = 3$ then:	$\Pr(LP a = BFS, q) = q * 0.01$
	$\Pr(LP a = BFA, q) = q * 0.04$
	$\Pr(LP a = BFL, q) = q * 0.05$
	$\Pr(SP a, q) = (1 - \Pr(LP a, q))/(q + 1) + penalty$
	$\Pr(AP a, q) = 1 - \Pr(LP a, q) - \Pr(SP a, q)$

In all cases, $\Pr(SP|a, q)$, $\Pr(AP|a, q)$ and $\Pr(LP|a, q)$ are eventually normalized in order to sum to one

References

1. Abdallah, S., & Lesser, V. (2004). Organization-based coalition formation. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems (AAMAS'04)* (pp. 1296–1297).
2. Agastya, M. (1997). Adaptive play in multiplayer bargaining situations. *Review of Economic Studies*, 64, 411–426.

3. Aumann, R. J., & Myerson, R. B. (1988). Endogenous formation of links between players and of coalitions: An application of the shapley value. In A. E. Roth, *The shapley value* (pp. 175–191). Cambridge: Cambridge University Press.
4. Banerjee, B., & Sen, S. (2000). Selecting partners. In *Proceedings of the fourth international conference on autonomous agents, Barcelona, Catalonia, Spain* (pp. 261–262).
5. Barto, A. G., Bradtke, S. J., & Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72(1–2), 81–138.
6. Bellman, R. E. (1957). *Dynamic programming*. Princeton: Princeton University Press.
7. Blankenburg, B., Dash, R. K., Ramchurn, S. D., Klusch, M., & Jennings, N. R. (2005). Trusted Kernel-based coalition formation. In *Proceedings of the fourth international joint conference on autonomous agents and multiagent systems (AAMAS'05)* (pp. 989–996).
8. Blankenburg, B., & Klusch, M. (2005). BSCA-F: efficient fuzzy valued stable coalition forming among agents. In *Proceedings of the IEEE international conference on intelligent agent technology (IAT)*.
9. Blankenburg, B., Klusch, M., & Shehory, O. (2003). Fuzzy Kernel-stable coalitions between rational agents. In *Proceedings of the second international joint conference on autonomous agents and multiagent systems (AAMAS'03)* (pp. 9–16).
10. Chalkiadakis, G. (2007). *A Bayesian approach to multiagent reinforcement learning and coalition formation under uncertainty*. PhD thesis, Department of Computer Science, University of Toronto, Canada.
11. Chalkiadakis, G., & Boutilier, C. (2003). Coordination in multiagent reinforcement learning: a Bayesian approach. In *Proceedings of the second international joint conference on autonomous agents and multiagent systems (AAMAS'03)* (pp. 709–716).
12. Chalkiadakis, G., & Boutilier, C. (2004). Bayesian reinforcement learning for coalition formation under uncertainty. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems (AAMAS'04)* (pp. 1090–1097).
13. Chalkiadakis, G., & Boutilier, C. (2007). Coalitional bargaining with agent type uncertainty. In *Proceedings of the twentieth international joint conference on artificial intelligence (IJCAI-07)* (pp. 1227–1232).
14. Chalkiadakis, G., & Boutilier, C. (2008). Sequential decision making in repeated coalition formation under uncertainty. In *Proceedings of the seventh international joint conference on autonomous agents and multiagent systems (AAMAS'08)* (pp. 347–354).
15. Chalkiadakis, G., Markakis, E., & Boutilier, C. (2007). Coalition formation under uncertainty: bargaining equilibria and the Bayesian core stability concept. In *Proceedings of the sixth international joint conference on autonomous agents and multiagent systems (AAMAS'07)* (pp. 400–407).
16. Chatterjee, K., Dutta, B., & Sengupta, K. (1993). A noncooperative theory of coalitional bargaining. *Review of Economic Studies*, 60, 463–477.
17. Conitzer, V., & Sandholm, T. (2003). Complexity of determining non-emptiness of the core. In *Proceedings of the eighteenth international joint conference on artificial intelligence (IJCAI-03)*.
18. Davis, M., & Maschler, M. (1965). The Kernel of a cooperative game. *Naval Research Logistics Quarterly*, 12, 223–259.
19. Dearden, R., & Boutilier, C. (1997). Abstraction and approximate decision theoretic planning. *Artificial Intelligence*, 89, 219–283.
20. Dearden, R., Friedman, N., & Andre, D. (1999). Model based Bayesian exploration. In *Proceedings of fifteenth conference on uncertainty in artificial intelligence* (pp. 150–159).
21. Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth national conference on artificial intelligence (AAAI-98)* (pp. 761–768).
22. DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.
23. Deng, X., & Papadimitriou, C. (1994). On the complexity of cooperative solution concepts. *Mathematics of Operation Research*, 19, 257–266.
24. Dieckmann T., & Schwalbe, U. (1998). *Dynamic coalition formation and the core*. Economics Department Working Paper Series, Department of Economics, National University of Ireland, Maynooth.
25. Duff, M. O. (2002). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst.
26. Fullam, K., Klos, T. B., Muller, G., Sabater, J., Schlosser, A., Topol, Z., et al. (2005). A specification of the agent reputation and trust (art) testbed: Experimentation and competition for trust in agent societies. In *AAMAS* (pp. 512–518).
27. Gerber, A. (2005). *Flexible cooperation between autonomous agents in dynamic environments*. PhD thesis, Saarland University, Germany.
28. Gillies, D. B. (1953). *Some theorems on n-Person games*. PhD thesis, Department of Mathematics, Princeton University, Princeton.

29. Horvitz, E. (1990). *Computation and action under bounded resources*. PhD thesis, Stanford University.
30. Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.
31. Kahan, J. P., & Rapoport, A. (1984). *Theories of coalition formation*. Hillsdale, NJ: Lawrence Erlbaum Associates.
32. Klusch, M., & Gerber, A. (2002). Dynamic coalition formation among rational agents. *IEEE Intelligent Systems*, *17*(3), 42–47.
33. Kraus, S., Shehory, O., & Taase G. (2003). Coalition formation with uncertain heterogeneous information. In *Proceedings of the second international joint conference on autonomous agents and multiagent systems (AAMAS'03)* (pp. 1–8).
34. Kraus, S., Shehory, O., & Taase, G. (2004). The advantages of compromising in coalition formation with incomplete information. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems (AAMAS'04)* (pp. 588–595).
35. Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: John Wiley and Sons.
36. Mas-Colell, A., Whinston, M., & Green, J. R. (1995). *Microeconomic theory*. Oxford: Oxford University Press.
37. Merida-Campos, C., & Willmott, S. (2004). Modelling coalition formation over time for iterative coalition games. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems (AAMAS'04)* (pp. 572–579).
38. Myerson, R. B. (1991). *Game theory: Analysis of conflict*. Cambridge: Harvard University Press.
39. Okada, A. (1996). A noncooperative coalitional bargaining game with random proposers. *Games and Economic Behavior*, *16*, 97–108.
40. Patel, J., Teacy, W. T. L., Jennings, N. R., Luck, M., Chalmers, S., & Oren, N., et al. (2005). Agent-based virtual organisations for the grid. *Multiagent and Grid Systems*, *1*(4), 237–249.
41. Poupard, P., & Vlassis, N. (2008). Model-based bayesian reinforcement learning in partially observable domains. In *International symposium on artificial intelligence and mathematics (ISAIM), Fort Lauderdale, FL*.
42. Price, B. (2003). *Accelerating reinforcement learning with imitation*. PhD thesis, University of British Columbia.
43. Puterman, M. L. (1994). *Markov decision processes*: Wiley.
44. Rapoport, A. (1970). *N-Person game theory*. MI: University of Michigan Press.
45. Rapoport, A., Kahan, J. P., Funk, S. G., & Horowitz, A. D. (1979). *Coalition formation by sophisticated players*. NY: Springer.
46. Russell, S., & Wefald, E. (1991). *Do the right thing: Studies in limited rationality*. Cambridge, MA: The MIT Press.
47. Sandholm, T., Larson, K., Andersson, M., Shehory, O., & Tohme, F. (1999). Coalition structure generation with worst case guarantees. *Artificial Intelligence*, *111*(1–2), 209–238.
48. Sandholm, T., & Lesser, V. R. (1997). Coalitions among computationally bounded agents. *Artificial Intelligence*, *94*(1), 99–137.
49. Satia, J. K., & Lave, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, *21*, 728–740.
50. Shapley, L. S. (1953). A value for n-Person games. In H. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II* (pp. 307–317). Princeton: Princeton University Press.
51. Shehory, O., & Kraus, S. (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence*, *101*(1–2), 165–200.
52. Shehory, O., & Kraus, S. (1999). Feasible formation of coalitions among autonomous agents in nonsuperadditive environments. *Computational Intelligence*, *15*, 218–251.
53. Shehory, O., Sycara, K., & Jha, S. (1997). Multiagent coordination through coalition formation. In *Agent theories, architectures and languages* (pp. 143–154).
54. Soh, L.-K., & Li, X. (2004). Adaptive, confidence-based multiagent negotiation strategy. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems (AAMAS'04)* (pp. 1046–1053).
55. Sondik, E. J. (1978). The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Operations Research*, *26*, 282–304.
56. Suijs, J., & Borm, P. (1999). Stochastic cooperative games: Superadditivity, convexity and certainty equivalents. *Journal of Games and Economic Behavior*, *27*, 331–345.
57. Suijs, J., Borm, P., De Wagenaere, A., & Tijs, S. (1999). Cooperative games with stochastic pay-offs. *European Journal of Operational Research*, *113*, 193–205.
58. Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*: MIT Press.

59. Teacy, W. T. L., Chalkiadakis, G., Rogers, A., & Jennings, N. R. (2008). Sequential decision making with untrustworthy service providers. In *Proceedings of the seventh international joint conference on autonomous agents and multiagent systems (AAMAS'08)*. (pp. 755–762).
60. von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
61. Wellman, M. P. (2006). Methods for empirical game-theoretic analysis. In *Proceedings of the 21st national conference on AI (AAAI-06)* (pp. 1552–1555).