IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 September 2022, accepted 22 September 2022, date of publication 28 September 2022,
date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210506

RESEARCH ARTICLE

Indirect Dynamic Negotiation in the Nash Demand Game

TATIANA V. GUY[1], **(Senior Member, IEEE), JITKA HOMOLOVÁ**[2], **AND ALEKSEJ GAJ**[3]
[1]Department of Information Engineering, Faculty of Economics and Management, Czech University of Life Sciences, 165 21 Prague, Czech Republic
[2]Department of Adaptive Systems, Institute of Information Theory and Automation, Czech Academy of Sciences, 182 00 Prague, Czech Republic
[3]Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, 120 00 Prague, Czech Republic

Corresponding author: Tatiana V. Guy (guy@ieee.org)

This work was supported in part by the Grant LTC18075.

ABSTRACT The paper addresses a problem of sequential bilateral bargaining with incomplete information. We proposed a decision model that helps agents to successfully bargain by performing indirect negotiation and learning the opponent's model. Methodologically the paper casts heuristically-motivated bargaining of a self-interested independent player into a framework of Bayesian learning and Markov decision processes. The special form of the reward implicitly motivates the players to negotiate indirectly, via closed-loop interaction. We illustrate the approach by applying our model to the Nash demand game, which is an abstract model of bargaining. The results indicate that the established negotiation: i) leads to coordinating players' actions; ii) results in maximising success rate of the game and iii) brings more individual profit to the players.

INDEX TERMS Learning, Markov decision process, Nash demand game, negotiation.

I. INTRODUCTION

Politics and business are considered traditional spheres of human negotiation. The internet and modern means of communication have extended human negotiation to new domains such as social networks, deliberative democracy, e-commerce, cloud-based applications, [1], [2]. Besides, automatic bargaining and negotiation, being inevitable in modern cyber-physical-social systems [3], have been established in variety of applications, like network negotiation, energy trading [4] and traffic management [5], multi-robot systems [6], manufacturing service allocation [7] and newly in ransomware negotiation [8]. While solving negotiation task, agents must take into account incomplete information and strategically interact with other, human or artificial, agents. Majority of the existing research however assumes negotiation with non-human agents.

Here we consider the simplest bilateral bargaining scenario with incomplete information often found in e-commerce [9]. A typical example is two self-interested agents (say, a buyer and a seller) bargaining on some goods or service. As soon as their price preferences differ, agents begin negotiations to achieve a mutually acceptable price. Either agent strives to satisfy own preferences as much as possible, but also has to take into account the opponent's preferences. Otherwise it is unlikely that an agreement can be reached.[1] Additional aspects of real-life bilateral bargaining to be considered are: i) *multi-attribute negotiation* when agents need to agree on goods/service characterised by several, possibly interrelated, attributes (say price of a product and terms of its delivery); ii) *limited negotiation time* as no agent can deliberate infinitely; iii) *absence of moderator* to coordinate the negotiation, so the agents must reach agreement themselves [11].

The negotiation has been widely addressed in diverse fields ranging from economy and sociology to computer science. An amount of works is much too large to survey them here. One can distinguish several main frameworks: game theoretic approach, negotiation protocols approach, evolutionary approach. Existing works however have different limitations preventing them from wide use. Game theoretic approach [11], [12], assumes that agents are perfectly rational and have common knowledge. Negotiation protocols approach, [13], needs the clear rules for negotiation, [14], and the results largely depend on the information available to the

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

[1] For details on modelling bargaining, see for instance [10].

agents about each other. Evolutionary approach, [15], being inspired by biological evolution, finds optimal negotiation via trial-error and agents should have access to policy of their opponents and their profits. Some approaches are based on an agent-coordinator responsible for assigning goods or services to agents. This coordinating (or planning) agent uses a negotiation mechanism to find the best share.

We consider a finite horizon bilateral sequential bargaining of two independent self-interested Bayesian decision making (DM) agents facing with incomplete information. The **key aspects of the targeted solution** are as follows.

- *Negotiation.* The purpose of negotiation is to enable agents to coordinate their actions/decisions. Thus negotiation is a *means* to achieve coordinated behaviour of the agents. We consider the ability to negotiate an *intrinsic* part of an agent and treat it accordingly. The proposed solution allows indirect negotiation via information feedback and further leads to coordinated behaviour without conventional (explicit) negotiation.
- *Domain-independence.* Existing solutions are either of domain-specific, [16], or domain-independent, [17]. The former ones may be more effective, but tailoring them to a new domain may often be useless. The ever-growing number of new applications make domain-tailored solutions less favourable. The considered Bayesian DM agent is inherently domain-independent.
- *Modelling and learning the opponent.* Incomplete knowledge is given by uncertainty regarding the opponent's preferences and behaviour. This uncertainty may prevent agents from reaching mutually beneficial agreement as well as own DM goals. The proposed solution uses Bayesian approach to dynamically learn opponent's model based on observed actions (bids).
- *Bounded rationality.* Assumption on perfect rationality used by game theoretic approach is not valid in real-life tasks. Moreover human agents often behave seemingly irrational due to cognitive or social factors [18]. Their DM is also influenced by emotional state [19] and personal traits [20]: self-interest, altruism, ability to cooperate. The proposed solution is general enough and has already proven to take into account human-like factors [21]. Thus the approach can serve both an artificial agent and a human.

Other important aspect of the negotiation problem concerns *limited deliberation*. Obviously, no agent can bargain indefinitely so the DM policy that is being designed must take that into account. It is hardly possible to set flexible limits on the length of negotiations, but we believe that the established internal feedback complemented by stopping rule can adaptively influence the length of negotiations. A natural decrease of the utility of goods/service over time can also be counteracted by introducing a kind of forgetting [22] in the utility function.

Main contributions. The paper contributes to research on bilateral bargaining in distributed settings. We propose a *self-interested probabilistic DM agent* maximising expected utility, that is able to purposefully negotiate. The developed agent is domain-independent, can serve to either human or artificial agents and is equipped with the following abilities (which indicate major contributions):

- ***Learning ability.*** To counteract incomplete knowledge and adapt to possible changes of its opponent, the agent is equipped with the learning ability. The algorithm is based on Bayesian approach and learns opponent's model from bargaining history, i.e. from the bids the opponent proposes during a negotiation. This allows to respect the opponent's dynamics as well as any other related uncertainty, cf. [23].
- ***Indirect negotiation.*** A key component of the proposed bargaining agent is a reward function that consists of two components. The first one respects a purely economical individual profit of the agent. The second component expresses degree to which bargaining agents exploit the game potential. It is important to note that the second component i) provides the agent with information feedback; ii) prompts the agents for indirect negotiation, and iii) set limits on the negotiation range. The trade-off between the individual profitability and the game potential is expressed by an agent-specific weight, cf. [24]. Naturally the opponent equipped with learning ability can model the weight and use this knowledge in next rounds. The weight expresses the agent preferences and partially reflects personal behavioural aspects of human bargaining. The latter opens an avenue for design of automated agents reflecting human traits, [25].
- ***Privacy preservation.*** The implicit nature of the resulting distributed interaction does not involve the exchange of any private data or models between players. Therefore the proposed approach fully preserves players' privacy.

The proposed solution also allows to incorporate prior knowledge of the opponent though does not require that. The methodology [26] makes it possible to use the available external or domain-specific knowledge to enrich the opponent's model. The paper also compares three types of prior knowledge reflecting typical cases and illustrate its use.

The paper continues our previous work [27] that assumes complete knowledge of the opponent model, which is rarely achievable in real-world applications. Thus the present paper focuses on learning the opponent model as well as on intrinsic motivation to cooperate. The last contribution of the work is that we have compared the performance of the proposed bargaining agents to agents employing heuristic models built on the extensive experimental meta-study [28].

Related research. The literature of negotiation constitutes a very large collection, and space limitations prevent it from being presented in its entirety here. Generally there are several models focusing on the explicit negotiation based either on game theory or negotiation theory. The proposed approach considers independent dynamical self-interested DM agents, with learning ability and special reward prompting ***indirect*** negotiation. The mentioned features are very practical and up to now missing within the otherwise well-elaborated and important area of the paper. Up to the authors best knowledge there is no similar approach. We use probabilistic models [29] of bargaining agents that interact in a closed-loop and admit Markov decision processes as a modelling methodology, cf. [30]. The area of agent negotiation and opponent modelling has a lot of achievements, see for instance [31], [32], [33]. The comprehensive survey can be found in [34] and in [35]. The recent paper [36] discusses main challenges and promises in the area. Most research on negotiating concerns *static* environments and focused on i) developing utility-based negotiation strategies for rational DM agents, see for instance [37], [38], and ii) creating agent-moderator helping DM agent in negotiation task, [17], [39], [40]. So far much less research describe negotiating in dynamic environment, see [41], [42]. The recent approach [43] uses a logistic regression for modelling the opponent, that requires collecting significant amount of data for learning and initialisation. Paper [44] uses a similar utility based on the bargaining principles though constructs a subgame that relies on the perfect equilibrium. The closely related work, dealt with opponent modelling, is probably [45]. It also employs Bayesian learning but relies on specific structure of preferences and policy of the opponent. Though work [46] also focuses on design of negotiation agents in dynamic and uncertain environments, it relies on a negotiation agent and proposes a set of heuristics to make negotiation decisions. Our model introduces an intrinsic mechanism that motivates the agent to negotiate while learning opponent's model via Bayesian approach. The resulting bargaining policy is optimal with respect to the resources available and individual preferences of the agents. It can also take into account human factors, which are important whenever human agents are involved.

We illustrate the approach using the Nash Demand Game (NDG) [12], a bilateral bargaining game for two players that should decide how to split given amount of money. The players simultaneously demand a certain portion of the amount they would like to get. The demand of one player is unknown to another one (an opponent). If the players' demands can be satisfied simultaneously, both players get the respective profit. Otherwise, they both get nothing. Despite its seeming simplicity, the NDG is a good model of dynamical resource allocation that achieves coordination without explicit negotiation. It also serves a big challenge for understanding human negotiation.

The remainder of the paper is organized as follows. Section II introduces notations and a mathematical background. Section III formulates the Nash Demand Game as MDP of a single player, introduces heuristic model of the opponent and prior models used in learning. Section IV describes and discuss simulated experiments. Section V and Section VI summarise the results obtained and outlines future research directions.

II. PRELIMINARIES

This section introduces and recalls necessary notions.

A. GENERAL CONVENTIONS

$\mathbb{N}, \mathbb{R}$	set of natural numbers, set of real numbers
$x_t \in \mathbf{X}$	value x from finite set $\mathbf{X}$ at discrete time t
$p(x)$	probability mass function of discrete random variable x
$p(x\|y)$	probability mass function of x conditioned on y
$E[x\|y]$	the expectation of x conditioned on y

Note that no notational distinction is made between a random variable and its realisation.

B. MARKOV DECISION PROCESS

We model player's decision making in the NDG via Markov Decision Process (MDP) framework [47]. MDPs were first introduced and developed in the operations research and economics [48]. Since that MDP framework has been widely used to describe and solve decision-theoretic problems. MDP allows to capture the underlying stochastics omnipresent in application domain and also allows to respect multiple DM criteria. Typical examples of using MDP framework include medical applications [49], predictive maintenance [50], power systems [51], more examples see [52].

The overall scenario is as follows. An player interacts with the environment by taking actions to achieve its[2] DM goal. The player is motivated by a reward it receives after each action taken. A finite state and action MDP is considered.

*Definition 1 (**MDP**):* The fully observable MDP is characterised by $\{\mathbf{T}, \mathbf{S}, \mathbf{A}, p, R\}$, where $\mathbf{T} = \{1, 2, \ldots, N\}, N \in \mathbb{N}$, is a set of decision epochs; $\mathbf{S}$ is a finite set of all possible environment states and $\mathbf{A}$ denotes a finite set of all actions available to the player. Function $p : \mathbf{S} \times \mathbf{S} \times \mathbf{A} \mapsto [0, 1]$ is the transition model $p(s_{t+1}|s_t, a_t)$ that moves the environment from state $s_t \in \mathbf{S}$ to state $s_{t+1} \in \mathbf{S}$ after the agent took action $a_t \in \mathbf{A}$; $R : \mathbf{S} \times \mathbf{S} \times \mathbf{A} \mapsto \mathbb{R}$ is a real-valued function representing the player's reward $R(s_{t+1}, s_t, a_t)$ after taking action $a_t \in \mathbf{A}$ in state $s_t \in \mathbf{S}$.

The transition model captures *environment dynamics* and is represented by a family of probability distributions $p(s_{t+1}|s_t, a_t)$, each denotes the probability that at time $t + 1$ the environment will move from s_t to s_{t+1} when action a_t is executed. The state transitions obey Markov property: the distribution over states at time $t + 1$ is independent of any previous state s_{t-j} and action $a_{t-j}, j \leq 1$ for fixed s_t and a_t.

The player's preferences are described by a reward function, R. The aim of the player is to choose a sequence of actions in order to maximise the total expected sum of rewards as described in the following section.

C. OPTIMAL DECISION POLICY

The player chooses action $a_t \in \mathbf{A}$ based on the *randomised DM rule* $p(a_t|s_t) : \mathbf{S} \mapsto \mathbf{A}$ in each decision epoch $t \in \mathbf{T}$.

[2] "It" is used as the generic pronoun. A device or an algorithm can be considered as the agent.

A sequence of DM rules forms *DM policy* $\pi_{t,h}$ at time t over decision horizon $h \in \mathbb{N}$, $s_\tau \in \mathbf{S}$, $a_\tau \in \mathbf{A}$:

$$\pi_{t,h} = \left\{ p(a_\tau|s_\tau) \middle| s_\tau, a_\tau, \sum_{a_\tau \in \mathbf{A}} p(a_\tau|s_\tau) = 1, \forall s_\tau \in \mathbf{S} \right\}_{\tau=t}^{t+h-1}. \quad (1)$$

MDP with finite horizon h evaluates the quality of DM policy by *expected total reward* defined as follows:

$$E\left[\sum_{\tau=t}^{t+h-1} R(s_{\tau+1}, s_\tau, a_\tau)|s_t\right] = \sum_{\tau=t}^{t+h-1} \sum_{\substack{s_{\tau+1} \in \mathbf{S} \\ s_\tau \in \mathbf{S} \\ a_\tau \in \mathbf{A}}} R(s_{\tau+1}, s_\tau, a_\tau) p(s_{\tau+1}, s_\tau, a_\tau|s_t), \quad (2)$$

where

$$p(s_{\tau+1}, s_\tau, a_\tau|s_t) = p(s_{\tau+1}|s_\tau, a_\tau) p(a_\tau|s_\tau) p(s_\tau|s_t).$$

The solution to MDP [47] is a sequence of DM rules, $\left\{p^{opt}(a_\tau|s_\tau)\right\}_{\tau=t}^{t+h-1}$, that maximises the expected reward (2) and forms the optimal decision policy:

$$\pi_{t,h}^{opt} = \arg \max_{\{\pi_{t,h}\} \in \boldsymbol{\pi}} E\left[\sum_{\tau=t}^{t+h-1} R(s_{\tau+1}, s_\tau, a_\tau)|s_t\right], \quad (3)$$

where $\boldsymbol{\pi}$ is a set of possible DM policies, see (1). The optimal policy (3) is computed by dynamic programming algorithm [48], [53], which requires knowledge of transition model $p(s_{\tau+1}|s_\tau, a_\tau)$.

D. LEARNING TRANSITION MODEL

In bilateral bargaining, the transition model is a model of the opponent, that is, it predicts the opponent's reaction to the player's action. Generally it describes the dynamics of the opponent's decision making. In real-life tasks, opponent model $p(s_{t+1}|s_t, a_t)$ is usually unknown to the player.[3] It reflects the player's knowledge about the behaviour of the opponent. Without lost of generality the model can be assumed time-invariant, i.e. $p(s_{t+1}|s_t, a_t) = p(s_t|s_{t-1}, a_{t-1})$ and can be learned from the observed data.

To simplify the presentation, let us drop out the time index and introduce the following temporary notations: $s' = s_{t+1}$, $s = s_t$ and $a = a_t$. The transition model then can be written $p(s'|s, a)$.[4]

We consider a *parametrised form* of the opponent's model with time-invariant parameter $\theta \in \Theta$

$$p(s'|s, a, \theta) = \theta_{s'sa}, \quad \theta_{s'sa} \in \Theta, \quad (4)$$

where Θ is a set of all possible θ's and $0 \leq \theta_{s'sa} \leq 1$, $\sum_{s' \in \mathbf{S}} \theta_{s'sa} = 1, \forall (s, a)|s \in \mathbf{S}$ and $a \in \mathbf{A}$.

Thus, parameter θ in (4) is an array defining transition probabilities $\theta_{s'sa}$ that opponent's state in the next time will equal s' whenever the previous state is s and the player takes action a. Our aim is to learn parameter θ, (4).

Let the player have belief $b(\theta)$ about the opponent's dynamics expressed via the probability density function of the parameter θ. While interacting with the opponent, the player updates belief about the parameter, $b(\theta)$, to a new value, $b'(\theta)$, given observed transition (s', s, a) as follows, see [54]:

$$b'(\theta) \propto b(\theta) p(s'|s, a, \theta) = b(\theta)\theta_{s'sa}. \quad (5)$$

Choosing belief $b(\theta)$ in conjugate form of Dirichlet distribution implies that the posterior (5) induced by Bayes' rule [54] is

$$\mathrm{Dir}(\boldsymbol{\nu}, \theta) \propto \prod_{s'sa} \theta_{s'sa}^{\nu_{s'sa}-1}. \quad (6)$$

In (6) concentration parameter $\boldsymbol{\nu} > 0$ is an array containing occurrences $\nu_{s'sa} > 0$ of triples (s', s, a). Each observation of a triplet (s', s, a) increases the corresponding entry, $\nu_{s'sa}$, by one.

Therefore, after $n \in \mathbb{N}$ observations $\{(s', s, a)\}_{n \in \mathbb{N}}$, update $\nu'_{s'sa}$ contains the actual occurrences of (s', s, a). Recalling (4), the expectation of (6) can be interpreted as Bayesian estimate of unknown parameter θ based on the observed data (i.e. transitions occurred):

$$E\left[p(s'|s, a, \theta)\middle|\nu'\right] = E\left[\theta_{s'sa}|\nu'\right] = \frac{\nu'_{s'sa}}{\sum_{s'} \nu'_{s'sa}}. \quad (7)$$

Recursive implementation of the prior statistics update is described in [55].

A real-life dynamic decision making requires an efficient and feasible learning that can be performed online. Markov models belong to the exponential family for which exact estimation is feasible. The estimation and prediction within this family is very simple, especially with the conjugate prior in the form of Dirichlet distribution. The needed update of functions (probability density functions, see (5)) is given by the algebraic recursive update of the finite dimensional sufficient statistics. This clarifies applicability of this learning in combination with decision making.

III. METHODOLOGY

A. MDP FORMALISATION OF NASH DEMAND GAME

The considered repetitive scenario of the game is as follows. Two structurally identical players $\mathcal{A}$ and $\mathcal{B}$ are bargaining on splitting an amount of money $q \in \mathbb{N}$. The roles of both players are the same. In each round, two stages are present: an *action* stage and a *reward* stage. During *action* stage, each player decides how much to claim from the total available amount. The players do not communicate and their interests can be competitive. At *reward* stage, the players announce their demanded shares, observe the demands of their opponents and reward is allocated. Note that in action stage each player has no information about their opponent's demand or preferences. The game runs for a fixed and known number of periods.

[3] It can be partially known or incorrectly specified.

[4] The new notation is valid within Section II-D only.

Let $q \in \mathbb{N}$ is a total amount to split. At the beginning of round $t \in \mathbf{T}$, each player $k \in \{\mathcal{A}, \mathcal{B}\}$ chooses action $a_t^k \in \mathbf{A}^k$ that is a demanded share of q in the round. The minimum demand equals 1 and the maximum is $q - 1$. If the sum of demands is less than or equal to q, both players get what they asked for, otherwise the players get zero reward.

Player's profit in round $t \in \mathbf{T}$ equals the amount of money player receives[5]:

$$\begin{aligned} z_t^{\mathcal{A}} &= a_t^{\mathcal{A}} \chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}), \\ z_t^{\mathcal{B}} &= a_t^{\mathcal{B}} \chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}), \end{aligned} \tag{8}$$

where $z_t^{\mathcal{A}}, z_t^{\mathcal{B}} \in \mathbf{Z}$ are profits of $\mathcal{A}$ and $\mathcal{B}$ respectively. $\mathbf{Z} = \{0, 1, 2, \ldots, q-1\}$ is a set of possible profits in one game round, and

$$\chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) = \begin{cases} 1 & \text{if } a_t^{\mathcal{A}} + a_t^{\mathcal{B}} \leq q, \\ 0 & \text{if } a_t^{\mathcal{A}} + a_t^{\mathcal{B}} > q. \end{cases} \tag{9}$$

The addressed distributed bargaining does not consider communication between the players or any agent-moderator. To find a fully *distributed* solution, the game is described from a point of view of a stand-alone player. Let us now formulate the discussed bargaining task of a stand-alone player, say player $\mathcal{A}$, as an MDP problem.

Definition 2 (Bargaining as an MDP Task): The bargaining scenario is modelled by tuple $\{\mathbf{T}, \mathbf{S}, \mathbf{A}, p, R\}$, see Definition 1, where $\mathbf{A} = \{1, 2, \ldots, q-1\}$ is a set of possible actions; $a_t^{\mathcal{A}} \in \mathbf{A}$ is *action* of player $\mathcal{A}$, i.e. a portion of q demanded by $\mathcal{A}$ at time $t \in \mathbf{T}$; $s_t = (a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}}) \in \mathbf{S}$ is a *state* observed by $\mathcal{A}$ at time t and $p(s_{t+1}|a_{t-1}^{\mathcal{A}}, s_t)$ is a transitional model that describes the state dynamics. Initial state $s_1 = (a_0^{\mathcal{A}}, a_0^{\mathcal{B}})$ is preset to the same demand $a_0^{\mathcal{A}} = a_0^{\mathcal{B}} = a_0$.

Reward as motivation for negotiation. Let reward of player $\mathcal{A}$ be defined as follows:

$$R_t^{\mathcal{A}} = a_t^{\mathcal{A}}(1 - \omega^{\mathcal{A}})\chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) - \omega^{\mathcal{A}} \mid q - (a_t^{\mathcal{A}} + a_t^{\mathcal{B}}) \mid . \tag{10}$$

The first term in (10) is a *pure economic profit* of player $\mathcal{A}$, cf. (8). The second term expresses *efficiency of using the game potential* at round t, i.e. whenever $a_t^{\mathcal{A}} + a_t^{\mathcal{B}} < q$ some amount remains unclaimed and thus lost for the players. The same situation happens when an agreement is not reached and the entire amount q is lost.

Obviously reward (10) ensures that, given fixed $a_t^{\mathcal{A}}$, player $\mathcal{A}$ will receive the maximum possible reward iff its opponent, $\mathcal{B}$, demands $q - a_t^{\mathcal{A}}$. The proposed form of reward, (10), "connect" $\mathcal{A}$'s action with that of $\mathcal{B}$ and thus encourages player $\mathcal{A}$ to *indirectly negotiate* with $\mathcal{B}$ during bargaining. The mechanism of dynamic indirect negotiation is as follows. Each player influences the amount left while their opponent observes this influence and changes their next demand. Let us assume that there is a tendency for some unclaimed amount to remain. Then, if one player has consumed a small portion of it, the other player will observe that and then may increase their demand in the next round. Another situation occurs when the joint claim of the players exceeds the available resources. Then any of the players may step back and reduce their demand in the next round. This behaviour can again lead to a large unclaimed amount and affects the future demands of the players. In particular, the desire to minimise the unclaimed amount, $\mid q - (a_t^{\mathcal{A}} + a_t^{\mathcal{B}}) \mid$, (10), forces player $\mathcal{A}$ to modify the current demand while taking into account the history of the opponent's claims. By doing so, in each round, each player dynamically *adapts* their demand to the foreseen demands of their opponent, that is indirectly negotiates with the opponent.

Weight $\omega^{\mathcal{A}} \in [0, 1]$ in (10) reflects $\mathcal{A}'s$ preferences between pure economic gain and exploiting the game's potential. The value $\omega^{\mathcal{A}} = 0$ implies player $\mathcal{A}$ considers pure economic profit only, while in case of $\omega^{\mathcal{A}} = 1$ player $\mathcal{A}$ cares about efficient use of the game potential. The $\mathcal{A}'s$ reward (10) thus equals

$$R_t^{\mathcal{A}} = \begin{cases} a_t^{\mathcal{A}} - \omega^{\mathcal{A}} \left(q - a_t^{\mathcal{B}}\right) & \text{if } a_t^{\mathcal{A}} + a_t^{\mathcal{B}} \leq q, \\ -\omega^{\mathcal{A}} \left(q - a_t^{\mathcal{B}}\right) & \text{otherwise.} \end{cases} \tag{11}$$

Definition 2 and considerations above describe DM of player $\mathcal{A}$. Easy to see that the same considerations can be applied to formalise decision making of player $\mathcal{B}$.

The conditional independence of the players' actions given by the game rules and the definition of the state, see Definition 2, imply

$$p(s_{t+1}|s_t, a_t^{\mathcal{A}}) = p(a_t^{\mathcal{A}}|a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}})p(a_t^{\mathcal{B}}|a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}}). \tag{12}$$

From player $\mathcal{A}$ point of view, the first factor in (12) is a part of $\mathcal{A}$'s optimal policy while the second factor models DM of player $\mathcal{B}$ and can be recursively estimated using Bayesian paradigm [54] as described in Section II-D.

B. HEURISTIC MODEL OF OPPONENT

The proposed approach formalised and solved bilateral dynamic bargaining of learning self-interested player within MDP framework (Section II-C). To verify the approach we propose a probabilistic bargaining model for non-learning and non-optimising opponent. The model is based on the reported experimental evidence obtained with human-players, see [28], [56]. For simplicity here we consider player $\mathcal{B}$ is serving as an opponent to $\mathcal{A}$.

Heuristic behaviour of $\mathcal{B}$ reflects the dependence of its future demand on the results of the previous round. Once the previous round demands are incompatible, that is $a_{t-1}^{\mathcal{A}} + a_{t-1}^{\mathcal{B}} > q$, player $\mathcal{B}$ tends to decrease next demand. If there are unclaimed money left in the previous round, $\mathcal{B}$, on the contrary, increases the next demand. The proportion (speed) of demands' increase/decrease may depend on personal traits (i.e. reflect the personality of $\mathcal{B}$).

The remainder of this section introduces model that reflects the behaviour of an opposing player, $\mathcal{B}$.

[5] Upper indexes indicate the player whom action or profit belongs to.

1) $\mathcal{B}$ HAD LOW DEMAND IN THE PREVIOUS ROUND

Consider the previous demand of player $\mathcal{B}$ is low, i.e. less than the fair split would have been, $a_{t-1}^{\mathcal{B}} \leq \frac{q}{2}$. The next demand (in sense of its mean value) then depends on the success of the previous round, i.e. whether demands in the previous round were compatible or not. Below we distinguish these two cases and provide the respective probabilistic description of $\mathcal{B}$'s actions.

i) **Incompatible Demands** ($a_{t-1}^{\mathcal{A}} + a_{t-1}^{\mathcal{B}} > q$): $\mathcal{B}$ tends to keep its next demand close to the previous one, $a_t^{\mathcal{B}}$, as the previous demand of $\mathcal{A}$ was certainly much higher than $a_{t-1}^{\mathcal{B}}$. Thus any further increase could cause players' demands to become incompatible again and implies zero profit. Therefore the new demand of player $\mathcal{B}$ can be modelled as follows:

$$p(a_t^{\mathcal{B}}|a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}}) \propto \exp\left(-\frac{(a_t^{\mathcal{B}} - a_{t-1}^{\mathcal{B}})^2}{2\sigma^2}\right) \tag{13}$$

while $a_{t-1}^{\mathcal{B}} \leq \frac{q}{2}$.

ii) **Compatible Demands** ($a_{t-1}^{\mathcal{A}} + a_{t-1}^{\mathcal{B}} \leq q$): opponent $\mathcal{B}$ will proportionally increase the next demand, expecting $\mathcal{A}$ to do the same in order to fully distribute the entire available amount, q. In other words player who received less in the previous round would also ask for proportionally less unclaimed money and vice versa. A model of $\mathcal{B}$ describing the new demand is then

$$p(a_t^{\mathcal{B}}|a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}}) \propto \exp\left(-\frac{K}{2\sigma^2}\right), \tag{14}$$

with $K = \left(a_t^{\mathcal{B}} - a_{t-1}^{\mathcal{B}} - \frac{a_{t-1}^{\mathcal{B}}}{a_{t-1}^{\mathcal{A}} + a_{t-1}^{\mathcal{B}}}(q - a_{t-1}^{\mathcal{A}} - a_{t-1}^{\mathcal{B}})\right)^2$ while $a_{t-1}^{\mathcal{A}} + a_{t-1}^{\mathcal{B}} \leq q$ and $a_{t-1}^{\mathcal{B}} \leq \frac{q}{2}$.

2) $\mathcal{B}$ HAD HIGH DEMAND IN THE PREVIOUS ROUND

Now let us consider a situation when the previous demand of $\mathcal{B}$ was high, i.e. its value was greater than the fair split would have been, $a_{t-1}^{\mathcal{B}} > \frac{q}{2}$. Then $\mathcal{B}$ decreases/increases demand while keeping own share proportional to the previous round in order to fully distribute the entire amount. A player who received less in the last round would ask for less of proportionally less unclaimed money and vice versa. Then a model of $\mathcal{B}'s$ new demand has the same form as (14).

C. PRIOR MODELS USED IN LEARNING

Our approach considers decision making of the player in question, $\mathcal{A}$, who models behaviour of the opponent, $\mathcal{B}$, and optimises own demand in order to maximise the accumulated profit. The ability to accurately predict the opponent's behaviour significantly affects the success of $\mathcal{A}'s$ decision making, (12). To learn a model of the opponent, $\mathcal{A}$ follows the approach described in Section II-D. It exploits knowledge available in the form of a parameter prior that quantifies $\mathcal{A}'s$ belief about dynamics of the opponent, $\mathcal{B}$. Following Bayesian paradigm this prior will be gradually updated with new data accumulated, see Section II-D and [55]. The choice of prior model is important, especially when a number of game rounds is limited. In implementation we use three prior models reflecting different knowledge $\mathcal{A}$ about $\mathcal{B}$:

- *a uniform prior distribution*. This model is used when $\mathcal{A}$ has little or no knowledge about the dynamics of $\mathcal{B}$
- *prior model describing "rational" heuristic*, see Section III-B. It is used when non-optimising $\mathcal{B}$ follows some heuristic and does not optimise. In that case prior model has the same structure as (13) or (14), but with different (larger) standard deviation σ.
- *pre-trained prior model.* The third way of building an a priori model mimics the natural learning process of human players, where the player first gathers some knowledge about the opponent's playing style and then updates this knowledge during the game. Practically it means we run game for 30 preliminary rounds and player $\mathcal{A}$ built prior model of $\mathcal{B}$ based on the data obtained during these rounds. This way of building prior is used whenever the both players optimise and learn.

IV. SIMULATED EXPERIMENTS

The proposed approach is illustrated with the Nash demand game, described in Section III-A, using simulated examples.[6]

We selected the most representative experiments from a much wider set of the experiments differing in the number of rounds and horizons. The selected experiments are long enough to perform learning (because very short runs will not be sufficient to learn the models used), while longer runs will add no significant information about the results.

A. GOAL OF THE EXPERIMENTS

The goal was to analyse the impact of the proposed distributed solution and indirect negotiation and to verify that player employing the proposed DM policy is capable of achieving better results than heuristic player playing the same role. The main objectives of the performed experiments are:

- illustrate the distributed DM approach in repetitive bargaining;
- show that the proposed form of the reward function leads to an indirect negotiation and to a coordinated course of actions of both players, that is, to a more efficient allocation of the available limited resources;
- demonstrate influence of weight ω in (10)
- show that DM policy with indirect negotiation brings higher profit to every player compare to the heuristic model.

B. COMMON SETTINGS OF THE EXPERIMENTS

Each game has 60 rounds and optimisation horizon $h \in \mathbb{N}$ equals 10 game rounds. The amount of money that players can split (if they reach an agreement) is $q = 10$ CZK per round. The reward (10) is evaluated for the optimal policy (3)

[6] The examples were implemented in MATLAB, The MathWorks, Inc.

resulted from the dynamic programming [48]. The initial state of each player $s_1 = (a_0^{\mathcal{A}}, a_0^{\mathcal{B}})$ is preset to $a_0^{\mathcal{A}} = a_0^{\mathcal{B}} = 3$.

The simulation is performed for 11 different values of weight ω, (10). Weight $(0 \leq \omega \leq 1)$ expresses a trade-off between the individual profitability and efficiency of using the game resources. It thus reflects the extent to which the player is negotiating. Zero value of ω in (10) models the situation when the player is interested only in economic profit. Other values of ω $(0 < \omega \leq 1)$ correspond to cases when the player maximises the personal profit while minimising the unclaimed amount of money.

C. EXPERIMENTS PERFORMED

The players used in the simulation are artificial agents with either heuristic DM model (see Section III-B) or proposed DM policy that optimises reward (10), see Section II-C. In each game at least one of the players uses the observed behaviour to update the opponent's model, see Section II-D. In order to display behaviour of our bargaining model, five typical cases were considered:

Test 1 : Both players are non-learning. The player in question, $\mathcal{A}$, is of the MDP type and uses the proposed DM policy optimising (10). Its opponent, $\mathcal{B}$, behaves heuristically, see Section III-B.

Test 2 : This case is similar to Test 1 but player $\mathcal{A}$ dynamically learns the opponent's model.

Test 3 : Both players are of the MDP type and non-learning. They have no knowledge of their opponent and do not model it either (i.e. they use uniform model).

Test 4 : Both players are of the MDP type, and use having non-informative prior for learning, see Section II-D.

Test 5 : This case is similar to Test 4, but the players use informative priors (i.e. opponent model trained during the preliminary phase).

D. APPROACH VERIFICATION

The players have played the game repeatedly with different settings. The results are summarised in graphs depicting *individual cumulative profits* of the players, total profit of the game, and success rates of game depending on the value of parameter ω. The *success rate* is defined as a number of game rounds in which the players' demands were compatible and thus satisfied. In other words, the value of the success rate shows how successfully the players collaborated, i.e. respected the opponent's actions. High values indicate high collaboration. The results show minimum, mean and maximum values of the individual cumulative profits and the game success rate. Note that

- The maximum success rate does not necessarily imply the maximum total profit of the game.
- Compatibility of the players' claims does not guarantee zero unclaimed amount in the game.
- It is not guaranteed either that the maximum profit will be obtained for the same value of the weight ω. Thus the total maximum (minimum) profit of the game is *not* equal to the sum of the individual maximum (minimum) profit of the players.

TABLE 1. Test 1: Player $\mathcal{A}$ optimises but not learn. Player $\mathcal{B}$ follows the heuristic model (13), (14).

	Cumulative profit			Success rate
	$\mathcal{A}$	$\mathcal{B}$	Total	%
Min. value	84.00	69.00	153.00	28.00
Mean value	130.55	141.55	272.09	51.52
Max. value	218.00	269.00	487.00	95.00

1) TEST 1: $\mathcal{A}$ IS A NON-LEARNING MDP PLAYER, $\mathcal{B}$ BEHAVES HEURISTICALLY

Player $\mathcal{B}$, behaves according to the heuristic model (13), (14) with $\sigma^2 = 1$.

Player $\mathcal{A}$ is of MDP type and uses DM policy (3) that optimises reward (10). In optimisation $\mathcal{A}$ uses model $p(s_{t+1}|s_t, a_t)$ having structure of the heuristic model, see Section III-B, but with different parameter $\sigma = 3$. This imitates a situation when $\mathcal{A}$ has partial or vague knowledge of the opponent.

Cumulative profits of the players $\mathcal{A}$ and $\mathcal{B}$ are shown in Figure 1 and Figure 2. Total cumulative profit and success rate of the game as a function of parameter $\omega^{\mathcal{A}}$ are shown in Figure 3 and Figure 4.

The players are successful in more than 51% of the rounds on average. The results show influence of parameter $\omega^{\mathcal{A}}$ on profit: the higher the parameter, the higher the profits of individual players and the higher the total profit of the game. This indicates a positive effect of the second term (10), which prompts $\mathcal{A}$ to *indirectly negotiate* with $\mathcal{B}$ by minimising the unclaimed amount in each round. As a result the players start to implicitly cooperate.

The results show the saddle value of parameter $\omega^{\mathcal{A}} = 0.5$ that provides the minimum values of $\mathcal{A}$'s profit and success rate of the game. The maximum is reached for $\omega^{\mathcal{A}} = 1$. Obviously, optimising player $\mathcal{A}$ earned slightly less on average than non-optimising player $\mathcal{B}$. It could be because player $\mathcal{B}$ used fixed decision making rules and $\mathcal{A}$ had to adapt to that.

2) TEST 2: $\mathcal{A}$ OPTIMISES AND LEARNS, $\mathcal{B}$ BEHAVES HEURISTICALLY

This experiment is similar to Test 1, see Section IV-D1, i.e. player $\mathcal{B}$ behaves accordingly to the heuristic model, Section III-B, and player $\mathcal{A}$ uses optimal DM policy minimising the proposed reward, (10). Unlike Test 1, player $\mathcal{A}$ is learning. $\mathcal{A}$ considers a uniform prior as $\mathcal{B}$'s transition model and dynamically updates it via data gathered, see Section III-C.

Cumulative profit and success rate obtained in Test 2 are shown in Figures 5-8 and Table 2. Obviously the learning has a positive impact on the game results. On average, the players are successful in more then 66% of all rounds - the average success rate is about 15% higher than in Test 1, as is the cumulative profit. The minimum values for individual profits

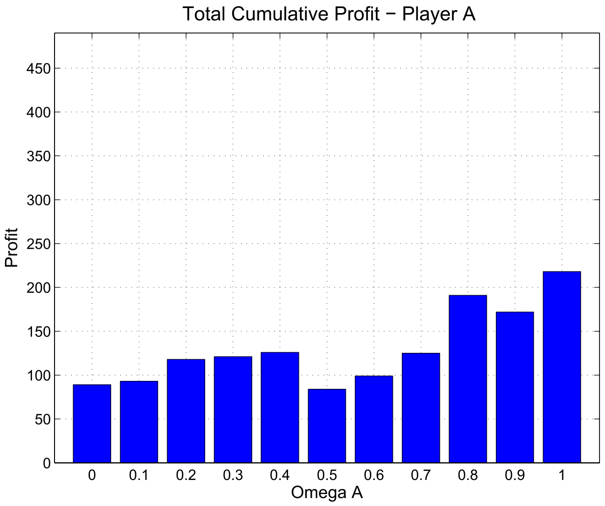

FIGURE 1. Test 1 - $\mathcal{A}$'s cumulative profit on weight $\omega^{\mathcal{A}}$.

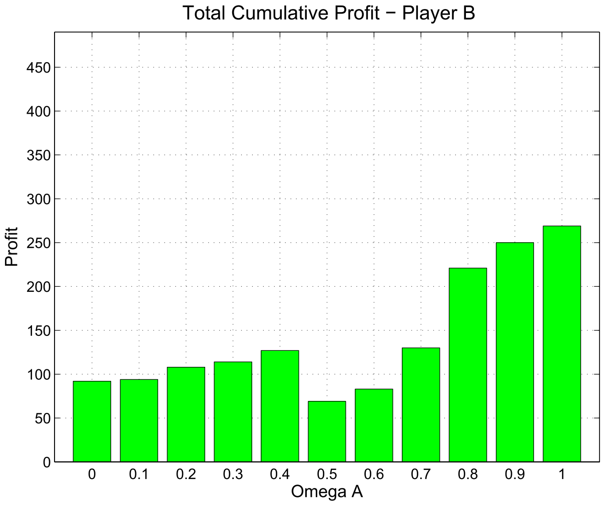

FIGURE 2. Test 1 - $\mathcal{B}$'s cumulative profit on weight $\omega^{\mathcal{A}}$.

TABLE 2. Test 2: Player $\mathcal{A}$ optimises and learns, player $\mathcal{B}$ follows the heuristic model (13), (14).

	Cumulative profit			Success rate
	$\mathcal{A}$	$\mathcal{B}$	Total	%
Min. value	170.00	168.00	338.00	62.00
Mean value	180.91	184.55	365.45	66.36
Max. value	190.00	216.00	406.00	75.00

and overall success rate are significantly higher cf. Table 1. On the other hand, their maximum values have noticeably decreased. The players have similar individual profits and their values weakly depend on parameter $\omega^{\mathcal{A}}$.

3) TEST 3: BOTH PLAYERS OPTIMISE BUT NONE LEARNS

This experiment considers both players are of MDP type and select DM policy maximising reward (10). However neither of the players is learning. They use a fixed uniform model (see Section III-C) that models the situation when there is no information about the opponent.

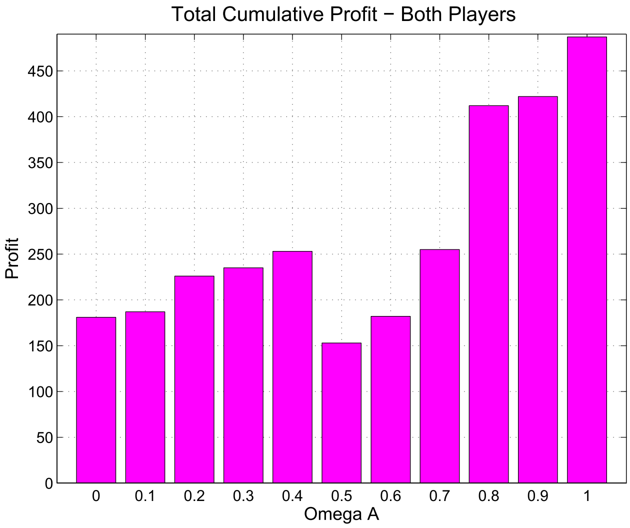

FIGURE 3. Test 1 - Total Profit of Players.

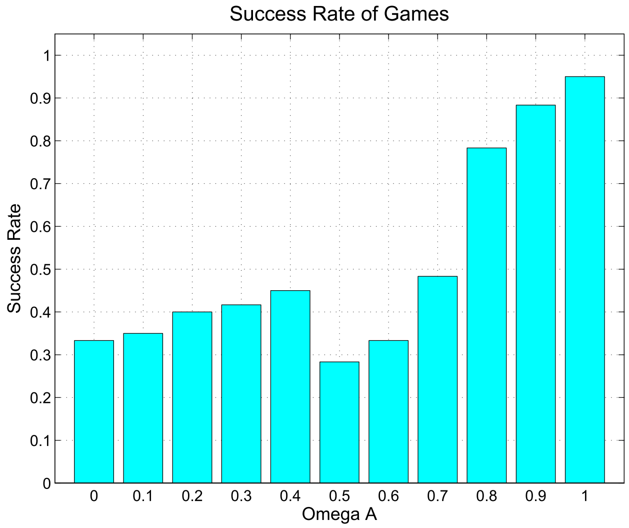

FIGURE 4. Test 1 - success rate of the games.

Cumulative profits and success rate of the game vs. parameters $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$ are shown in Figures 9-12 and Table 3.

The results illustrate positive impact of i) optimal bargaining compare to heuristic behaviour, cf. results of Test 1 and Test 2 and ii) proposed reward (10) that prompts on indirect negotiation. Even with non-informative prior knowledge, the players get higher profit. If the players' weights are $\omega^{\mathcal{A}} \geq 0.5$ and $\omega^{\mathcal{B}} \geq 0.5$ the success rate is 100% and overall game profit gained is close to the maximum possible (600 CZK), see Table 3. By other words: when the players care about the optimal allocation of the resources (by assigning high weights to the second term in reward (10)), the bargaining is more profitable. On average, the players are successful in more than 76% of all rounds.

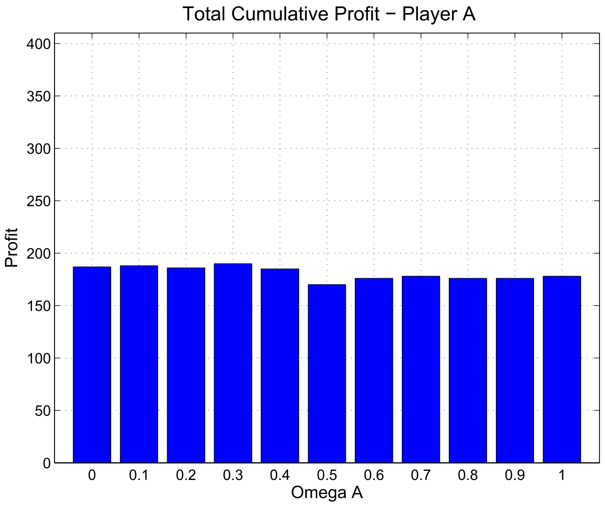

FIGURE 5. Test 2 - $\mathcal{A}$'s cumulative profit on weight $\omega^{\mathcal{A}}$.

TABLE 3. Test 3: Both players optimise but none learns.

	Cumulative profit			Success rate
	$\mathcal{A}$	$\mathcal{B}$	Total	%
Min. value	179.00	179.00	504.00	67.00
Mean value	219.12	219.12	438.23	76.58
Max. value	298.00	298.00	596.00	100.00

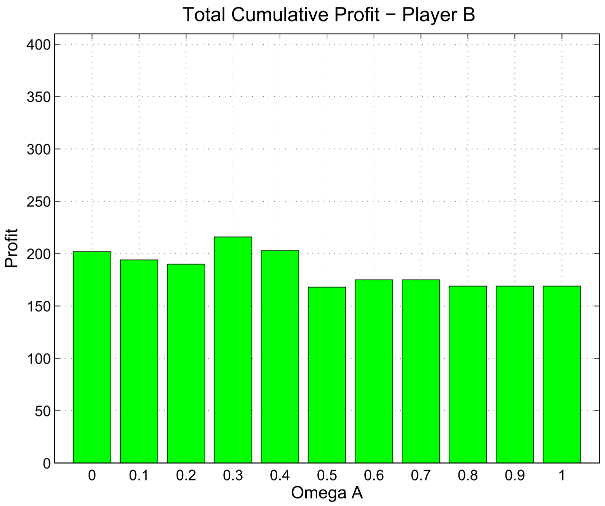

FIGURE 6. Test 2 - $\mathcal{B}$'s cumulative profit on weight $\omega^{\mathcal{A}}$.

4) TEST 4: PLAYERS OPTIMISE AND LEARN WITH UNIFORM PRIOR

This experiment is similar to Test 3, i.e. both players are of MDP type and maximise reward (10). Unlike Test 3, the ability to learn the opponent's model has been added to the players. The agents dynamically enhance their non-informative (uniform) priors based on the data observed during the game. Thus each player i) learns their opponent; ii) searches for optimal demand; iii) indirectly negotiates via minimising unshared resources.

TABLE 4. Test 4: Both players optimise and learn with uniform prior.

	Cumulative profit			Success rate
	$\mathcal{A}$	$\mathcal{B}$	Total	%
Min. value	152.00	152.00	304.00	52.00
Mean value	267.38	267.38	534.76	92.18
Max. value	310.00	310.00	596.00	100.00

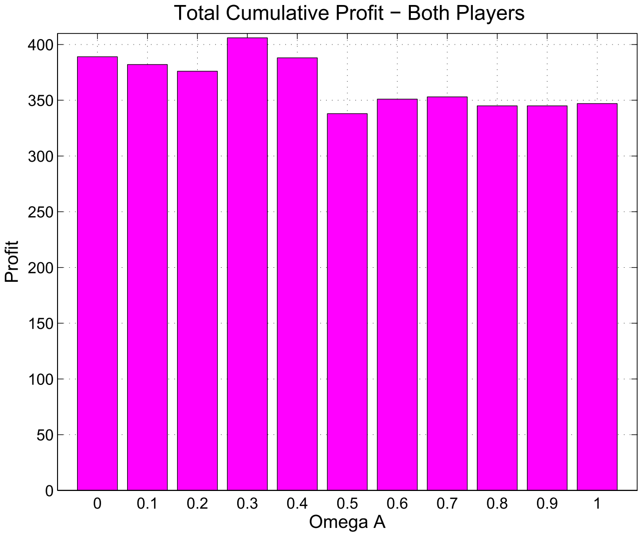

FIGURE 7. Test 2 - Overall profit of the players.

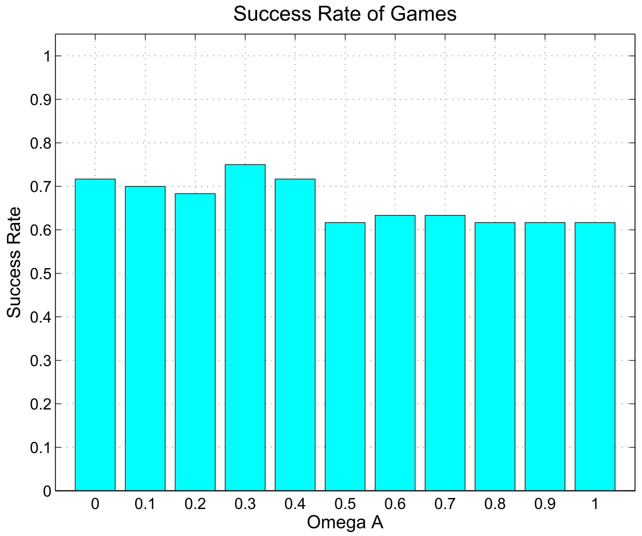

FIGURE 8. Test 2 - Success rate of the games.

Cumulative profits and success rate of the game in dependence on parameters $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$ are shown in Figures 13-16 and Table 5.

The results show significant improvement due to the learning. The minimum values of the individual profits and the success rate decreased but their maximum values increased on average, see Table 4. The significant improvement occurred when $\omega^{\mathcal{A}} \leq 0.5$ and $\omega^{\mathcal{B}} \leq 0.5$, see Figures 13-15. Compare to Test 3, learning ability brought higher individual profits as

TABLE 5. Test 5: Both players have informative prior, learn and optimise.

	Cumulative profit			Success rate
	$\mathcal{A}$	$\mathcal{B}$	Total	%
Min. value	152.00	152.00	304.00	52.00
Mean value	282.40	282.64	565.27	97.48
Max. value	328.00	328.00	596.00	100.00

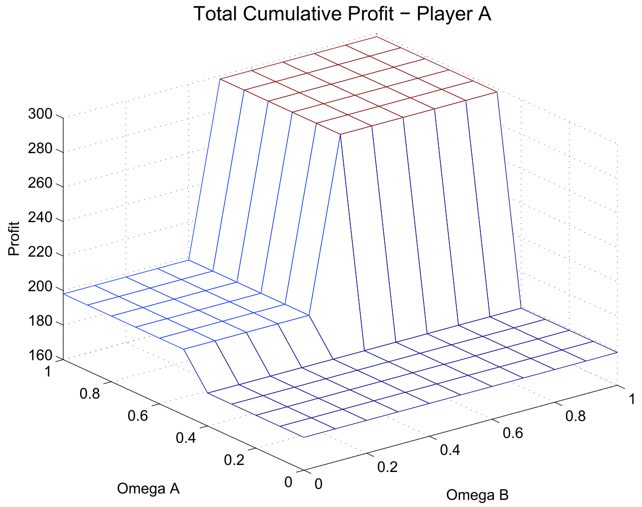

FIGURE 9. Test 3 - $\mathcal{A}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.

well as higher success rates achieved for relatively low values of ω. Thus learning helps even when the player's willingness to negotiate (expressed by ω) is low.

5) TEST 5: INFORMATIVE PRIOR INFORMATION

This experiment is a modification of Test 4 with each of the players having meaningful prior knowledge of the opponent. First, to get informative prior, the players played 30 training rounds during which they gained prior models of their opponents. Then, Test 3 has been performed with the resulting prior instead of uniform distribution.

Cumulative profits and success rate of the game in dependence on parameters $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$ are shown in Figures 17-20 and Table 5.

The results show further improvement, see Table 5, cf. Tests 3-4. The minimum values of profits and success rate do not change but the maximum and mean values noticeably increased, cf. Test 4 (Section IV-D4). The players achieve much higher individual profits for low values of weights ω because they coordinated their demands to make them almost always compatible (see Figure 20).

V. DISCUSSION

Section IV describes simulation results obtained on the NDG. It can be seen that our DM model can help the players to effectively bargain and counteract the incomplete knowledge. The main advantages of the proposed DM model are as follows:

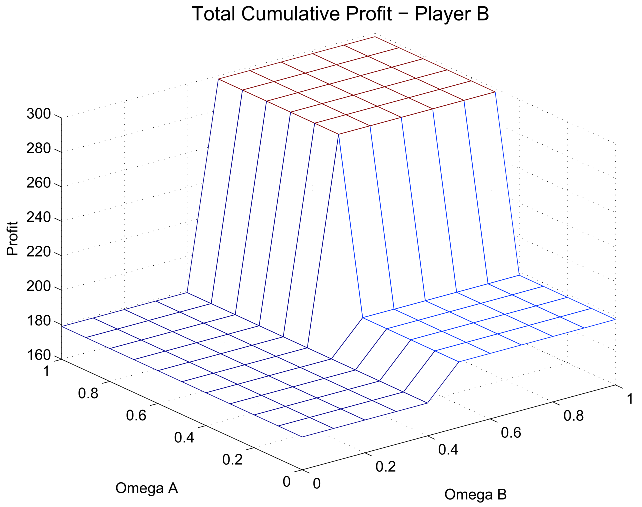

FIGURE 10. Test 3 -$\mathcal{B}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.

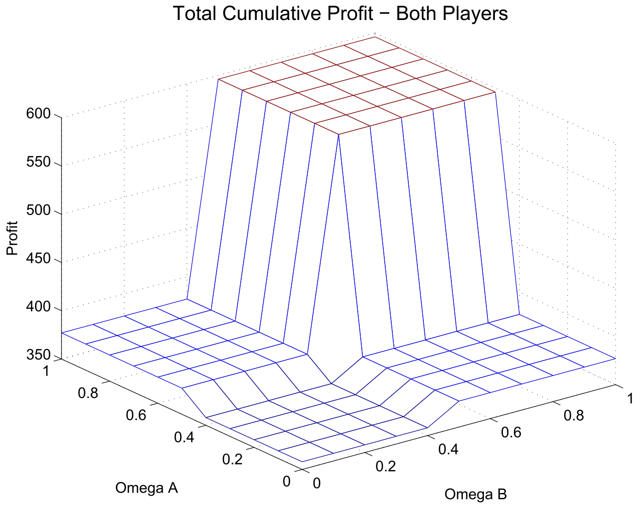

FIGURE 11. Test 3 - Overall profit of the players.

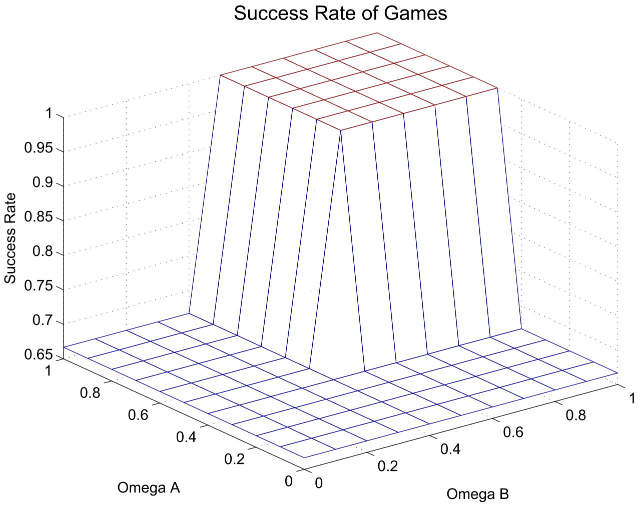

FIGURE 12. Test 3 - Success rate of the games.

- The proposed reward function respects individual economic profit of the bargaining agent and the unclaimed

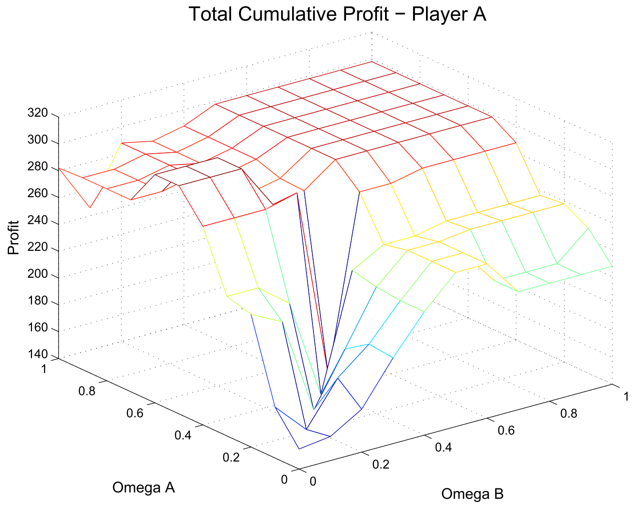

FIGURE 13. **Test 4 - $\mathcal{A}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.**

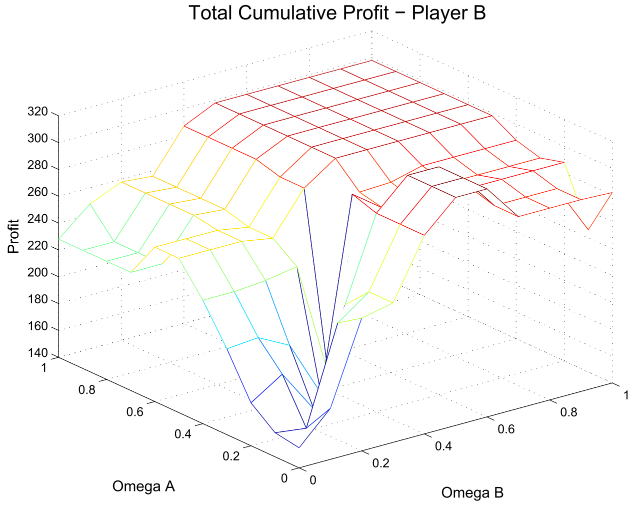

FIGURE 14. **Test 4 - $\mathcal{B}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.**

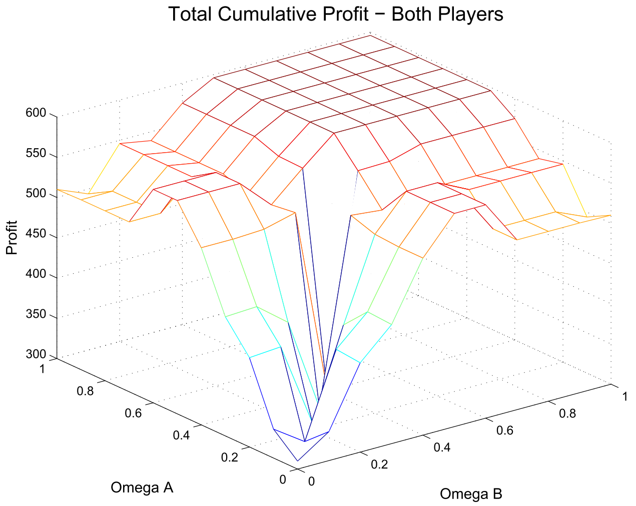

FIGURE 15. **Test 4 - Overall profit of the players.**

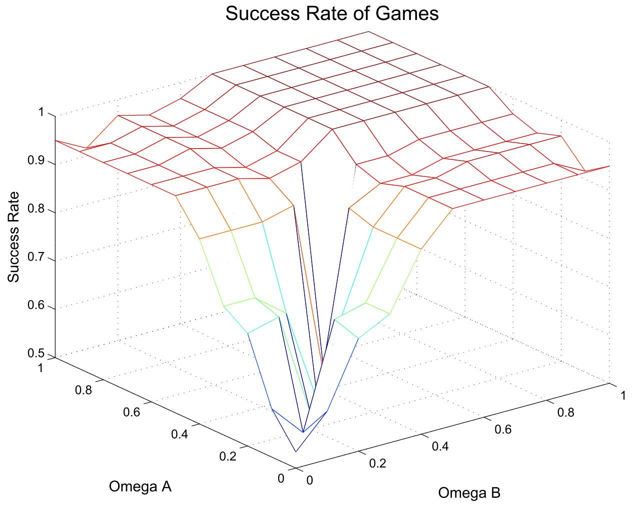

FIGURE 16. **Test 4 - Success rate of the games.**

amount of money from the previous round. As the opponent's past actions enter the reward (10), the optimal policy of the agent implicitly respects them. And vice versa: the optimal policy of the opponent respects agent's actions. Hence both players are forced to implicitly cooperate.

- The weight ω in (10) expresses trade-off between the individual profitability and efficiency of using game potential. At the same time it also reflects agent's preferences and partially style of playing (personal traits). High values of the weight in the player's reward (10) indicate a high interest of the agent in efficient use of game resources, i.e. in minimising the remaining unclaimed amount. In each round thus the reward *encourages* the agent to dynamically "adapt" its current demand to the predicted demand of the opponent. In the next round, the resulting profit[7] together with the updated opponent's model, is used in (2), (3) to select a new demand. This is the essence of the proposed *indirect dynamic negotiation*.
- Compared to the heuristic bargaining model, Section III-B, our optimal DM policy increased the mean value of the player's individual profit by more than 50% (in the case of an uninformative prior) and by about 65% (informative prior).
- Learning significantly improves the bargaining results. However optimising but not learning agent can have worse individual results compare with the heuristic opponent. The reason is that the optimising agent implicitly cooperates with the opponent during bargaining but does not use the correct opponent model for this.[8] On contrary, the opponent does not cooperate and it uses a fixed heuristic model. As a result, the agent's effort brings more profit to the opponent than to itself.

[7] which reflects the effect of the previous round and thus provides a feedback to the agent.

[8] And therefore cannot predict the opponent.

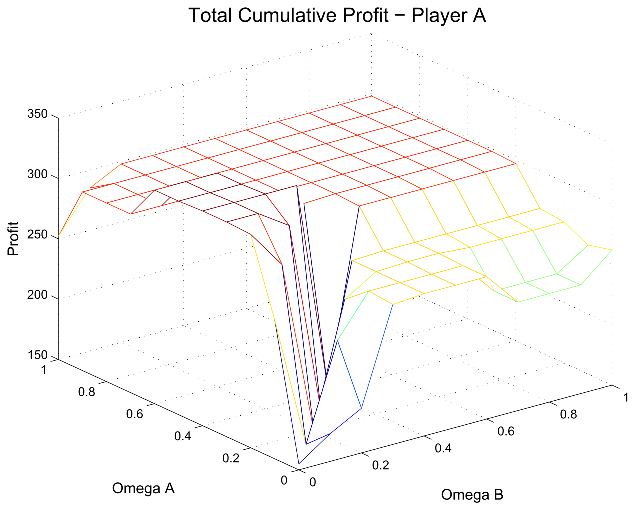

FIGURE 17. Test 5 - $\mathcal{A}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.

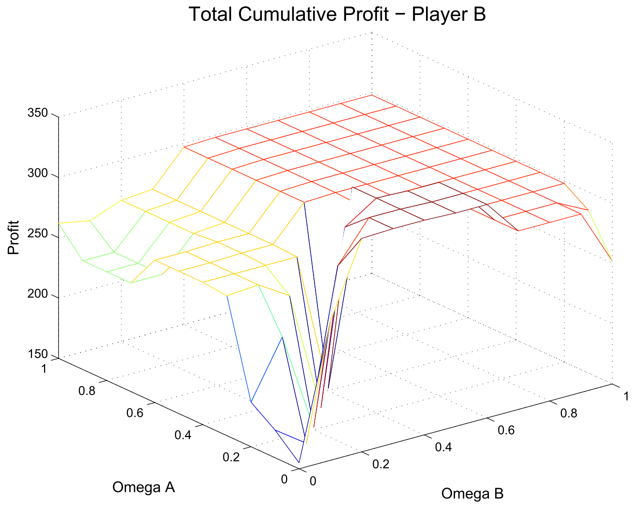

FIGURE 18. Test 5 - $\mathcal{B}$'s cumulative profit in dependence on weights $\omega^{\mathcal{A}}$ and $\omega^{\mathcal{B}}$.

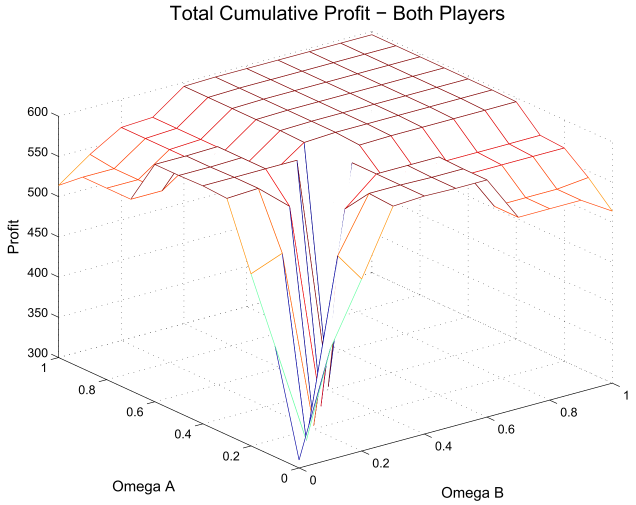

FIGURE 19. Test 5 - Overall profit of the players.

- The best bargaining results were achieved if both players are learning and employ the proposed bargaining policy. Informative prior used in learning can significantly improve the agent's profit.

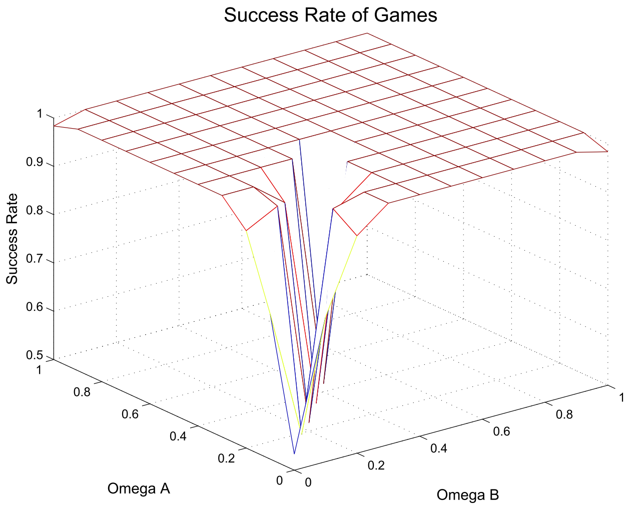

FIGURE 20. Test 5 - Success rate of the games.

The proposed solution can be further extended i) to cover multi-issue bargaining; ii) to respect human non-rationality given by social and cognitive aspects; iii) to respect emotional state of the agent that has been proved to significantly influence DM [57].

VI. CONCLUDING REMARKS

The paper addresses a problem of sequential bilateral bargaining with incomplete information. We proposed DM model that helps agents to successfully bargain by performing indirect negotiation and learning the opponent's model. Methodologically the paper casts heuristically-motivated bargaining of a self-interested independent agent into a framework of Bayesian learning and Markov decision processes. The proof of the main results is based on the standard methodology. However, the problem formulation and the gained solution are novel and practically important. The special form of the reward *implicitly* motivates the players to negotiate indirectly, via closed-loop interaction. At the same time the proposed method is privacy-preserving, since it does not require the exchange of data or models between the bargaining agents. We illustrate the approach by applying our model to the Nash demand game, which is an abstract model of bargaining. The paper provides our original formulation and solution of the practically important DM scenario. It presents the initial study that confirms that our formulation is meaningful and gives the promising results. The results indicate that the introduced DM model: i) leads to coordinating the players' actions and to their indirect negotiation; ii) results in maximising success rate of the game and iii) brings more individual profit to the players compare to the heuristic model.

The proposed bargaining policy minimises losses caused by: (i) insufficient use of the resources; (ii) demands that exceed the total resources available; and (iii) incomplete knowledge.

The results obtained indicate possibility to create a realistic and applicable methodology of cooperation and negotiation in *flatly* organised networks of interacting agents without a fixed structure, cf. [58]. We believe that our approach is suitable for non-cooperative, multi-agent networks, since we provide an easy way to implicit cooperation. The solution does not rely on a central authority and the proposed DM model outperforms a heuristic model whenever both agents are rational, learning and follow the optimal strategy.

In future work we would like:

- to cover the multi-issue bargaining;
- to extend the approach to a multi-agent settings;
- to implement the approach for other bargaining rules than Nash demand game.

Further foreseen challenge is learning weights of individual players based on their bargaining history. The weights indirectly reflect agent's model of bargaining and preferences. Moreover the weights may depend on the agent's personality [59], which allows taking into account the influence of personality traits on decision making.

ACKNOWLEDGMENT

The authors would like to thank Miroslav Kárný for his thorough feedback on an earlier draft and for useful discussions.

REFERENCES

[1] Y.-Y. Yang and X.-M. Xie, "Research on the effectiveness of network negotiation based on evolutionary game model," *IEEE Access*, vol. 8, pp. 194623–194630, 2020.

[2] E. David, E. Gerding, D. Sarne, and O. Shehory, *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets: AAMAS Workshop, AMEC 2009, Budapest, Hungary, May 12, 2009, and IJCAI Workshop, TADA 2009* (Lecture Notes in Bus. Information Processing). Pasadena, CA, USA: Springer, 2010.

[3] L. Li, K. Robert Lai, and S. Zhu, "Data-driven behavior-based negotiation model for cyber-physical-social systems," *IEEE Access*, vol. 7, pp. 83319–83331, 2019.

[4] S. Chakraborty, T. Baarslag, and M. Kaisers, "Automated peer-to-peer negotiation for energy contract settlements in residential cooperatives," *Appl. Energy*, vol. 259, Feb. 2020, Art. no. 114173.

[5] B. López, B. Innocenti, and D. Busquets, "A multiagent system for coordinating ambulances for emergency medical services," *IEEE Intelligent Systems*, vol. 23, no. 5, pp. 50–57, 2008.

[6] T. Ito, H. Hattori, M. Zhang, and T. Matsuo, *Rational, Robust, and Secure Negotiations in Multi-Agent Systems* (Studies in Computational Intelligence). Berlin, Germany: Springer, 2008.

[7] K. Kang, B. Q. Tan, and R. Y. Zhong, "Multi-attribute negotiation mechanism for manufacturing service allocation in smart manufacturing," *Adv. Eng. Informat.*, vol. 51, Jan. 2022, Art. no. 101523.

[8] P. Ryan, J. Fokker, S. Healy, and A. Amann, "Dynamics of targeted ransomware negotiation," *IEEE Access*, vol. 10, pp. 32836–32844, 2022.

[9] F. Ren and M. Zhang, "A single issue negotiation model for agents bargaining in dynamic electronic markets," *Decis. Support Syst.*, vol. 60, pp. 55–67, Apr. 2014.

[10] J. Kennan and R. Wilson, "Strategic bargaining models and interpretation of strike data," *J. Appl. Econometrics*, vol. 4, no. S1, pp. 87–130, 1989.

[11] A. Rubinstein, "Perfect equilibrium in a bargaining model," *Econometrica*, vol. 50, no. 1, pp. 97–109, Jan. 1982.

[12] J. F. Nash, Jr., "The bargaining problem," *Econometrica*, vol. 18, no. 2, pp. 155–162, 1950.

[13] H. Raiffa, *The Art and Science of Negotiation*. Cambridge, U.K.: Cambridge Univ. Press, 1982.

[14] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA, USA: MIT Press, 1994.

[15] T. Ellingsen, "The evolution of bargaining behavior," *Quart. J. Econ.*, vol. 112, no. 2, pp. 581–602, May 1997.

[16] S. G. Ficici and A. Pfeffer, "Modeling how humans reason about others with partial information," in *Proc. 7th Int. Joint Conf. Auton. Agents Multiagent Syst. (AAMAS)*, 2008, pp. 315–322.

[17] R. Lin, S. Kraus, J. Wilkenfeld, and J. Barry, "Negotiating with bounded rational agents in environments with incomplete information using an automated agent," *Artif. Intell.*, vol. 172, nos. 6–7, pp. 823–851, Apr. 2008.

[18] M. A. Neale and M. H. Bazerman, "Negotiator cognition and rationality: A behavioral decision theory perspective," *Organizational Behav. Hum. Decis. Processes*, vol. 51, no. 2, pp. 157–175, Mar. 1992.

[19] M. van't Wout, L. J. Chang, and A. G. Sanfey, "The influence of emotion regulation on social interactive decision-making," *Emotion*, vol. 10, no. 6, pp. 815–821, 2010.

[20] G. F. Loewenstein, L. Thompson, and M. H. Bazerman, "Social utility and decision making in interpersonal contexts," *J. Personality Social Psychol.*, vol. 57, no. 3, pp. 426–441, Sep. 1989.

[21] T. Guy, M. Kárný, A. Lintas, and A. Villa, "Theoretical models of decision-making in the ultimatum game: Fairness vs. reason," in *Advances in Cognitive Neurodynamics (V)*, R. P. X. Wang, Ed. Cham, Switzerland: Springer, 2016.

[22] R. Kulhavý and M. Kárný, "Tracking of slowly varying parameters by directional forgetting," *IFAC Proc. Volumes*, vol. 17, no. 2, pp. 687–692, Jul. 1984.

[23] B. An, N. Gatti, and V. Lesser, "Bilateral bargaining with one-sided two-type uncertainty," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, 2009, pp. 403–410.

[24] G. Haim, Y. Gal, B. An, and S. Kraus, "Human–computer negotiation in a three player market setting," *Artif. Intell.*, vol. 246, pp. 34–52, May 2017.

[25] R. Lin, Y. Oshrat, and S. Kraus, *Automated Agents That Proficiently Negotiate With People: Can We Keep People out of the Evaluation Loop*. Berlin, Germany: Springer, 2012, pp. 57–80.

[26] A. Quinn, M. Kárný, and T. V. Guy, "Optimal design of priors constrained by external predictors," *Int. J. Approx. Reasoning*, vol. 84, pp. 150–158, May 2017.

[27] J. Homolová, E. Zugarová, M. Kárný, and T. V. Guy, "On decentralized implicit negotiation in modified ultimatum game," in *Multi-Agent Systems and Agreement Technologies* (Lecture Notes in Computer Science), vol. 10767, F. Belardinelli and E. Argente, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-01713-2_25.

[28] D. J. Cooper and E. G. Dutcher, "The dynamics of responder behavior in ultimatum games: A meta-study," *Experim. Econ.*, vol. 14, no. 4, pp. 519–546, Nov. 2011.

[29] M. Kárný and T. V. Guy, *On Support of Imperfect Bayesian Participants*. Berlin, Germany: Springer, 2012, pp. 29–56.

[30] M. S. Fagundes, S. Ossowski, M. Luck, and S. Miles, "Using normative Markov decision processes for evaluating electronic contracts," *AI Commun.*, vol. 25, no. 1, pp. 1–17, 2012.

[31] P. Faratin, C. Sierra, and N. R. Jennings, "Using similarity criteria to make negotiation trade-offs," in *Proc. 4th Int. Conf. MultiAgent Syst.*, Jul. 2000, pp. 119–126.

[32] D. Zeng and K. Sycara, "Bayesian learning in negotiation," *Int. J. Hum.-Comput. Stud.*, vol. 48, no. 1, pp. 125–141, Jan. 1998.

[33] L. Wu, S. Chen, X. Gao, Y. Zheng, and J. Hao, "Detecting and learning against unknown opponents for automated negotiations," in *PRICAI 2021: Trends in Artificial Intelligence*, D. Pham, T. Theeramunkong, G. Governatori, and F. Liu, Eds. Cham, Switzerland: Springer, 2021, pp. 17–31.

[34] T. Baarslag, M. J. Hendrikx, K. V. Hindriks, and C. M. Jonker, "Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques," *Auto. Agents Multi-Agent Syst.*, vol. 30, pp. 849–898, Sep. 2015.

[35] U. Kiruthika, T. S. Somasundaram, and S. K. S. Raja, "Lifecycle model of a negotiation agent: A survey of automated negotiation techniques," *Group Decis. Negotiation*, vol. 29, no. 6, pp. 1239–1262, Dec. 2020.

[36] T. Baarslag, M. Kaisers, E. H. Gerding, C. M. Jonker, and J. Gratch, *Self-Sufficient, Self-Directed, and Interdependent Negotiation Systems: A Roadmap Toward Autonomous Negotiation Agents*. Cham, Switzerland: Springer, 2022, pp. 387–406.

[37] S. Fatima, M. Wooldridge, and N. Jennings, "Optimal negotiation of multiple issues in incomplete information settings," in *Proc. 3rd Int. Joint Conf. Auton. Agents Multiagent Syst., (AAMAS)*, vol. 3, N. Jennings, C. Sierra, L. Sonenberg, and M. Tambe, Eds. 2004, pp. 1080–1087.

[38] D. E. Kröhling, O. J. A. Chiotti, and E. C. Martínez, "A context-aware approach to automated negotiation using reinforcement learning," *Adv. Eng. Informat.*, vol. 47, Jan. 2021, Art. no. 101229.

[39] R. Lin, Y. Gev, and S. Kraus, "Bridging the gap: Face-to-face negotiations with an automated mediator," *IEEE Intell. Syst.*, vol. 26, no. 6, pp. 40–47, Nov. 2011.

[40] B. An, K. M. Sim, C. Y. Miao, and Z. Q. Shen, "Decision making of negotiation agents using Markov chains," *Multiagent Grid Syst.*, vol. 4, no. 1, pp. 5–23, May 2008.

[41] S. S. Fatima, M. Wooldridge, and N. R. Jennings, "Approximate and online multi-issue negotiation," in *Proc. 6th Int. Joint Conf. Auto. Agents Multiagent Syst. (AAMAS)*, 2007, pp. 947–954.

[42] F. Ren and M. Zhang, "A single issue negotiation model for agents bargaining in dynamic electronic markets," *Decis. Support Syst.*, vol. 60, pp. 55–67, Apr. 2014.

[43] A. Stern, S. Kraus, and D. Sarne, "A negotiating strategy for a hybrid goal function in multilateral negotiation," 2022, *arXiv:2201.04126*1.

[44] Z. Feng, C. Tan, J. Zhang, and Q. Zeng, "Bargaining game with altruistic and spiteful preferences," *Group Decis. Negotiation*, vol. 30, no. 2, pp. 277–300, Apr. 2021.

[45] K. Hindriks and D. Tykhonov, "Opponent modelling in automated multi-issue negotiation using Bayesian learning," in *Proc. 7th Int. Joint Conf. Auto. Agents Multiagent Syst.*, 2008, pp. 331–338.

[46] B. An, V. Lesser, and K. Sim, "Strategic agents for multi-resource negotiation," *Auto. Agents Multi-Agent Syst.*, vol. 23, pp. 114–153, Jul. 2010.

[47] M. L. Puterman, *Markov Decission Processes*. Hoboken, NJ, USA: Wiley, 1994.

[48] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[49] J. Boger, J. Hoey, P. Poupart, C. Boutilier, G. Fernie, and A. Mihailidis, "A planning system based on Markov decision processes to guide people with dementia through activities of daily living," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 323–333, Apr. 2006.

[50] M. Feng and Y. Li, "Predictive maintenance decision making based on reinforcement learning in multistage production systems," *IEEE Access*, vol. 10, pp. 18910–18921, 2022.

[51] H. Song, C. C. Liu, J. Lawarree, and R. W. Dahlgren, "Optimal electricity supply bidding by Markov decision process," *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 618–624, May 2000.

[52] W. T. Scherer, S. Adams, and P. A. Beling, "On the practical art of state definitions for Markov decision process construction," *IEEE Access*, vol. 6, pp. 21115–21128, 2018.

[53] W. B. Powell, *Approximate Dynamic Programming*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.

[54] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System Identification*, P. Eykhoff, Ed. Oxford, U.K.: Pergamon Press, 1981, pp. 239–304.

[55] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Cham, Switzerland: Springer, 2006.

[56] J. Henrich, R. Boyd, S. Bowles, C. Camerer, and E. Fehr, *The Handbook of Experimental Economics*. Princeton, NJ, USA: Princeton Univ. Press, 2001.

[57] I. Kapoutsis, R. Volkema, and A. Lampaki, "Mind the first step: The intrapersonal effects of affect on the decision to initiate negotiations under bargaining power asymmetry," *Frontiers Psychol.*, vol. 8, p. 1313, Aug. 2017.

[58] T. Wu, X. Liu, J. Qin, and F. Herrera, "Trust-consensus multiplex networks by combining trust social network analysis and consensus evolution methods in group decision-making," *IEEE Trans. Fuzzy Syst.*, early access, Mar. 10, 2022, doi: 10.1109/TFUZZ.2022.3158432.

[59] M. Fiori, A. Lintas, S. Mesrobian, and A. E. P. Villa, *Effect of Emotion and Personality on Deviation from Purely Rational Decision-Making*. Berlin, Germany: Springer, 2013, pp. 129–161.

TATIANA V. GUY (Senior Member, IEEE) received the Dipl.-Eng. degree in control and automation from the Kiev Polytechnic Institute, and the Ph.D. degree in cybernetics from Czech Technical University, Prague. Since 2013, she has been the Head of the Adaptive Systems Department, Institute of Information Theory and Automation, Prague. She has an appointment as a Researcher at the Czech University of Life Sciences. Her current research interests include distributed decision making, nature-inspired cooperation, and transfer learning.

JITKA HOMOLOVÁ received the B.Sc. degree in financial mathematics from Charles University, Prague, Czech Republic, and the M.Sc. and Ph.D. degrees in informatics from Czech Technical University, Prague. Since 2004, she has been with the Institute of Information Theory and Automation, Prague. Her main research interests include Bayesian identification, Markov decision processes, bargaining, and multi-agent systems.

ALEKSEJ GAJ is currently pursuing the M.Sc. degree in applied mathematics with the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague. Since 2018, he has been involved in the research work at the Adaptive Systems Department, Institute of Information Theory and Automation, Prague. His research interests include cooperation in multi-agent systems, and use of quantum mechanics for decision making.