Learning and Reasoning in Cognitive Radio Networks

Liljana Gavrilovska, Vladimir Atanasovski, Irene Macaluso, and Luiz DaSilva

***Abstract*—Cognitive radio networks challenge the traditional wireless networking paradigm by introducing concepts firmly stemmed into the Artificial Intelligence (AI) field, i.e., learning and reasoning. This fosters optimal resource usage and management allowing a plethora of potential applications such as secondary spectrum access, cognitive wireless backbones, cognitive machine-to-machine etc. The majority of overview works in the field of cognitive radio networks deal with the notions of observation and adaptations, which are not a distinguished cognitive radio networking aspect. Therefore, this paper provides insight into the mechanisms for obtaining and inferring knowledge that clearly set apart the cognitive radio networks from other wireless solutions.**

***Index Terms*—Knowledge, Learning, Reasoning, Game theory, Reinforcement learning, Policy based reasoning.**

I. Introduction

The core idea of cognitive radios is based on the cognitive cycle, according to which radios must be able to observe their operating environment, then decide how to best adapt to it, and act accordingly. As the cycle repeats, the radio should be able to learn from its past actions. The principle rests on the radio's ability to observe, adapt, reason, and learn.

A little over ten years since cognitive radios have been first proposed by Mitola [1], the research literature on cognitive radios is vast. It has, however, tended to focus on the first two aspects (observation, adaptation) and less so on the last two (reasoning, learning). The 'observe' portion of the cognitive cycle is exemplified by work on sensing for opportunistic access to spectrum. The 'adapt' portion can manifest itself when the radio, based on its sensed operating environment, performs channel selection, power and topology control, adaptive modulation and coding, or some combination thereof.

In this paper, we focus on the 'reason' and 'learn' aspects of cognition. Those aspects lend themselves to multi-disciplinary analysis, taking advantage of advances made in game theory, artificial intelligence, multi-objective reasoning, and policy systems, among others.

We start with a brief review of the framework according to which cognitive radios are understood. We then discuss two alternate mathematical views of learning mechanisms. Game theory offers us the mathematical tools to model interactions among autonomous players (in the context of this paper, typically cognitive radios seeking to maximize some objective function). Through dynamic game models, we can study how the radio's actions are affected by past experiences: in a way, how the radio 'learns' from its past actions and those of others.

Dynamic games have been applied to problems that model the interaction among secondary users competing for opportunistic access to the spectrum, as well as those that model interactions between primary and secondary users. Such models range from repeated games [2], [3], [4] to stochastic games [5], [6], [7] and evolutionary games [8], [9]. We briefly define each of those classes of games and discuss some example applications.

We then look at reinforcement learning, both by a single agent and by multiple agents, and how advances in that field can be applied to cognitive radio and dynamic spectrum access problems. Reinforcement learning has been applied to a variety of problems in the context of the cognitive radio literature, including dynamic channel selection [10], transmission power adaptation for spectrum management [11], cooperative sensing in ad hoc networks [12], and multicarrier aggregation [13].

The next portion of this survey concerns itself with reasoning, a fundamental aspect of every 'intelligent' entity, regardless of being biological or artificial. Reasoning may be broadly classified into being *instinctive* or *cognitive* [14]. Instinctive reasoning is driven by emotions and is therefore an inherent characteristic of biological entities (including humans). Cognitive reasoning requires the power of cognition, i.e. the complex interaction of knowledge (past and present), learning and the associated inference mechanisms, stripping the emotions from the entire process. This leads to an increased reasoning time, but also to an improved reasoning result (i.e. a more meaningful and 'intelligent' solution). This paper refers to the notion of cognitive reasoning in its application in wireless networking.

The primary responsibility of the cognitive reasoning is the choice of a set of actions that lead to efficient decision-making. Therefore, the cognitive reasoning is often viewed as a decision process using historical as well as current knowledge of the environmental context. Additionally, the process learning must be powerful enough to enrich the knowledge base, to foster increased efficiency of the subsequent reasoning. As a result, there is a tight coupling among knowledge, learning, and reasoning in the cognitive sense.

Cognitive reasoning may be investigated at three levels of abstraction: *conceptual*, *formal* and *realizational* [15]. The conceptual abstraction requires models capable of capturing the specific and possible nuances within the reasoning entity. An example is a cognitive agent [15], which can be a living entity, a group of living entities, or a technical system. The

L. Gavrilovska and V. Atanasovski are with the Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University in Skopje, Macedonia, e-mail: (liljana;vladimir@feit.ukim.edu.mk).

I. Macaluso and L. DaSilva are with CTVR, Trinity College Dublin, Ireland (e-mail:irene.macaluso@gmail.com, dasilval@tcd.ie). L. DaSilva is also with Virginia Tech, USA.

formal theory requires frameworks and logic to interpret the interactions among the elements involved in the cognitive reasoning. The formalism is crucial for handling the various plausible reasoning methods. It also ensures that the reasoning itself is self-contained and independent from the actual enabling technology [16]. Finally, the realizational theory should encompass the envisioned application and the environmental practical limitations of an operating cognitive engine. For instance, in the context of cognitive radios, [17] introduces a cognitive wrapper, envisioned as a realizable cognitive entity with scalable intelligence and designer-specified learning and reasoning algorithm capabilities, while [18] discusses the aspect of reasoning robustness, which is extremely important for practical realizations.

There are numerous applications of cognitive reasoning in the telecommunications domain. For instance, [19] and [20] and elaborate on the usage of reasoning for network monitoring and management. These applications address the problem of scalability and showcase the potential of cognitive reasoning to handle various network incidents timely and efficiently. Reference [21] introduces a reasoning framework for enabling smart homes, with reasoning as an intelligent interpreter of data coming from various electronic devices in homes. Moreover, reasoning can provide an unambiguous interface for the consumers to track and, possibly, intervene in the home environment, allowing for increased intelligence and energy-awareness. Of particular relevance to our paper, the main application of cognitive reasoning to wireless communications has been in the area of efficient and flexible spectrum management [22].

Cognitive reasoning is a focal aspect of cognitive radio networks. It fosters the development and/or the extraction of contextual and environmental awareness towards an optimal solution to a particular problem. A reasoning output would then be a timely and intelligent answer to a problem set based on previous actions and consequences, current observations and objectives, and the descriptions of the used data-types [17], [23]. However, the reliability of the reasoning output strongly depends on the accurate estimation of the environmental context, which needs to be carefully analyzed in different cognitive networking applications [24].

This paper enumerates and discusses some possible frameworks for reasoning in cognitive radio networks, from Bayesian networks to case-based reasoning. We give special attention to policy-based reasoning, as it is particularly applicable to cognitive radios operating in new and dynamically changing spectrum regimes. In the conclusions, we offer our views on some of the open research areas in reasoning and learning for cognitive networks.

In each section, we combine some fundamental discussion of the principles of learning and reasoning with some examples of how they can be applied to cognitive radios and dynamic spectrum access.

II. Cognitive Radio Framework

Prior to the introduction of the cognitive radio networking paradigm, learning and reasoning mechanisms were not customarily built into wireless network architectures. Rather, the observations and the adaptations were governed and fostered by hard-coded rules inside the terminals' firmware. The introduction of cognitive radio networks allowed for the incorporation of learning and reasoning mechanisms as a distinct characteristic of the cognitive cycle within.

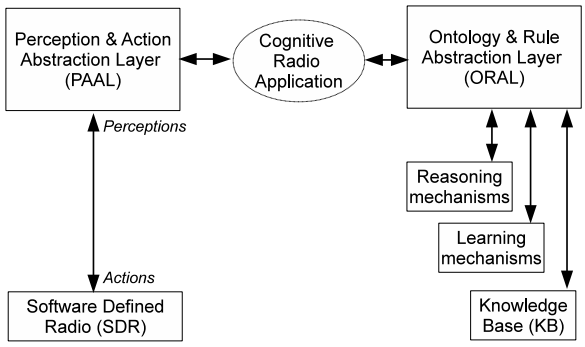

Fig. 1. General Cognitive Radio Architecture.

Learning mechanisms are responsible for building up *knowledge* and *knowledge bases*. However, the knowledge by itself would be useless in cognitive radio networks unless there is a form of *inference* that determines how various pieces of knowledge can be translated into actionable decisions. The inference is enabled by *reasoning mechanisms*, resulting in a tight coupling between learning, knowledge and reasoning.

The aspects of learning and reasoning, knowledge and knowledge bases, as well as observations and adaptations, are intertwined in a general cognitive radio architecture, abstracted in Fig. 1. The Cognitive Radio Application requires a Software Defined Radio (SDR) in order to fulfill its functionalities. These functionalities may include spectrum mobility, spectrum handovers, adaptations based on perceived past and predicted future environmental changes, etc. However, a crucial cornerstone of the general cognitive radio architecture is the *need for platform independence* of the knowledge and the application itself. Therefore, the Perception and Action Abstraction Layer (PAAL) is introduced as a mediator that allows translation of the SDR observables and the actions into a *platform-independent knowledge representation*. This allows independence of the actual cognitive radio application from the plethora of market-available SDR devices that use different software wrappers.

On the other hand, the acquired knowledge also necessitates independence from the potential application. As a result, the Ontology and Rule Abstraction Layer (ORAL) is foreseen as a presenter of knowledge (i.e. ontologies and rules) in a *platform-independent implementation manner*. Finally, the Knowledge Base (KB) stores the acquired knowledge and the set of actions that were or are to be executed.

The general cognitive radio architecture from Fig. 1 can be instantiated into various specific architectures [25] that incorporate different numbers of loops in the cognitive cycle, different duration of the learning process, adaptations to different conditions and parameters, etc. Every instantiation

may also include some specific functionalities (e.g. genetic algorithms), but they all adhere to the common principles elaborated in the Introduction.

The next section details the most prominent learning mechanisms within modern cognitive radio networking.

III. Decision Making and Learning Mechanisms

We take a broad view of learning to study the adaptations performed by a network of cognitive radios. Our discussion encompasses both the application of machine learning techniques to cognitive networks and game-theoretic analysis of simple adaptation mechanisms that can be shown to converge to an equilibrium (such as a Nash equilibrium or one of its variations in cooperative and non-cooperative game theory). Recent work [26] investigates the intersection between machine learning and game theory.

A. Game-theoretic analysis

Game theoretic models account for multi-agent decision making, including cases where each player decides on her actions based on observing the history of actions selected by other players in previous rounds of the game. This allows us to model a learning process by each player and whether this learning ultimately leads to a stable state for all. Games that model competition and cooperation that evolve with time are called *dynamic games*. The parallel to cognitive networks should be clear: in many game theoretic models of cognitive networks, the players in the game are the cognitive radios that form the network. These radios take actions such as setting their transmit power or selecting a channel in which to operate. Such actions are based on the radio's observations of its environment (e.g., channel availability, frame error rate, or interference). As time progresses, a radio can learn from the outcome of its past actions and from observing the actions of other radios in the network, and modify its actions accordingly.

1) Repeated games: The simplest game theoretic model that captures these concepts is that of a repeated game. A repeated game is one in which each stage of the game is repeated, usually with an infinite time horizon. Let $\mathbf{N}$ denote the set of radios in a network, and the vector $\mathbf{a^{(k)}}$ denote the N-dimensional vector of actions taken by the players in the k^{th} stage of the game. In each stage k, a player's strategy seeks to maximize her utility function, while taking into account the history of actions collected in the vector $\mathbf{h^{(k)}} = (\mathbf{a^{(1)}}, \mathbf{a^{(2)}}, \ldots, \mathbf{a^{(k)}})$. In other words, a player's strategy can be expressed as a mapping from histories to actions: $a_i^{(k)} = f_i(\mathbf{h^{(k-1)}})$. The expected utility is typically discounted by a factor $0 < \delta < 1$, meaning that a payoff in future stages of the game is worth less than the same payoff in the current stage.

The cognitive radio process is often described by the OODA loop (observe-orient-decide-act). In Fig. 2, we map the four steps in this reasoning process to the formalism of a repeated game.

A simple example may be in order at this point. Consider a number of cognitive radios competing for channels that are available when the primary licensee for the frequency band is not active. Mapping this problem into the repeated games formulation above, the radios are the players in the game, their action is the selection of one of C channels, and these selections may depend on the history of primary user activity, as well as on the pattern of channel utilization by other secondary users (for example, a channel that has a history of being frequently occupied by the primary may be avoided by all secondary users).

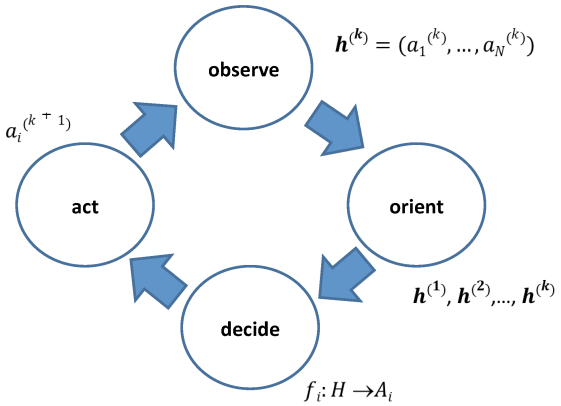

Fig. 2. Starting at the top of the diagram, cognitive radio i observes other radios' actions at the k^{th} stage: these actions are collected into a history vector. The history vectors for all previous stages are considered during the orientation step. The radio then decides on an action by applying a strategy that maps from the set of histories to the radio's action set. Finally, the radio performs an action during the $(k+1)^{th}$ stage of the repeated game, and the cycle repeats.

The well-known concept of Nash equilibrium is readily applied to repeated games: In a Nash equilibrium strategy profile, no player can unilaterally increase her expected payoff by selecting a different strategy.

In the study of economic incentives for cognitive radios and networks, [27] considers an oligopoly spectrum market, with license holders competing to provide services to secondary users. This is modeled as a repeated game: with associated incentives and punishment for deviating, the authors show that it is possible to sustain a Nash equilibrium that maximizes the providers' profit.

Channel selection in opportunistic spectrum access has also often been modeled as a repeated game. Wu et al. [2] model the sharing of open spectrum as a repeated game; they consider a punishment scheme and show that a more efficient equilibrium can be reached when autonomous radios interact repeatedly, as opposed to when they interact in a single stage game (in general, a well known result in game theory, an example of which is the repeated Prisoner's Dilemma). They go further and also consider incentives for cognitive radios to truthfully report their operating conditions in negotiating access to spectrum: relying on mechanism design, the authors of [2] design cheat-proof strategies for dynamic spectrum sharing. The selection of the best spectrum opportunities by secondary users of some spectrum band is modeled as a repeated game in [3]. In that model, secondary users will have to vacate their

current channel whenever a primary user becomes active, and the authors consider a cost associated with switching channels. A subgame perfect equilibrium [28], a Nash Equilibrium that is also an equilibrium for every proper subgame of the original game, is one way to characterize the likely outcome of such a game.

In [4], the authors use repeated games to model the evolution of reputation among secondary users, when one of them is chosen to manage the spectrum made available by the primary user. In several of the applications above, repeated interactions among a set of cognitive radios allow for the design of incentive mechanisms that lead to a more efficient equilibrium. A different question is whether there are simple ways for radios, by observing others' actions and the utility resulting from its own actions, to converge to a Nash equilibrium. We treat that question next.

2) Potential games: The class of games called *potential games* is of particular interest in the context of learning. If a dynamic adaptation problem can be modeled as a potential game, then if radios follow a simple adaptation algorithm (which we will discuss in more detail shortly) they are guaranteed to reach a solution that is stable from the point of view of the entire network.

To introduce potential games, let us start with the concept of a *potential function*. A potential function V maps from the action set of all players, $\mathbf{A} = \mathbf{A_1} \times ... \times \mathbf{A_N}$, into the real numbers: $V\colon \mathbf{A} \rightarrow \mathbb{R}$. A unilateral change in action by one player has the same effect on that player's utility $u_i(\mathbf{a})$ as it has on the potential function. Formally, for all players $i \in N$ and all $a_i, b_i \in \mathbf{A_i}$:

$$V(a_i, \mathbf{a_{-i}}) - V(b_i, \mathbf{a_{-i}}) = u_i(a_i, \mathbf{a_{-i}}) - u_i(a_i, \mathbf{a_{-i}}).$$

(Here, we adopt standard notation in game theory, with $\mathbf{a_{-i}} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$ representing the vector of all players' actions, except player i.) A game for which a potential function can be found is called an *exact potential game*.

A weaker concept of potential function is that of an *ordinal potential function*. That function also maps the action set of all players to the real numbers, but with the following property:

$$V(a_i, \mathbf{a_{-i}}) - V(b_i, \mathbf{a_{-i}}) > 0 \iff u(a_i, \mathbf{a_{-i}}) - u(a_i, \mathbf{a_{-i}}) > 0.$$

A game for which such a function can be found is called an *ordinal potential game*. It is easy to see that every potential game is an ordinal potential game, but also that the converse is not true. But how are potential games relevant to our discussion of learning in networks of cognitive radios? Because potential games have desirable properties in terms of the existence of a Nash equilibrium and the convergence to that equilibrium through simple adaptations. For instance, all finite potential games have at least one Nash equilibrium in pure strategies (a finite game is a game where the player and action sets are finite). More generally, if the strategy space for the game is compact and the potential function continuous, then the game has at least one pure strategy Nash equilibrium. Just as importantly, from the point of view of learning, is that the players are guaranteed to reach these equilibria through best reply and better reply dynamics.

To introduce better and best reply dynamics, let us consider a repeated game where at each stage exactly one player is offered the opportunity to take action. A player is said to follow a best reply strategy if her selected action maximizes her utility, given the other players' current actions. With a *better reply strategy*, the player will always select an action that provides an improvement in utility with respect to her previous action, again given the other players' current actions.

Some of the seminal work in applying potential games to cognitive radio problems was done by Neel [29]. A number of problems in multi-channel communications can be modeled as potential games. For example, when the utility function of each radio considers the social welfare of the network (e.g., by attributing a cost to the radio from interference caused to others, as well as interference suffered from others) a potential function naturally emerges. This is the case in the work on channel selection by [30].

Even when players have utility functions that reflect their own selfish interests, rather than those of the network, in a number of cases of interest to dynamic spectrum access and cognitive network games the model results in a potential game.

Thomas et al. [31], for example, model the topology control problem for an ad hoc network where nodes can select a channel to operate on from a finite set of available channels. This topology control mechanism consists of two phases: in the first phase, radios select a transmit power level with energy efficiency and network connectivity in mind; in the second, they select channels, with interference minimization objectives. The authors are able to show that both problems (power control and channel selection) can be formulated as ordinal potential games, and best-response dynamics are guaranteed to converge to an equilibrium.

3) Other dynamic games: More general formulations of dynamic games have also been applied to cognitive radio and dynamic spectrum access problems.

The dynamic game model in [32], for example, is used to model uncertainty about observed strategies adopted by other players. A primary and multiple secondary users (SUs) interact, with the former setting prices for access to the spectrum and the latter selecting how large a portion of spectrum to use. Since each secondary user only interacts with the primary, it cannot get a complete picture of the strategies and payoffs of other secondary users. Each SU therefore gradually adapts its selection of how much of the spectrum to occupy based on the marginal benefit it can observe from this selection. A learning rate parameter adjusts the speed with which adaptations can be made. This parameter will impact the stability region of the learning algorithm, as well as its sensitivity to the selection of the initial strategy.

A particular class of dynamic games that has found recent applications to the dynamic spectrum access problem is that of stochastic games. In stochastic games, the environment changes in response to the actions of all players. This is captured by the introduction of a state space and a stochastic process that models the game's transitions among states. Each player's stage payoff depends on the current state of the game as well as on all players' actions.

The work in [5] considers a set of radios performing

distributed and opportunistic access to the channel, wherein in each time slot one radio can be scheduled per spectrum hole (and, when the primary user is present, none of the secondaries is allowed to transmit). The authors model this problem as a switching control game, a type of stochastic game where the state space can be partitioned into disjoint subsets such that, whenever the game is in state S_i, the transition probabilities depend only on player i's actions. The decisions of each of the radios can be then described by a finite sequence of Markov decision processes.

Both centralized and distributed stochastic games are formulated in [6], where radios compete for spectrum opportunities with and without help from a central spectrum moderator, respectively. In [7], the same authors model bidding policies for secondary users competing for spectrum controlled by a spectrum broker, again using the formalism of stochastic games and considering that each secondary user can only observe a partial history of previous usage of spectrum.

Another variation of stochastic games gives rise to evolutionary game theory. This is inspired by evolutionary biology and the idea that an organism's genes largely determine its fitness to the environment in which they exist. The more fit the organism, the higher the likelihood that it will produce offspring, increasing the representation of its genes in the overall population. The process of mutation is also modeled through random changes to the players' strategies over time.

An evolutionary game is proposed in [8] to study behavioral dynamics in cooperative spectrum sensing, where each sensing agent (possibly belonging to different providers) must decide whether to contribute to the overall picture of spectrum availability. Reference [9] applies evolutionary games to the problem of network selection by radios facing the choice of multiple wireless access technologies.

After having briefly summarized some of the game theoretic models used to analyze multi-agent decision making and the process of arriving at stable outcomes (critical for cognitive radios operating in a network), we turn our attention to the application of reinforcement learning to cognitive radios.

B. Reinforcement-learning techniques

Reinforcement learning (RL) plays a key role in the literature on multi-agent learning. In fact the nature of the task itself, i.e. learning a mapping between situations and actions while interacting with other agents, makes the use of supervised learning techniques quite difficult. In a dynamic and non-stationary environment it would be challenging, sometimes even impossible, to provide the agents with the correct actions associated with the current situation. The RL paradigm is more versatile in the multi-agent domain, as it allows the agents to autonomously discover the situation-action mapping through a mechanism of trial and error. An RL agent learns by exploring the available actions and refining its behavior using only an evaluative feedback, referred to as the *reward*. In other words, in the RL paradigm an agent learns by interacting with its environment, which in the multi-agent domain also includes other agents. The learning mechanism is driven by the rewards. Generally an agent is expected not just to take into account the immediate reward, but also to evaluate the consequences of its actions on the future in order to maximize its long-term performance. Delayed reward and trial-and-error constitute the two most significant features of RL.

1) Single-agent RL: Multi-agent reinforcement learning (MARL) evolved from the single-agent RL setting. In the single agent case, RL is usually performed in the context of Markov decision processes (MDP). In a typical RL scenario the agent represents its perception at time k as a state $\mathbf{x_k} \in \mathbf{X}$, where $\mathbf{X}$ is the finite set of environment states. The agent interacts with the environment by performing actions. Each action $a_k \in \mathbf{A}$, where $\mathbf{A}$ is the finite set of actions of the agent, could trigger a transition to a new state. The agent will receive a reward as a result of the transition, according to the reward function $\rho\colon \mathbf{X} \times \mathbf{A} \times \mathbf{X} \rightarrow \mathbb{R}$. The agent's task is to devise a policy, i.e. a sequence of (state, action) pairs, to maximize the expected discounted reward. In the context of MDP, it has been proved that an optimal deterministic and stationary policy exists [33]. The problem of learning the optimal policy for the single-agent RL scenario has been addressed both in the case where the state transition and reward functions are known (model-based learning) and in the case where they are not (model-free learning). Most MARL algorithms are based on Q-learning [34], a model-free algorithm that estimates an optimal action-value function. An action-value function, named Q-function, is the expected return of a state-action pair for a given policy. The optimal action-value function, Q^*, corresponds to the maximum expected return for a state-action pair. Once it estimated Q^*, the agent can select the optimal actions by using a greedy policy, i.e. the policy that for every state the agent selects the action with the corresponding highest Q-value. The updating rule of the Q-function is:

$$\begin{aligned} Q_{k+1}(\mathbf{x}_k, a_k) \quad = \quad & (1-\alpha_k) Q_k(\mathbf{x}_k, a_k) + \\ & \alpha_k \left[r_{k+1} + \gamma \max_{a'} Q_{k+1}(\mathbf{x}_{k+1}, a') \right] \end{aligned}$$

where γ is the discount factor, $\alpha_k \in [0,1]$ is the learning factor, and $r_{k+1} = \rho(\mathbf{x}_k, a_k, \mathbf{x}_{k+1})$. As it can be noted, the updating rule of Q-learning does not require knowledge about the reward or the transition functions: only the observed reward is used to update the Q-values. In a stationary environment the learned Q-function converges to Q^* if all the state-action pairs are visited an infinite number of times and under the stochastic approximation conditions on the sequence of the learning factors α_k [34].

Incidentally, there are clear connections between MDPs and game theoretic models, in particular stochastic games. A stochastic game is a dynamic game for which state transitions are probabilistic, allowing us to model uncertainty in the players' operating environment. While an MDP models a single agent's decisions, in a stochastic game there are multiple agents, and their actions, the next state, and rewards depend on the vector of all players' actions ([35] offers a good treatment of stochastic games and their relationship to MDP). Fig. 3 provides one way to position reinforcement learning and game theoretic models with respect to the number of agents considered and to the cardinality of the state space.

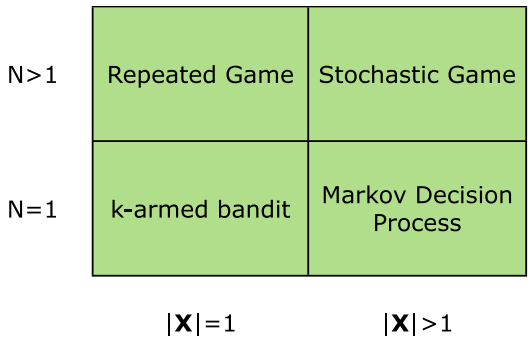

Fig. 3. The relationship between repeated games, stochastic games, MDPs, and multi-armed bandit problems is illustrated by this matrix, with N indicating the number of players and $\mathbf{X}$ the state space.

In the next sub-section, we will treat reinforcement learning from a multi-agent point of view.

2) MARL: A possibility that has been extensively explored in the MARL domain is the straightforward use of the Q-learning algorithm while ignoring the presence of the other agents acting in the same environment and considering the results of this interaction as noise. In the following we will use the term "independent Q-learning" to refer to this approach. However, because of the non-stationarity of the environment caused by the presence of other agents, the theoretical result on convergence no longer holds. This means that for some games the agents may exhibit cyclic behavior. Despite its limitations, the independent Q-learning approach has been widely adopted in the cognitive radio literature. In some cases (e.g., [36]), the issues related to convergence are acknowledged and simulation results are presented to show that the agents achieve an equilibrium. In other cases (e.g., [37]), the question of convergence is not discussed.

An intuitive extension of the independent Q-learners approach is to maintain a Q-value for each combination of the states and actions of all agents. However this approach requires that the other agents' actions be observable. Most importantly, the curse of dimensionality, which already poses serious challenges in the single-agent domain, raises even more important issues in this case.

Furthermore, the main issue with the use of independent Q-learners is that the update is based on the agent's own maximum payoff in the next state. This is hardly justified in the multi-agent domain, as the agent's payoff in the next state depends on the other agents' actions [38].

Various attempts have been made to find a different update rule, more suitable to the multi-agent case. A useful example to better understand the strong relationship between MARL and game theory is the Nash Q-learning algorithm [35]. This approach clearly acknowledges the interactive nature of the learning involved in the multi-agent domain by modeling the MARL problem as a stochastic game. In particular, a modified version of the Q-learning rule is proposed. Each agent updates its Q-table using the expected return corresponding to the NE of the stage games corresponding to the states of the stochastic game. This approach, however, requires that each agent be able to compute an NE in every stage game given by all the agents' Q-tables. This means that each agent has to maintain the Q-tables for all the other agents, i.e. it has to observe the other agents' actions and rewards. Moreover, all agents have to agree on using the same NE. This requires a coordination mechanism for all but a restricted class of games where all the agents achieve the maximum expected return in correspondence to the same NE. It is unclear how strong a role this or similar approaches based on game theoretic analysis will play in the context of cognitive radio applications, due to their strict requirements and their sensitivity to noisy observations. More recent models for games of imperfect private or public monitoring can be used to model such noisy observations, but they come at the cost of significant increased complexity.

A common feature of most MARL algorithms is the use of a discrete state-action space. This is a heritage from the classic single-agent approach. Moreover, generally algorithms derived from Q-learning algorithm can only learn deterministic policies. A notable exception to the above observations is Hyper-Q [39], where the agent state includes an estimate of the other agents' strategies. As the Q-function evaluates the other agents' mixed strategy, Hyper-Q employs a function approximator.

A possible solution to these limitations is provided by direct policy search methods. This class of algorithms tries to directly learn the optimal policy, without attempting to approximate the value function. In other words, the learning problem is modeled as an optimization problem with unknown objective function. The policy is generally represented as a parametric function, and different approaches can be adopted to explore the strategy space (see [40] and references therein).

Some of the solutions proposed in the MARL literature use an opponent-independent closed-form solution for the matrix games (see for example [41]). In cognitive network applications this class of approaches is unlikely to play a key role but for a limited set of scenarios. In fact cognitive network applications are characterized by the intrinsic heterogeneity of the radios' behavior, due for example to hardware limitations. This feature will favor agents that are aware of and therefore can exploit other players' strategies. In this respect a class of approaches that learn a model of the other agents' strategies is of particular interest. Typically an agent chooses the best response based on its current model of the other agents' strategies. It then refines this model after observing the other agents' play. Examples of this class of approaches are fictitious play [42] and Joint Action Players [35]. In some cases the model of the other agents' strategy is simply based on a frequentist approach: an agent counts the number of times that another agent has selected a certain action. A simple but effective extension in the case of non-stationary strategies is the Exponential Moving Average, which assigns greater weight to the most recent observations and allows each player to react more quickly to the dynamics of the other players. More sophisticated techniques, based on a Bayesian approach, can also be used.

A number of MARL algorithms have been proposed that can only deal with repeated stateless games (see [40] and references therein). In the CR literature independent Q-learning has

also been used in this fashion [43]. It should be noted that the delayed reward, which is an essential feature of RL in general and Q-learning in particular, is no longer part of this simplified scenario. In the case of repeated games, more suitable RL schemes, such as learning automata [44], should be adopted. A learning automaton is a reinforcement learning scheme where each agent is a policy iterator, i.e. it directly updates its action probabilities based on the environment response. We have recently applied learning automata to the problem of distributed channel selection in the context of frequency-agile radios that are able to operate in multiple frequency bands simultaneously [13].

The general update rule is [44]:

$$\begin{aligned} p_i(t+1) = \quad & p_i(t) - (1-\beta_t) f_i(p(t)) + \beta_t g_i(p(t)) \quad \forall a(t) \neq a_i \\ p_i(t+1) = \quad & p_i(t) + (1-\beta_t) \sum_{j \neq i} f_j(p(t)) \\ & -\beta_t \sum_{j \neq i} g_j(p(t)) \qquad\qquad a(t) = a_i \end{aligned}$$

where the functions f and g are the reward and the penalty function, respectively, and $\beta_t \in [0,1]$ is the reward received by the agent at time t (with $\beta = 0$ corresponding to a favorable outcome). Different choices of the reward and penalty functions lead to different reinforcement schemes. Among them, the linear reward inaction scheme is of particular interest in that it has been proved to converge to a pure NE for special types of finite stochastic games, such as two-player zero sum games, N-player games with common payoff, and particular general sum N-player games [45]. For a review on the use of learning automata for adaptive wireless networks the reader is referred to [46].

When using the linear reward inaction scheme, the agent modifies its policy only when it receives a favorable feedback from the environment. In particular the penalty function is null, while the reward function is a linear function of the action probabilities. It should be noted that a linear reward inaction scheme can only converge to pure Nash equilibria [45].

Among the learning schemes that can only converge to a pure Nash equilibrium, the trial-and-error learning algorithm [47] is concise and of simple implementation. In fact, each player only maintains the last selected action and the corresponding perceived utility. At each time, each player decides to either perform the last selected action with probability $1-\epsilon$ or to randomly select another action with probability ϵ. If the player observes a strict increase in the payoff, the new strategy is adopted. If all players adopt trial-and-error learning, a pure Nash equilibrium is played at least $1-\epsilon$ fraction of the times, for any $\epsilon > 0$ [47]. In [48], the authors applied this result to the discrete power allocation problem, and observed that the number of iterations required to be close to a Nash equilibrium depends on ϵ and on the structure of the observed payoffs.

In general, RL algorithms select an action with probability proportional to the total reward received in the past as a result of choosing that action. In order to achieve a balance between exploration and exploitation, whilst avoiding the most unsatisfactory actions, a softmax action selection rule is generally adopted, where actions are ranked and weighted according to their estimated utility. The most commonly used softmax action selection rule is based on the Boltzmann-Gibbs distribution. In this case, the exploration/exploitation tradeoff is controlled by the temperature parameter τ. High values of τ determine a random action selection; low values of τ favour the selection of actions corresponding to higher rewards; $\tau \to 0$ corresponds to the greedy action selection scheme. A congestion game is a game where resources are limited and the utility of a player depends on which resources she chooses and how many other players chose the same resource. In [48] it is shown that, for congestion games, a learning scheme using the Boltzmann-Gibbs distribution to update the players' strategy almost surely converges to Nash equilibria.

A different approach, namely regret matching [49], also considers the hypothetical rewards the agent would have received by selecting actions it did not play. The agent associates to each action a regret, i.e. the difference between the average reward the agent would have received by always playing that action and the actual average reward. The agent then selects an action with probability proportional to the corresponding regret. Only actions with positive regret are considered. Although regret matching has been proved to converge to correlated equilibria in self-play, it makes strong assumptions on the agents' inputs. In fact, in order to compute the regret, each agent has to be able to observe all the other agents' actions.

The concept of regret is also used as an alternative evaluation criterion for learning algorithms. The no-regret criterion is verified when the average regret is less than or equal to zero against all other agents' strategies. For example in [50] the authors examine the performance of two algorithms for distributed channel selection providing bounds on the regret experienced by the secondary users while learning a channel access policy.

3) Pros and cons of different learning techniques: In [26] the authors present an interesting and useful comparison of some of the learning techniques discussed above with respect to the algorithms' requirements (computational complexity and assumptions on the agents' inputs) and to the convergence properties. However the analysis of RL is not conclusive, as RL is family of algorithms whose convergence properties and requirements depend on the particular implementation.

One of the fundamental issues common to all RL approaches is the convergence time when the dimension of the state-action space is beyond that of a toy problem. This aspect has not received sufficient attention in the CR literature. One exception is [10], where the spectrum pool, which corresponds to the action space, is randomly partitioned into different subsets in order to expedite the exploration stage. Although successful in facilitating the exploration, the obvious risk of this approach is that the CRs might converge to a suboptimal, and potentially inefficient, policy, as the exploration stage is blindly limited to a subset of the available actions. The problem of scaling up reinforcement learning has been well studied in the machine learning community. A possible solution is to use function approximation [51]. This approach allows an agent to generalize from previously observed states

and actions to an approximation of the action-value function for state-action pairs that have never been observed by the agent. This approach has been adopted in [11], where CRs use function approximation to determine channel assignment and transmission powers for large state problems.

As a final comment, the application of learning techniques to cognitive network problems should include an assessment of whether there is sufficient structure in the observation of the changing wireless environment (e.g., spectrum utilization patterns of a primary user) to justify trying to learn from these observations. We have tackled this question in [52], where we show the correlation between the Lempel-Ziv complexity of observed spectrum use and the benefits of a reinforcement learning approach in the secondary users' selection of a channel for opportunistic use.

IV. Reasoning Mechanisms

After the previous section's discussion of the relevant learning mechanisms within the cognitive networking context, this section will focus on the inference mechanisms needed to relate the acquired and the learned knowledge. These inference mechanisms are represented by *reasoning mechanisms*, which are also a quintessential part of the cognition process.

The field of reasoning is popular among psychologists, philosophers, and cognitive scientists. The development of cognitive networking imposes reasoning as a challenge for technologists and networking scientists as well. It is expected that a simple mapping of the reasoning process from other science fields will also fit the cognitive networking world. While this may seem mostly true, there are clear differences in the cognitive networking context that must be taken into account when analyzing the reasoning mechanisms and their associated aspects. We focus on those differences.

A. Cognitive frameworks and associated reasoning

As already discussed in section II, the general cognitive radio architecture depicted in Fig. 1 may be instantiated in various specific realizations. This proves to have a profound effect on the process of reasoning, since different cognitive architectures incorporate various approaches within the cognitive cycle.

Cognitive frameworks are generally classified into being *basic* or stemming from the *unified theory of cognition* [25], [53]. The basic ones can be symbolic, connectionist or hybrid. The frameworks stemming from the unified theory of cognition can be either simple, e.g. Observe-Orient-Decide-Act (OODA) and Critique-Explore-Compare-Adapt (CECA), or complex, e.g. SOAR, Storm and ACT-R [25].

The implementation of a specific cognitive framework reflects on the associated reasoning within. For instance, the OODA framework relies on a feedback loop to model adaptations to changing environmental conditions. The reasoning, i.e. the decision-making, involves identification of the available hardware configuration changes, identification of the best option to meet the new situation and implementation of the reconfiguration changes on the hardware in a constant feedback loop. This framework was originally developed by the US Department of Defense in order to describe the methodology that fighter pilots utilize during aerial combat and is applicable to reactive situations. The CECA framework expands the OODA framework to adequately describe a proactive decision-making process. The reasoning here is based on social cognition, i.e. multiple entities working on complex problems. This framework does not rely on reactive external observations, but focuses on proactive goal-oriented situations. Both OODA and CECA frameworks are applicable in cognitive radio networks.

The SOAR framework is a complex and powerful software suite designed to approximate rational behavior. Its complexity limits its application in cognitive radio networks. The Storm framework extends SOAR towards the development of Biologically Inspired Cognitive Architectures (BICA) and may be suitable for applications in cognitive radio networks. Finally, the ACT-R framework theorizes the way human cognition functions. It allows users to represent tasks and measure the time to perform a task and the accuracy of a task. This has potential application to decision-making in cognitive radio networks.

This section focuses on the reasoning and its possible types, methods, and realizations. Specific practical implementations within a complete cognitive framework will also be mentioned whenever applicable.

B. Reasoning types

There is a lack of straightforward logical categorization of the reasoning types in the cognitive networking world, as a result of the technical implementation peculiarities of a particular cognitive networking solution and the corresponding limitations, as well as of the potential applications and the corresponding requirements. Therefore, Table I briefly elaborates the most prominent reasoning types used within the field of cognitive networking today.

TABLE I
Classification of relevant reasoning types for cognitive networking

Reasoning type	*Explanation*
Proactive	Takes actions only when there is an indication of an impending problem; used where time constraints are more relaxed
Reactive	Prepares actions based on expected necessities for immediate actions and is more suitable for dynamic environments
Inductive	Forms hypotheses that seem likely based on detected patterns (conducive for cognitive radios)
Deductive	Forgoes hypotheses and only draws conclusions based on logical connections
One-shot	Selects a final action based on immediately available information
Sequential	Chooses intermediate actions and observes the response of the system following each action
Centralized	Higher degree of relationship between the inputs (actions) and the outputs (observations)
Distributed	Lower degree of relationship between the inputs (actions) and the outputs (observations)

C. Reasoning methods

The process of reasoning necessitates enablers (i.e. methods) of the inference goals. The most relevant ones within the

cognitive networking context are [54]:

- **Distributed constraint reasoning** - further classified as Distributed Constraint Satisfaction Problem (DisCSP) or Distributed Constraint Optimization Problem (DCOP). The former attempts to find any of a set of solutions that meets a set of constraints, whereas the latter attempts to find an optimal solution to a set of cost functions.
- **Bayesian networks** - a method of reasoning under uncertainty that can be a result of limited observations, noisy observations, unobservable states, or uncertain relationships between inputs, states, and outputs within a system.
- **Metaheuristics** - an optimization method that teams simpler search mechanisms with a higher-level strategy that guides the search. This method commonly employs randomized algorithms as part of the search process and, as a result, may arrive at a different solution each time it runs. Metaheuristics are a powerful method for tackling Non-deterministic Polynomial-time hard (NP-hard) problems.
- **Heuristics** - a method that exploits problem-specific attributes and may lead to increased performance of certain heuristic techniques. This method is not generic as the previous three.

A special form of a reasoning method is represented by *multi-objective reasoning*, which is used when there are multiple, potentially competing goals in the inference process. Therefore, multi-objective reasoning is essentially a multi-objective optimization problem and, as such, follows the characteristics of multi-objective optimizations. Recently, there are attempts to combine this approach with the DCOP.

D. Some specific reasoning realizations

Combining a specific reasoning type with a specific reasoning method (along with the inevitable and intertwined learning mechanisms) instantiates a specific reasoning realization that can be effectively deployed in a cognitive networking context. Some of the most relevant reasoning realizations are elaborated below.

Case-Based Reasoning (CBR). CBR [55] is a combination of reasoning and learning. The knowledge base is termed as the case base, where cases are representations of past experiences and their outcomes. The case base possesses a structured content in order to be easily shared among different entities within the cognition process and the cognitive network itself (termed agents). The sharing allows usage of past experiences and makes the network more robust and resilient. Usually, CBR involves 4-stage cycle: *retrieve*, *reuse*, *revise* and *retain*. CBR was used in a practical realization of the cognitive radio architecture for IEEE 802.22 applications [56]. The CBR engine is the focal point allowing decision-making in situations when secondary users must vacate a spectrum for a primary incumbent user.

Subsumption reasoning. Subsumption reasoning [57] essentially represents a decomposition of the target goal into smaller sub-goals and ideally with regard to their complexity. The decomposition leads to a set of layered modules operating in parallel that build upon each other, i.e. a hierarchical approach. Higher-level behaviors are assumed to function at a longer time scale and take advantage of complex optimization and learning functions such as partial plan generation and time series learning. Lower-level behaviors provide a tight coupling between the sensory input data and the actuation and often employ reactive learning algorithms with little to no state, such as self-organizing maps, decision trees, or hard codes input to output mappings. As a result, each layer realizes a sub-goal of the more complex overall goal.

Fuzzy logic reasoning. Fuzzy logic reasoning [58] relies on Fuzzy Logic (FL), which is a multivalued logic that allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low, etc. In an FL system, the knowledge of a restricted domain is captured in the form of *linguistic rules*, i.e. the relationships between two goals are defined using *fuzzy inclusion* and *non-inclusion* between the supporting and hindering sets of the corresponding goals. FL is helpful in very complex processes and is already applied in various telecommunications domains (e.g. QoS routing, caching, RRM, etc.). Lately, it has become popular for efficient reasoning (i.e. decision-making) in cognitive networks.

One of the most promising approaches to providing FL-assisted reasoning in cognitive radio networks is the usage of Fuzzy Cognitive Maps (FCMs) [59]. FCMs represent a means for modeling systems through the causal relationships that characterize them. Graphically, they are rendered as directed graphs in which a node represents a generic concept (e.g. an event or a process) and edges between any two nodes indicate that there is a causal relation between them. Their advantage lies in the power to handle feedback loops (unlike Bayesian networks), the straightforward inference method (simple multiplication and thresholding), and the ability to merge into a combined FCM that can smoothen discrepant biases stemming from the merging FCMs. However, there may be some disadvantages in practical cognitive network applications, since the inference of causality between events based only on observational data (without any a priori knowledge) is not immediate. Additionally, the abductive reasoning, i.e. the process of stating which causes are responsible for a given effect, is an NP-hard problem.

FCMs can be very useful in cognitive networks, facilitating cross-layering and using this information for reasoning within the cognition cycle. Reference [60] introduces a mathematical methodology able to represent the complex interactions among various protocol stack layers based on FCMs. The methodology is then applied to a sample test case of a VoIP WiFi system. The authors investigated the number of system-supported VoIP calls using given specific quality constraints on different protocol stack layers. The FCM framework was used to represent the correlation among the operating parameters on different layers and increase the overall system performance. Reference [61] extends the work in [60] by analyzing the scalability issues within the FCM framework when there exists a high number of cross-layer interactions. The authors discuss and propose a method for distinction among the cross-layer interactions that carry valuable information for the cognitive process. This should ensure that the reasoning in cognitive networks could converge to a solution before the environmental

conditions change, thus minimizing the reasoning time.

Relational reasoning. Relational reasoning [62] relies on the *relational structure* of propositional knowledge and the semantic features of objects and relational roles. It is enabled by the notion of *similarity*, which is a fundamental construct in cognitive science and inherently possesses featural and relational aspects. These aspects may allow for the development of relational analogies, which is common in human reasoning. However, [63] argues that these relational analogies may be modeled using sound scientific models capable of increasing the scientific literacy for the cognitive reasoning process.

A special reasoning realization that has attracted increased attention within the cognitive networking world is **policy based reasoning** [64]. It relies on the concept of dynamically derivable and interchangeable policies that surpass the traditional hardcoded firmware in current devices, offering higher flexibility and efficiency for the cognition process. The policies are expressed using a specific policy language consisting of a set of clearly defined *ontologies*. An ontology language defines the meaning of terms in vocabularies and their relationships [65]. Policy based reasoning is starting to be adopted by academia, industry and standardization [66] and regulatory bodies and may become a cornerstone of future efficient cognitive networking. Therefore, the following sub-sections will focus on a specific, already developed and operating, architectural realization of the policy based reasoning concept, along with its potential applications.

E. Case study: policy based reasoning

Fig. 4 depicts a fully functional architectural instantiation of the policy based reasoning concept [67]. The architecture embraces policies, expressed in CoRaL [64], coming from various stakeholders (e.g. operators, regulators, and users), offering options for each of them to express their specific goals. Moreover, the architecture supports dynamic resource management through dynamic policy changes that reflect the different behavior of the terminals. The full set of policies is efficiently reasoned and the reasoning output is presented to the resource management system (represented by the Cognitive Resource Manager - CRM) as an available solution set.

1) Architectural components and interfaces: The proposed policy system architecture comprises three main elements: a *policy server*, a *Policy Engine (PE)* and a *Policy Handling Toolbox (PHT)*. The first one is located on the network side, while the other two are terminal-based policy elements (Fig. 4).

Policy Server. The policy server is the central policy repository in the network, storing policies coming from the operator and regulator sides. It comprises:

- **Policy Server Database (PSD)** - for keeping track of all active users and active policies in the network and the user/policies associations.
- **Policy Server Database Handler (PSDH)** - for managing the database (storing policies and registering users into the database), disseminating the policies to the users, and informing them about policy changes.
- **Policy Manager (PM)** - for extracting the policies from the database, making the appropriate changes (add/change/delete policies) and reflecting the changes back in the PSD.

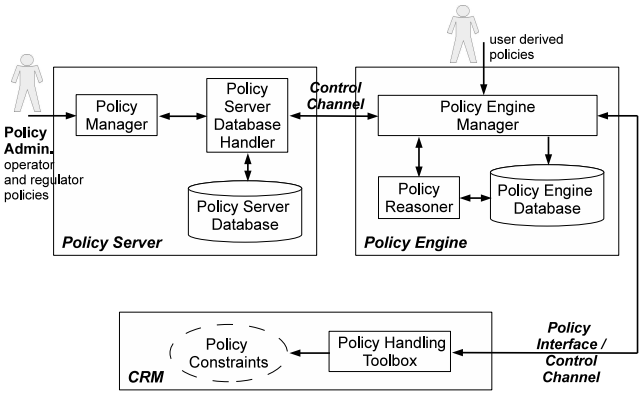

Fig. 4. Policy system architecture.

Policy Engine (PE). The PE is the *policy decision point* in the proposed policy architecture. It is located in the terminal and is responsible of reasoning on the set of active policies and presenting the reasoned result to the CRM. In order to be capable of performing the previously mentioned assignments, the policy engine consists of three components:

- **Policy Engine Database (PED)** - for local storage of the operator and regulator policies dedicated to the host user, as well as the locally derived user policies.
- **Policy Engine Manager (PEM)** - for handling the communication of the PE with the other policy network entities.
- **Policy Reasoner (PR)** - for performing the reasoning process on the set of policies in PED after every received policy query, thus providing the solution space to the CRM. The PR used in the proposed policy framework is the XG Prolog PR [68] with modified and extended functionalities. The crucial improvement to the XG PR is the support of ”why not permitted?” response from the reasoning process. This is important because it highly improves the conformance checking process and, as a result, it minimizes the time required to converge to a permitted solution.

Policy Handling Toolbox (PHT). The PHT is an integral part of the CRM which is the *policy enforcing point* in the architecture. The CRM is responsible for optimization, learning and decision making. The PHT creates and sends policy language-specific requests to the PE. In the opposite direction, the policy replies are received and provided in CRM-understandable fashion.

Interfaces. As illustrated on Fig. 4, the policy architecture yields two key interfaces, the **policy interface** (supporting the local communication between the PE and the CRM) and the **control channel interface** (handling the communications between the policy server and the PEs of the nodes).

The policy architecture also includes the specification of a custom *policy protocol* [67], which defines the communication between the policy components via the defined policy and control channel interfaces.

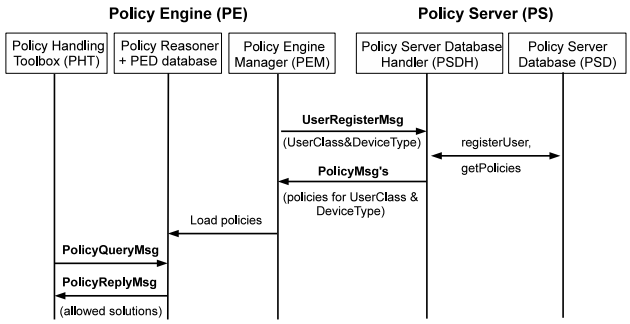

Fig. 5. User registration and policy checking process.

2) Policy architecture functionalities: The elaborated policy architecture incorporates many functionalities [67]. This subsection briefly describes some of them that are crucial for the subsequent understanding and elaboration of the policy-based reasoning applications.

User and/or terminal classification. The organization of the PSD provides a feature for policy classification and dissemination based on the users' class and device type. Each terminal registers to the policy server at start up, announcing its user class and device type. As a response, the relevant policies are received from the policy server (Fig. 5).

Efficient policy checking mechanism. The policy architecture has an efficient and flexible policy conformance checking mechanism. When the policy request is not permitted, a list of alternative solutions is formed utilizing the "why not permitted?" response (Fig. 5).

Dynamic policy management. The proposed architecture offers a framework for dynamic network resource management utilizing policies. When there is a policy change in the PSD (either manually input or emergency-triggered), the changes are immediately distributed to the users (terminals) of interest, so the changes can be reflected in their behavior instantly (Fig. 6).

For more extensive details on the elaborated policy architecture and its functionalities, the reader is referred to [67].

F. Applications of policy-based reasoning

The potential applications of the policy-based reasoning concept and the previously elaborated policy architecture are firmly stemmed in the cognitive networking context. They allow crucial cognitive networking operations such as spectrum opportunity detection, spectrum mobility, spectrum management, etc. This sub-section discusses some of the possible applications, along with results obtained on a testbed implementation of the policy-based architecture.

1) Spectrum handover: The ability to perform spectrum handover (i.e. switching between different channels or spectrum bands) is an essential cognitive networking concept. The policy-based reasoning can significantly facilitate fast and accurate spectrum handover, fostering the cognitive networking viability and wide range deployment in various scenarios. Fig. 7 depicts testbed results on the throughput of an RTP over

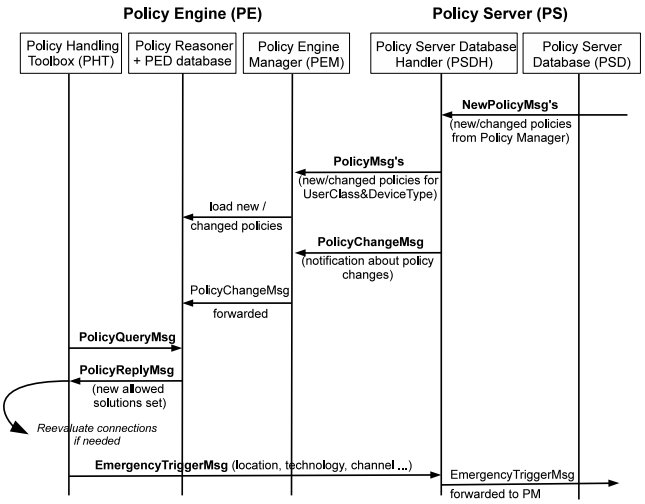

Fig. 6. Reporting of policy changes.

UDP based streaming application while performing policy controlled channel switching in an IEEE 802.11 ISM environment (WiFi). The testbed consists of a central PS performing channel and policy management for two USRP2 [69] enabled laptops aiming to establish communication using the following storyboard:

1) The PS sends both USRP2 nodes predefined policies specifying the allowed WiFi channels.
2) The source USRP2 forms a spectrum map (top-down power ranking of available channels) and chooses the best solution for the RTP over UDP streaming (in this case WiFi channel 3).
3) A policy that forbids WiFi channel 3 is manually input in the PSD.
4) The policy change is immediately sent to the USRP2 nodes, enforcing their reconfiguration in order to change the channel and perform appropriate spectrum handover.
5) The PR calculates a new solution and passes it to the USRP2s. The source USRP2 repeats step 2 and combines both pieces of information to select the best channel solution (in this case, WiFi channel 1).

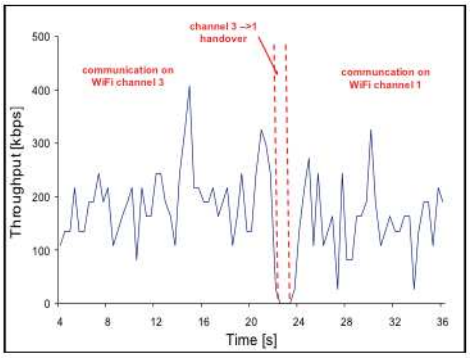

Fig. 7. RTP over UDP streaming throughput.

The *application handover delay* is around 1.5s including all actions performed during the channel switching. However, the *system reaction time to policy changes* (exchange of policy related messages and performing the reasoning) is only around 200ms. The rest of the handover time is due to the actual USRP2 characteristics.

2) Spectrum opportunity detection: The policy-based reasoning architecture can be used to efficiently detect spectrum opportunities and translate them into policies, which will easily govern the cognitive network behaviour afterwards [70]. Figs. 8-10 depict 2.4GHz ISM band channel occupancy measurements in an indoor scenario in a time period between 9:30 and 21:30 for typical working days. The results show that channel 10 (2457 MHz) experiences the highest utilization during the day and is not suitable for potential secondary usage. The frequency ranges 2400-2408 MHz and 2470-2480MHz are practically underutilized and are subject to potential secondary usage. The frequencies in the range 2408-2430 MHz can also be used for secondary access, because of relatively low utilization. However, one should be cautious not to harm potential primary users in this range, and therefore a CSMA/CA medium access should be used with a backoff slot higher than the standard IEEE 802.11 MAC. Finally, the frequency range 2426-2448 MHz is available only after 17:00, with lower transmission power levels in order not to violate the SINR requirements of potential primary users.

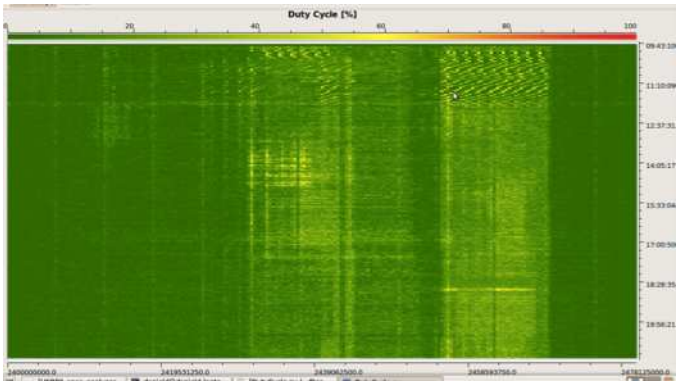

Fig. 8. Duty cycle (%) measurements on the 2.4GHz ISM band (x-axis) in time period between 9:30 and 21:30 hours (y-axis).

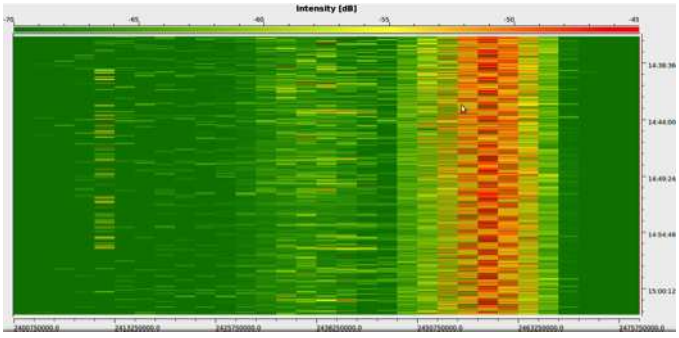

Fig. 9. Received signal power readings (intensity in dB) on the 2.4GHz ISM (x-axis) band before 17:00 (y-axis).

The previous elaboration can be translated into CoRaL

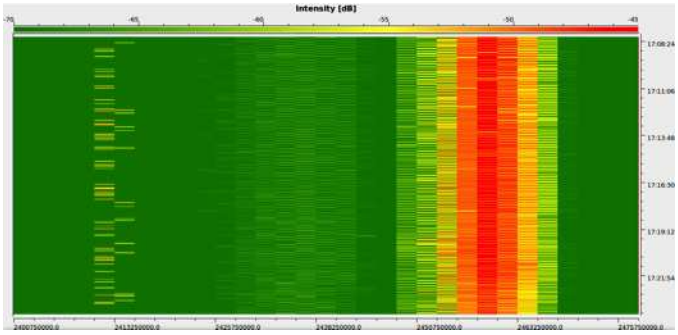

Fig. 10. Received signal power readings (intensity in dB) on the 2.4GHz ISM band (x-axis) after 17:00 (y-axis).

policies specifying the secondary spectrum access conditions:

policy specOpp1 is
use request_params;
defconst loc1 : Location = loc(42.004, 21.408, 0.0);
allow if
distance(onLocation(req_transmission),loc1)=<20 //20m from the loc1 point
{centreFrequency(req_transmission) in {2401..2407} or //in MHz
centreFrequency(req_transmission) in {2471..2479}} and //in MHz
meanEIRP(req_transmission)=<30 and
bandwidth(req_transmission)=<2.5; //in MHz
end

policy specOpp2 is
use request_params;
defconst loc1 : Location = loc(42.004, 21.408, 0.0);
allow if
distance(onLocation(req_transmission),loc1)=<20 //20m from the loc1 point
centreFrequency(req_transmission) in {2409..2429} and //in MHz
meanEIRP(req_transmission)=<30 and
bandwidth(req_transmission)=<2.5 and
//in MHz macType(req_datalink) == csmaca and
backoff(req_datalink)>=10 //in ms
end

policy specOpp3 is
use request_params;
defconst loc1 : Location = loc(42.004, 21.408, 0.0);
defconstallowedPeriod : TimePeriod;
startTime(allowedPeriod,"T17:00:00");
endTime(allowedPeriod,"T08:00:00");
allow if
distance(onLocation(req_transmission),loc1)=<10 and //10m from the loc1 point
inTimePeriod(onTime(req_transmission), allowedPeriod) and
centreFrequency(req_transmission) in {2427..2447} and //in MHz
meanEIRP(req_transmission)=<30 and

bandwidth(req_transmission)=<2.5 and //in MHz
macType(req_datalink) == csmaca and
backoff(req_datalink)>= 10 //in ms
end

The first policy (i.e. *specOpp1*) specifies that transmissions are allowed within 20m of the defined location ”loc1”, on central frequencies in the ranges 2401-2407MHz and 2471-2479MHz using a mean Equivalent Isotropically Radiated Power (EIRP) of 30dBm and a bandwidth of 2.5MHz. The second policy (i.e. *specOpp2*) specifies that transmissions be allowed within 20m of the defined location ”loc1”, on central frequencies in the range 2409-2429MHz using a mean EIRP of 30dBm and a bandwidth of 2.5MHz. Additionally, this policy specifies that the nodes use CSMA/CA as a MAC procedure with backoff time slot duration of 10ms. Finally, the third policy (i.e. *specOpp3*) allows transmissions within 10m of the defined location ”loc1” in the time period 17:00-08:00, on central frequencies in the range 2427-2447MHz using a mean EIRP of 30dBm, a bandwidth of 2.5MHz and a CSMA/CA MAC procedure with a backoff time slot duration of 10ms.

The policy system can afterwards use these CoRaL spectrum policies in order to regulate the secondary access to the 2.4 GHz ISM band for multiple secondary users. The following sub-section elaborates this aspect in more detail.

3) Spectrum sharing: The derived policies can be efficiently used to share the available spectrum among multiple secondary users. The potential of the policy-assisted spectrum sharing application is investigated with a laboratory testbed comprising several unaware secondary USRP2 based users that try to access and use the 2.4 GHz ISM band. The usage of this band is regulated according to the rules of the active secondary system policies specified in the PS residing on a desktop computer. The desktop computer is also enabled with a sensing capability so that it can dynamically derive and change *secondary spectrum policies*. Furthermore, the desktop computer is enabled with reasoning capabilities and performs the policy reasoning, resulting in the secondary USRP2 based users getting already reasoned information in the form of an available solutions set. Whenever a policy change occurs (because of a change in the environment, manual change, etc.), the new solution set is calculated (reasoned) and the secondary users are informed about the changes and the new solutions.

The secondary users’ policies are dynamically planned considering spectrum occupancy history (similar graphs as in Figs. 8-10). The policy server keeps two tables, i.e. a *short term occupancy decisions table*, saving the channel vacancy decisions in the last several minutes, and a *medium term history* reflecting the spectrum availabilities in the last couple of hours. Then, a channel is considered as an opportunity if the duty cycle of the channel activity is below a predefined threshold for 10% of the time in the short term history. However, the entire frequency band in the short term history is divided into 1 and 2 MHz non-overlapping channels, in proportion 30% (at most) and 70% (at least if possible) of the available spectrum. The 1 MHz channels are the ones that, although unoccupied in the short term table, were detected as used in the medium term history. Therefore, from a spectrum opportunity detection point of view, these channels are treated as riskier than the 2 MHz channels. Furthermore, when the targeted 70% for the spectrum assigned to 2 MHz channels is not fulfilled, i.e. some of the channels are repossessed by the primary system, secondary users can occupy the bands already assigned to non-priority 1 MHz channels. The testbed comprises two priority classes of USRP2 based secondary users, i.e. a higher priority class (aiming to establish real-time video streaming communication) and a lower priority class (targeting file transfer communication). The first class is allowed to use 2 MHz and 1 MHz channels, while the second class is only allowed to use 1 MHz channels.

Fig. 11 depicts the assigned bandwidth through time for the two types of channels, the priority and non-priority channels, as well as the detected available bandwidth through time [71]. It can be concluded that the priority channel assignment is more static through time and, therefore, the higher priority class would experience fewer forced terminations by the primary system. This is due to the fact that the priority users have the ”exclusive right” to the medium term history of the duty cycle in the bands of interest. Another reason is that whenever the targeted number of priority channels falls below the current assigned (due to environment changes, 70% of the current available bandwidth), the exceeding number of priority channels are not released, in order not to force termination on the priority users. In contrast, the assigned non-priority bandwidth through time follows the available bandwidth curve, i.e. is more dynamic through time and adapts to more dynamic environment changes.

The results from the policy-assisted secondary sharing show that the proposed scheme is flexible and efficient, since it enables dynamic secondary system channel allocation and classification using policies. The channel classification into priority and non-priority secondary channels ensures that priority users (or applications) will experience higher QoS than the non-priority ones.

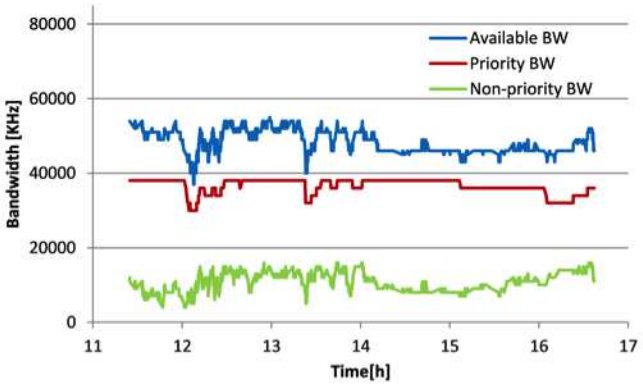

Fig. 11. Available bandwidth vs. assigned bandwidth for the priority and non-priority channels through time.

G. Pros and cons of different reasoning mechanisms

Different reasoning mechanisms exhibit different behavior in ’intelligent’ wireless networks. There is no single reasoning

approach that can suit and accommodate the plethora of possible applications of cognition in wireless networks. Therefore, it is often extremely important to have as much as possible a priori knowledge of the environment (i.e. observations) so that proper actions (i.e. outputs) are inferred.

Proactive reasoning is applicable to wireless environments that have relaxed time constraints. This implies that the channel characteristics are not rapidly changing, allowing for increased reasoning time and more reliable reasoning results. The proactive reasoning schemes are often combined with sequential and centralized reasoning mechanisms in order to use the available time for several intermediate reasoning results and relying on more closely related inputs/outputs of the system. This ensures that the system's reaction upon every intermediate reasoning result is carefully scrutinized and used in the process of converging towards an optimal reasoning solution. Examples of such reasoning approaches include cognitive wireless backhauling or secondary spectrum access in TV white spaces. These scenarios assume a more static environmental context in the spatial locations of interest and application, thus allowing for proactive, sequential and centralized reasoning.

In contrast to that, dynamic wireless environments exhibit fast changes, leading to time restrictions when it comes to cognitive reasoning. In this case, reactive reasoning is more suitable, as it can shorten the reasoning time and perform the reasoning within the specified time constraints. Reactive reasoning does not rely on past knowledge, but forms imminent actions based on immediately available information or on the expected need for immediate actions. This form of reasoning is often combined with one-shot and distributed reasoning schemes and is expected to become a major cognitive reasoning mechanism for future dynamic cognitive radio networks. Examples of these types of reasoning can be found in cognitive ad-hoc networks and cognitive cellular networks. These scenarios assume very dynamic wireless environments imposing serious time limitations on the reasoning part of the cognitive cycle.

Table II summarizes the main advantages and disadvantages of the specific reasoning realizations in cognitive radio networks today.

TABLE II
SUMMARY OF REASONING REALIZATIONS IN COGNITIVE RADIO NETWORKS

Reasoning realization	*Advantages*	*Disadvanatages*
CBR	Robust and resilient; Allows distributed sharing of knowledge	Inductive; Not suitable for highly dynamic environments
FL reasoning	Application in complex systems; Fosters easier cross-layer optimizations	Need for a priori knowledge; Abductive
Subsumption reasoning	Decomposition of the decision-making process; Ability to handle complex optimization tasks	Requires more implementation resources
Relational reasoning	Close to human cognition; Easy expression of analogies	Requires sound scientific models for analysis; Analogies can be misleading
Policy-based	Easy expression of various environmental limitations; Easy management of environmental context expressed in policies; Very scalable; Simple reasoning engine (comparisons of policies in a database)	Not suitable for dynamic environments; Reasoning time can be long if many policies exist in the system

V. CONCLUSIONS

The focal point of cognitive radio networks that clearly distinguishes them from other wireless networking solutions is their ability to learn, build knowledge bases, and reason upon stored knowledge. They tightly integrate these concepts into a unified networking framework able to 'learn' its environmental surrounding and taken actions and efficiently 'reason' in order to infer knowledge and perform optimal decision-making. Additionally, the cognitive framework also comprises the characteristics of autonomous 'observation' of the radio environment and optimal 'adaptation' to current conditions based on past perceived or maybe even future predicted actions. Therefore, cognitive radio networks are envisioned as a multidisciplinary engineering challenge integrating concepts from artificial intelligence and wireless networking sciences.

This paper focuses on the learning and reasoning challenges within cognitive radio networks. It discusses prominent learning mechanisms able to efficiently model the behavior of cognitive radio nodes. Furthermore, the paper gives a broad overview of the notion of reasoning, discusses the most relevant reasoning mechanisms and frameworks today, and in particular focuses on policy based reasoning as an efficient and implementable mechanism for deploying cognitive behavior in wireless networks.

A. Future Directions

The introduction of learning and reasoning into cognitive radio networks still faces some challenges. Some of the theoretical algorithms have high implementation complexity, limiting their practical implementation. This also gives rise to the problem of real-time algorithm convergence, of upmost importance for practical deployments.

Game theory has proven to be a powerful tool to model adaptations by cognitive radios, as well as emerging market mechanisms to support dynamic sharing of spectrum. Some of the challenges include translating the theoretical results about stability of adaptations and convergence to an equilibrium into practical and scalable adaptation mechanisms and an enforceable spectrum sharing etiquette. The field of mechanism design, so successful in the design of auction rules, may hold the key to broader protocol design for cognitive radios coexisting and competing for scarce resources, even in the absence of monetary incentives. Our research community has also only scratched the surface in the analysis of the impact

of partial or even inaccurate information on the actions taken by these radios. Relatively recent developments in games of imperfect public and private information have the potential to yield new insight into what we have termed the price of ignorance [72].

In terms of machine learning, one interesting aspect that requires further investigation in a cognitive network scenario is the concept of delayed reward. Most of the cognitive radio literature focuses on the maximization of the immediate reward, whereas the RL paradigm aims at optimizing the long term performance by taking into account the consequences of the agent's actions into the future. An important factor which needs further attention is the convergence time of many of the RL learning algorithms discussed in the paper. This aspect will become more and more significant with the increase of the degrees of freedom of a CR, i.e. of the cardinality of the action space of a CR. For example, if we consider the combined channel and power selection problem, the dimension of the action space for a realistic scenario does not allow the use of the traditional look-up table approach to store the value function. It is not unrealistic to envisage a scenario where a cognitive network will be required to dynamically perform carrier aggregation and, therefore, to decide how many and which channels it should access, thus further (in a combinatorial manner) increasing the number of decision variables.

There are several cautionary perspectives when discussing practical applications of cognitive reasoning in wireless networks. As already mentioned, the implementation complexity may seriously limit the entire solution, thus an optimal tradeoff between resources and expected outcomes is a must. Furthermore, the process of cognitive reasoning is inevitably time consuming, giving rise to the aspect of *reasoning time*. It is common to think that longer reasoning times yield better results, but this may become problematic in dynamic environments (especially in wireless networks). Namely, longer reasoning time may result in environmental changes that would need to be taken into account anew, thus leading to an ever increasing delay and, sometimes, even non-convergence of the reasoning process. In this sense, it is extremely important to address the number of reasoning inputs that will be used for the process. An efficient reasoning engine assumes careful selection of important and unimportant knowledge within the cognitive cycle, making the reasoning closely intertwined with the learning. Finally, the choice of the reasoning framework and approach requires accurate estimation of the environmental context and is strongly affected by and dependent on the precision of the other cognitive cycle elements.

As the field of cognitive radio networks attracts increased academic and industry interest, new standards must foster platform independence and cover the plethora of currently envisioned scenarios and potential applications of the cognitive radio paradigm. The IEEE DySPAN Standards Committee [73] is intensively working on these challenges, attempting to provide a common terminology, provide coexistence and conformance mechanisms, and propose efficient policing of cognitive radio networks along with appropriate policy languages and necessary ontologies. All these aspects provide a broad foundation for research in the area of cognitive radio networks, especially in the distinct sub-areas of learning and reasoning,. The end-goal is to provide autonomous and cognitive behavior of wireless networks in the future wireless interconnected world.

ACKNOWLEDGMENT

This work was partially supported by COST Action IC0902, on Cognitive Radio and Networking for Cooperative Coexistence of Heterogeneous Wireless Networks and the EC FP7 ICT-257626 NoE ACROPOLIS. It is also based upon works supported by the Science Foundation Ireland under Grant No. 10/CE/I1853. Authors V. Atanasovski and L. Gavrilovska express their gratitude to Mr. D. Denkovski for his fruitful collaboration in the development of and the experimentation with the policy testbed.

REFERENCES

[1] J. Mitola, "Cognitive Radio An Integrated Agent Architecture for Software Defined Radio," Ph.D. dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2000.

[2] Y. Wu, B. Wang, K. Liu, and T. Clancy, "Repeated open spectrum sharing game with cheat-proof strategies," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1922–1933, 2009.

[3] I. Malanchini, M. Cesana, and N. Gatti, "On spectrum selection games in cognitive radio networks," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2009, pp. 1–7.

[4] O. Raoof, Z. Al-Banna, and H. Al-Raweshidy, "Competitive spectrum sharing in wireless networks: a dynamic non-cooperative game approach," *Wireless and Mobile Networking*, pp. 197–207, 2009.

[5] J. Huang and V. Krishnamurthy, "Transmission control in cognitive radio as a markovian dynamic game: Structural result on randomized threshold policies," *IEEE Transactions on Communications*, vol. 58, no. 1, pp. 301–310, 2010.

[6] M. van der Schaar and F. Fu, "Spectrum access games and strategic learning in cognitive radio networks for delay-critical applications," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 720–740, 2009.

[7] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 1904–1919, 2009.

[8] B. Wang, K. Liu, and T. Clancy, "Evolutionary game framework for behavior dynamics in cooperative spectrum sensing," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2008, pp. 1–5.

[9] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: an evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, 2009.

[10] T. Jiang, D. Grace, and P. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET*, vol. 5, no. 10, pp. 1309–1317, 2011.

[11] C. Wu, K. Chowdhury, M. Di Felice, and W. Meleis, "Spectrum management of cognitive radio using multi-agent reinforcement learning," in *9th International Conference on Autonomous Agents and Multiagent Systems: Industry track*, 2010, pp. 1705–1712.

[12] B. Lo and I. Akyildiz, "Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks," in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2010, pp. 2244–2249.

[13] I. Macaluso, L. DaSilva, and L. Doyle, "Learning Nash Equilibria in Distributed Channel Selection for Frequency-agile Radios," in *Workshop on Artificial Intelligence for Telecommunications and Sensor Networks*, 2012.

[14] A. Rubinstein, "Instinctive and cognitive reasoning: A study of response times," *EconPapers*, no. 2006.36, 2006.

[15] O. M. Anshakov and T. Gergely, *Cognitive Reasoning*. Springer, 2010.

[16] L. Bass, J. Ivers, M. Klein, and P. Merson, "Reasoning frameworks," Carnegie Mellon, Software Engineering Institute, Tech. Rep. CMU/SEI-2005-TR-007, 2005.

[17] K. E. Nolan, P. Sutton, and L. Doyle, "An encapsulation for reasoning, learning, knowledge representation and reconfiguration cognitive radio elements," in *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, 2006.

[18] E. Adamapoulou, K. Demestichas, P. Demestichas, and M. Theologou, "Enhancing cognitive radio systems with robust reasoning," *International Journal of Communications Systems*, vol. 21, no. 3, 2008.

[19] N. Samaan and A. Karmouch, "Circumscriptive context reasoning for automated network management operations," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2006.

[20] S. Musman, "Using parallel distributed reasoning for monitoring computer networks," in *IEEE Military Communications Conference (MILCOM)*, 2010.

[21] Y.-G. Cheong, Y.-J. Kim, S. Y. Yoo, H. Lee, S. Lee, S. C. Chae, and H.-J. Choi, "An ontology-based reasoning approach towards energy-aware smart homes," in *IEEE Consumer Communications and Networking Conference (CCNC)*, 2011.

[22] B. Bahrak, A. Deshpande, M. Whitaker, and J.-M. Park, "Bresap: A policy reasoner for processing spectrum access policies represented by binary decision diagrams," in *IEEE Intl. Symp. on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2010.

[23] Q. Mahmoud, *Cognitive Networks: Towards Self-Aware Networks*. Wiley, 2007.

[24] X. Y. Wang, A. Wong, and P.-H. Ho, "Extended knowledge-based reasoning approach to spectrum sensing for cognitive radio," *IEEE Transactions on Mobile Computing*, vol. 9, no. 4, 2010.

[25] A. Amanna and J. H. Reed, "Survey of cognitive radio architectures," in *IEEE SoutheastCon*, 2010.

[26] L. Rose, S. Lasaulce, S. Perlaza, and M. Debbah, "Learning equilibria with partial information in decentralized wireless networks," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 136–142, 2011.

[27] D. Niyato and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of nash equilibrium, and collusion," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 192–202, 2008.

[28] A. MacKenzie and L. DaSilva, "Game theory for wireless engineers (synthesis lectures on communications)," 2006.

[29] J. Neel, J. Reed, and R. Gilles, "Convergence of cognitive radio networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 4, 2004, pp. 2250–2255.

[30] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," in *IEEE Intl. Symp. on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2005, pp. 269–278.

[31] R. Thomas, R. Komali, A. MacKenzie, and L. DaSilva, "Joint power and channel minimization in topology control: A cognitive network approach," in *IEEE Intl. Conf. on Communications (ICC)*, 2007, pp. 6538–6543.

[32] D. Niyato and E. Hossain, "Competitive spectrum sharing in cognitive radio networks: a dynamic game approach," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2651–2660, 2008.

[33] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. The MIT press, 1998.

[34] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[35] M. Bowling and M. Veloso, "An Analysis of Stochastic Game Theory for Multiagent Reinforcement Learning," in *Technical report CMU-CS-00-165, Computer Science Department, Carnegie Mellon University*, 2000.

[36] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823–1834, 2010.

[37] K. Yau, P. Komisarczuk, and P. Teal, "A context-aware and intelligent dynamic channel selection scheme for cognitive radio networks," in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, 2009.

[38] Y. Shoham, R. Powers, and T. Grenager, "Multi-agent reinforcement learning: a critical survey," in *Tech. Rep. Comput. Sci. Dept., Stanford University, Stanford, CA*, 2003.

[39] G. Tesauro, "Extending Q-learning to general adaptive multi-agent systems," in *Advances in neural information processing systems 16*, 2004.

[40] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156–172, 2008.

[41] J. Hu and M. Wellman, "Nash Q-learning for general-sum stochastic games," *The Journal of Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.

[42] D. Fudenberg and D. Levine, *The theory of learning in games*. MIT press, 1998.

[43] K. Yau, P. Komisarczuk, and P. Teal, "Performance Analysis of Reinforcement Learning for Achieving Context Awareness and Intelligence in Mobile Cognitive Radio Networks," in *IEEE International Conference on Advanced Information Networking and Applications (AINA)*, 2011.

[44] K. Narendra and M. Thathachar, *Learning automata: an introduction*. Prentice-Hall, Inc., 1989.

[45] P. Sastry, V. Phansalkar, and M. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 5, pp. 769–777, 1994.

[46] P. Nicopolitidis, G. Papadimitriou, A. Pomportsis, P. Sarigiannidis, and M. Obaidat, "Adaptive wireless networks using learning automata," *IEEE Wireless Communications*, vol. 18, no. 2, pp. 75–81, 2011.

[47] H. Young, "Learning by trial and error," *Games and economic behavior*, vol. 65, no. 2, pp. 626–643, 2009.

[48] S. Lasaulce and H. Tembine, *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. Academic Press, 2011.

[49] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.

[50] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[51] L. Baird, "Residual Algorithms: Reinforcement Learning with Function Approximation," in *12th International Conference on Machine Learning*, 1995, pp. 30–37.

[52] I. Macaluso, T. Forde, L. DaSilva, and L. Doyle, "Impact of cognitive radio: Recognition and informed exploitation of grey spectrum opportunities," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 85–90, 2012.

[53] A. Newell, *Unified Theory of Cognition*. Harvard University Press, 1994.

[54] D. H. Friend, "Cognitive networks: Foundations to applications," Ph.D. dissertation, PhD Thesis, VirginiaTech, 2009.

[55] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations and system approaches," *AI Communications*, vol. 7, no. 1, 1994.

[56] A. He, J. Gaeddert, K. K. Bae, T. R. Newman, J. H. Reed, L. Morales, and C.-H. Park, "Development of a case-based reasoning cognitive engine for ieee 802.22 wran applications," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 13, no. 2, 2009.

[57] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, 1991.

[58] M. J. Kaur, M. Uddin, and H. K. Verma, "Analysis of decision making operation in cognitive radio using fuzzy logic system," *International Journal of Computer Applications*, vol. 4, no. 10, 2010.

[59] B. Kosko, "Fuzzy cognitive maps," *Int. J. Man-Mach. Stud.*, vol. 24, no. 1, pp. 65–75, 1986.

[60] C. Facchini and F. Granelli, "Towards a model for quantitative reasoning in cognitive nodes," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2009.

[61] C. Facchini, F. Granelli, and N. L. S. da Fonseca, "Identifying relevant cross-layer interactions in cognitive processes," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[62] E. Taylor and J. E. Hummel, "Finding similarity in a model of relational reasoning," *Elsevier Cognitive Systems Research*, 2009.

[63] D. F. Sibley, "A cognitive framework for reasoning with scientific models," *Journal of Geoscience Education*, vol. 57, no. 4, 2009.

[64] D. Elenius and et al., "Coral - policy language and reasoning techniques for spectrum policies," in *8th IEEE International Workshop on Policies for Distributed Systems and Networks*, 2007.

[65] B. Chandrasekaran and et al., "What are ontologies, and why do we need them?" *IEEE Intelligent Systems*, vol. 14, no. 1, 1999.

[66] "IEEE 1900.5 working group on policy language and policy architectures for managing cognitive radio for dynamic spectrum access applications," Information available at: http://grouper.ieee.org/groups/dyspan/5/index.htm.

[67] D. Denkovski, V. Pavlovska, V. Atanasovski, and L. Gavrilovska, "Novel policy reasoning architecture for cognitive radio environments," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[68] "XG Prolog Policy Engine," Available at: http://xg.csl.sri.com/prolog.php.
[69] "Universal Software Radio Peripheral 2 (USRP2)," Information available at: http://www.ettus.com.
[70] D. Denkovski, V. Atanasovski, and L. Gavrilovska, "Policy enforced spectrum sharing for unaware secondary systems," in *4th International Conference on Cognitive Radio and Advanced Spectrum Management (CogART)*, 2011.
[71] "EC FP7 QUASAR (248303) project. Deliverable D2.2: Methodology for assessing secondary spectrum usage opportunities," 2010.
[72] R. Komali, R. Thomas, L. DaSilva, and A. MacKenzie, "The price of ignorance: distributed topology control in cognitive networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1434–1445, 2010.
[73] "IEEE DySPAN Standards Committee (DySPAN-SC)," Information available at: http://grouper.ieee.org/groups/dyspan.

Liljana Gavrilovska currently holds the position of full professor and Head of the Institute of Telecommunications at the Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University in Skopje. She is also Head of the Center for Wireless and Mobile Communications (CWMC) working in the area of telecommunication networks and wireless and mobile communications. She has received her B.Sc, M.Sc and Ph.D. from Ss Cyril and Methodius University in Skopje, University of Belgrade and Ss Cyril and Methodius University in Skopje, respectively. Prof. Gavrilovska participated in numerous EU funded projects such as ASAP, PACWOMAN, MAGNET, MAGNET Beyond, ARAGORN, ProSense, FARAMIR, QUASAR and ACROPOLIS, NATO funded projects such as RIWCoS and ORCA and several domestic research and applicative projects. Her major research interest is concentrated on cognitive radio networks, future mobile systems, wireless and personal area networks, cross-layer optimizations, broadband wireless access technologies, ad hoc networking, traffic analysis and heterogeneous wireless networks. Dr. Gavrilovska is author/co-author of more than 150 research journal and conference publications and technical papers and several books. She is a senior member of IEEE.

Vladimir Atanasovski currently holds the position of assistant professor at the Institute of Telecommunications at the Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University in Skopje. He has received his B.Sc, M.Sc and Ph.D. from Ss Cyril and Methodius University in Skopje, in 2004, 2006 and 2010, respectively. Dr. Atanasovski participated in numerous EU funded projects such as PACWOMAN, MAGNET, ARAGORN, ProSense, FARAMIR, QUASAR and ACROPOLIS, NATO funded projects such as RIWCoS and ORCA and several domestic research and applicative projects. Dr. Atanasovski is an author/co-author of more than 90 research journal and conference publications and technical papers. His major research interests lie in the areas of cognitive radio networks, resource management for heterogeneous wireless networks, traffic analysis and modeling, cross-layer optimizations, ad-hoc networking and nanonetworks.

Irene Macaluso is a Research Fellow at CTVR - The Telecommunications Research Centre based at Trinity College, Dublin. Dr. Macaluso received her Ph.D. in Robotics from the University of Palermo in 2007. Dr. Macaluso's current research interests are in the area of cognitive radio networks, with particular focus on the application of machine learning to wireless resource management and reconfigurable wireless networks.

Luiz A. DaSilva holds the Stokes Professorship in Telecommunications in the Department of Electronic and Electrical Engineering at Trinity College Dublin. He is also a Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech, USA. His research focuses on distributed and adaptive resource management in wireless networks, and in particular cognitive radio networks, dynamic spectrum access, and the application of game theory to wireless networks. Prof. DaSilva is currently a principal investigator on research projects funded by the National Science Foundation in the United States, the Science Foundation Ireland, and the European Commission under Framework Programme 7. He is a co-principal investigator of CTVR, the Telecommunications Research Centre in Ireland. He has co-authored two books on wireless communications and in 2006 was named a College of Engineering Faculty Fellow at Virginia Tech.