Reinforcement Learning For Nash Equilibrium Generation

David Cittern
Imperial College London, UK
david.cittern10@imperial.ac.uk

Abbas Edalat
Imperial College London, UK
a.edalat@imperial.ac.uk

Abstract

We propose a new conceptual multi-agent framework which, given a game with an undesirable Nash equilibrium, will almost surely generate a new Nash equilibrium at some predetermined, more desirable pure action profile. We provide convergence proofs for the two-agent, two-action game to show how applying a targeted, operant conditioning-like mechanism to one or both of the agents will almost surely guarantee the emergence of this new equilibrium as a result of changing the payoff matrix, when the agent(s) targeted for reinforcement learn independently according to a standard model-free algorithm. We consider both the case in which the additional reward is the result of an internal (re-)appraisal, such that the new equilibrium is stable independent of the continued application of the procedure; and when it represents additional reward that comes directly from the environment, which results in equilibrium decay if the reinforcement process is terminated. We also consider how the evolution of the game structure is affected by the relative values placed on immediate/future rewards, and the set of outcomes that are reinforced, and we discuss our results within the context of internally-generated reward and state representations.

1 Introduction

One of the biggest challenges in game theory and multi agent systems is the problem of how independent and self-interested agents who do not communicate can be guided towards stable behaviour at some particular action profile that has been deemed desirable [36]. The problem is complicated by the fact that existing Nash equilibria can sometimes correspond to outcomes that are either inefficient, or undesirable within a wider social context. Moving forwards, the issue is likely to become more significant. For example, it has been argued that purely reward-seeking agents will inevitably become adversarial in nature towards humans [4]. In addition, with increasing autonomy and generality, "the design of reward mechanisms eliciting desired agent behaviour become both more important and more difficult" (the "reward engineering principle" [11]).

We propose a new conceptual framework in which an external agent acts as a reinforcer by deterministically applying targeted payoff reinforcement to some subset of the agents ("reinforced agents"). Here we focus on the two-agent, two-action case, however the framework can be generalised to n-agents with almost sure (i.e. probability 1) convergence still guaranteed (see Appendix). A high-level ordinal state representation ("Q-state") is introduced over the evolving payoff matrix ("M-state") of each reinforced agent. The reinforced agents then use their Q-states in order to learn to adapt their behaviour to the changing environment in a trial-and-error process. For independent Q-learning agents, we will show how a predetermined new, strict Nash Equilibrium (NE) in pure actions will almost surely emerge under the directed influence of the external agent.

Reinforcement learning has traditionally considered all reward as emanating directly from the environment. Recently, however, frameworks such as Intrinsically Motivated Reinforcement Learning (IMRL) have begun to distinguish between extrinsic reward (tied to task-related, extrinsic motivation) and intrinsic reward (generated according to fulfilment of agent-specific goals) [8]. In particular, some IMRL-based frameworks have considered emotion as the appraisal mechanism driving intrinsic reward generation [33] [41]. Under the former (traditional) model, the additional reward in our framework can be considered as being temporary in nature and, thus, on termination of the process the equilibrium will gradually decay. In contrast, under the second (more biologically plausible) perspective, we can consider the external agent as sending signals in order to influence the internal appraisal mechanisms underlying the generation of intrinsic reward, so that changes to the payoff matrix can be considered as being more permanent in nature. Although we don't consider the specifics of the appraisal/reappraisal mechanism here, our results provide motivation for implementations based on these principles.

The remainder of this paper is organised as follows. In Section 2 we outline related work, before detailing our new framework in Section 3 and providing convergence proofs and related results in Section 4. In Section 5 we present simulation results for the evolution of a particular game representing a parent-child attachment relationship. Finally, in Section 6, we provide a summary and some suggestions for future work.

2 Related Work

A game-theoretic mediator is a reliable, external agent, who can optionally play on behalf of each participant in the game, but who cannot enforce behaviour. Many variants have been considered, for example in [25] the inclusion of the mediator changes the structure of the game, resulting in the emergence of a "mediated equilibrium" which is additionally a strong equilibrium (i.e. resistant to coalition deviation). In [16] the performance of the concept of soft correlated equilibrium is studied in congestion games, where the game is considered to be in equilibrium if no agent can gain unilaterally by deviating from the mediator's recommendation. Closely related are arbitration frameworks, in which the agents *must* conform to the external agent's (arbitrator's) recommendations on how to act (e.g. final offer arbitration [14]).

Within these frameworks, the mediated outcome will only persist so long as the mediating agent remains present; a requirement that is undesirable under many scenarios. In addition, there is a reliance on all of the agents being aware of, and both able and willing to communicate with, the mediator. In order to address these issues, we propose a framework under which an external agent conditions behaviour in either one or both of the agents by reinforcing the payoffs for certain outcomes following their occurrence. Our framework achieves a form of conflict resolution by modelling the gradual evolution of a new, stable equilibrium which, once generated, will (under assumptions of internal reappraisal) persist regardless of whether or not the external agent continues to exert an influence over the game, and which additionally does not necessarily require that the external agent be in contact with all agents.

Dynamic games in which the payoffs change have previously been considered in Stochastic Games (SG) [34], which are the n agent generalisation of a Markov Decision Process (MDP) [18]. In an SG the payoff matrix forms the current state and can change between each round according to some stationary probabilistic function over the current state and joint action choices. They have been studied as a framework for non-cooperative multi-agent reinforcement learning [23]: mixed SGs in particular, in which no constraints are imposed on the reward functions of the agents, are a suitable model for self-interested agents acting within a dynamic environment [7].

Whilst the environment in a multi-agent SG is stationary over state and action vectors in an objective sense, for each agent it is non-stationary since its sensing of the environment subsumes the behaviour of the remaining agents, and thus optimal policy convergence is not always guaranteed for independent Q-learners [38]. With this in mind, multi-agent extensions to Q-learning have often involved some form of opponent modelling, and have considered Q values over joint actions rather than just the agent's own action (we refer to [3] for a good comparative evaluation of techniques). An influential example is Nash-Q learning [19], which, under certain conditions (including that the agent observes not only its own reward but those of all other agents as well), will provably converge on an NE policy. In contrast to Nash-Q learning, our framework does not require opponent modelling.

In reinforcement learning, shaping rewards are additional rewards that are supplied to an agent with the aim of increasing convergence rate to an optimal policy. Potential-Based Reward Shaping (PBRS) is an influential variant in which additional reward is defined according to the difference between a potential function over pre and post-transition states. In particular, it has been shown that PBRS does not change the underlying optimal policy for a single-agent MDP [28]. Subsequently, PBRS was extended to potential advice over both states and actions [39]. PBRS has additionally been considered in the multi-agent domain, where importantly it has been proven that, although these additional rewards affect exploration, they do not change the NE policy in the underlying SG [10]. In contrast, we consider convergence to a new NE, using additional reward signals based only on the target action profile and observed outcomes. The work in [42] considered how additional incentive rewards could be associated with MDP states in order to induce a particular, pre-specified policy, however only the single-agent setting is considered. [26] determined the minimum amount of additional reward required to induce cooperative behaviour between two stateless Q-learning agents playing a Prisoner's Dilemma game.

In IMRL [8], a distinction is made between extrinsic reward (related to extrinsic motivation and external, task-related goals) and intrinsic reward (related to intrinsic motivation and agent-specific goals, such as learning or exploration). A recent extension to the IMRL framework, called Emotion-based Intrinsically Motivated Reinforcement Learning (EIMRL), has proposed that the reward signal should constitute both extrinsic and intrinsic reward, with the intrinsic reward being generated according to a variation of emotional appraisal theory (in particular novelty, goal relevance, control and valance dimensions of appraisal) [33]. This separation of reward is more biologically plausible and in line with findings from neuroscience, which recognises an intricate reward process. Phasic dopamine release in midbrain dopamine neurons in the ventral tegmental area has long been linked to habitual learning, through the encoding of a model-free reward prediction error, with respect to homeostatic-based primary reward signals from the hypothalamus [20]. However, such signals are likely to encapsulate errors with respect to more than just primary reward. For example, it is currently believed that the orbitofrontal cortex and amygdala, circuits central to emotional processing, play a key role in the computation of reward predictions and errors, and that these are projected to midbrain dopamine neurons via the striatum [32, p.361].

A recent study applied a variant of EIMRL to a multi-agent spatial prisoner's dilemma [41]. The authors found experimentally that agents using particular appraisal mechanisms, taking into account a metric of social fairness, could come to behave in a more cooperative way compared to agents without such a socially-aware appraisal mechanism. Although we don't consider the specifics of appraisal mechanisms here, our conceptual framework and results provide motivation for the development of internal appraisal mechanisms following its principles. Such mechanisms will provide a robust way to change multi-agent behaviour in a dynamic, on-line and predictable manner, in cases in which rewards are intrinsically generated and cannot be altered directly by an outside influence, but instead can only be influenced in a gradual way using perceptual signals (as in biological systems).

3 Framework

We consider the game in Fig. 1 with a set of 2 agents $A = \{\alpha_1, \alpha_2\}$, where the action set for agent α_1 is $B_1 = \{\beta_{11}, \beta_{12}\}$, and for agent α_2 is $B_2 = \{\beta_{21}, \beta_{22}\}$. The initial payoff matrices are U and V for α_1 and α_2 respectively. We assume that this game has an initial NE in pure actions at $NE_{initial} = (\beta_{12}, \beta_{22})$, i.e. that $U_{22} \geq U_{12}$ and $V_{22} \geq V_{21}$. Our goal is to generate a new, strict NE in pure actions at $NE_{target} = (\beta_{11}, \beta_{21})$. Either one or both of the agents $\alpha_i \in R \subseteq A$ will take the role of "reinforced" agents, and any remaining agent $\alpha_j \in A \setminus R$ will be a "reactive" agent.

We now introduce a source of additional reward. As discussed previously, this additional reward can be considered as either coming directly from the environment (e.g. from an agent external to the game), or alternatively as resulting from an internal (re-)appraisal mechanism. The second case could correspond to, for example, a process of self-reflection or self-therapy within a human agent, or to a reappraisal triggered by a particular environmental signal. From this point onwards, we refer to the source of additional reward simply as the "external agent". The external agent is considered as deterministically exerting an influence over the reinforced agents partaking in the game. We assume that the reactive and reinforced agents do not

		α_2	
		β_{21}	β_{22}
α_1	β_{11}	U_{11}, V_{11}	U_{12}, V_{12}
	β_{12}	U_{21}, V_{21}	U_{22}, V_{22}

Figure 1: The initial game with a NE in pure actions at (β_{12}, β_{22})

communicate with each other, but the reinforced agent(s) are free to communicate with the external agent, who is not considered as being an agent that partakes in the game itself.

3.1 Reactive Agents

We assume that reactive agents are unaware of any change to the structure of the game being played and that they therefore continue to play according to their initial, static payoff matrix. It is also assumed that these agents will be playing an iterative strategy with a reactive characteristic; i.e. that they pick their move in each round based on their analysis of the pattern of play of the other agents. Such a reactive iterative strategy is one in which the reactive agent maximises its expected payoff based on a probability distribution over the last t moves that the other agents have chosen. In this study we concentrate on the simple case of this iterated strategy for which $t = 1$, i.e. a Best Response to Last Move (BRTLM) iterated strategy, whereby a reactive agent picks the action corresponding to its highest payoff, based on the assumption that the other agent will play the same move that it played in the previous round.

3.2 Reinforced Agents

The reinforced agents (either one or both of the agents) must change the value that they place on individual outcomes in order for a new pure action NE to emerge. Each reinforced agent therefore plays a dynamic game in which its payoff matrix changes according to a payoff reinforcement rule, and we model a learning process for each reinforced agent so that the reinforcement of certain desirable action combinations gradually leads all agents into a stable pattern of play at NE_{target}. Note that the NE for any stage game depends only on the relationship between the ordinal (i.e. relative) ranking of the payoffs, and not on their magnitude. It is therefore fair to assume that the agents consider the ranking over the potential payoffs when they come to choose their actions, and thus it is natural to use an ordinal representation for the agent's state. We also note that, in a practical sense, ordinal utility is much easier to elicit and measure than cardinal utility.

Definition 3.1. *For $U, V \in {\mathbb{R}^+}^{2\times 2}$ we define $U \equiv V$ iff:*

$$\begin{aligned} &U_{mn} < U_{m'n'} \Leftrightarrow V_{mn} < V_{m'n'} \quad \textit{and} \\ &U_{mn} = U_{m'n'} \Leftrightarrow V_{mn} = V_{m'n'} \end{aligned}$$

A complete set of equivalence classes for $\equiv$ is contained in ${\mathbb{N}_4^+}^{2\times 2}$, where $\mathbb{N}_4^+ = \{1, 2, 3, 4\}$, and for convenience we use this representation. Let $M \in {\mathbb{R}^+}^{2\times 2}$ be the current payoff matrix for $\alpha_i \in R$ (the "M-state"). We introduce the canonical representation of M under $\equiv$ by:

$$[M] := M_{/\equiv} \in E \subset {\mathbb{N}_4^+}^{2\times 2}$$

which we call the "Q-state", where:

$$\begin{aligned} E = \{X \in {\mathbb{N}_4^+}^{2\times 2} \mid \min_{mn}(X_{mn}) = 1,\ \forall m, n : \\ (X_{mn} = 1 \text{ or } \exists\ m', n' : X_{m'n'} = X_{mn} - 1)\} \end{aligned}$$

The Q-state $[M]$ is thus a dense ranking over α_i's payoff matrix M (such that equal payoffs receive the same ordinal value and the next outcome receives the immediately following ordinal value), with minimum ordinal value 1 in each Q-state. The current state for α_i is given by the (M-state, Q-state) tuple $(M, [M])$.

We say that reinforced agent α_i plays a "reinforced game", which is defined by the state transition system and transition rules given in Fig. 2. The state transition system is a 4-tuple, defined fully by the state space, α_i's initial state $(M^0, [M^0])$, reinforcement set η_i and reinforcement parameter $r_i > 1$. The initial state consists of α_i's initial payoff matrix M^0 (α_i's initial M-state) along with its equivalence $[M^0]$ (α_i's initial Q-state). The reinforcement set η_i is the set of outcomes that will trigger reinforcements in α_i's payoff matrix M, and the reinforcement parameter $r_i > 1$ specifies the magnitude of these reinforcements.

At each discrete step in time (representing a moment in which a decision is to be made) α_i's system is in some particular Q-state ($[M]$). We assume

$$\boxed{\quad\Big(\mathbb{R^+}^{2\times 2} \times E,\ \big(M^0,\ [M^0]\big)\ ,\ \eta_i,\ r_i\Big)} \tag{i}$$

$$\boxed{\begin{array}{l}
(M,[M]) \overset{(\beta_{1p},\beta_{2q})}{\longrightarrow} (M,[M]) \ \text{ if } \ (\beta_{1p},\beta_{2q}) \notin \eta_i \\
(M,[M]) \overset{(\beta_{1p},\beta_{2q})}{\longrightarrow} (M',[M']) \ \text{ if } \ (\beta_{1p},\beta_{2q}) \in \eta_i \\[1em]
\text{with } {M'}_{mn} = \begin{cases} r_i\ M_{mn} & \text{if } m = p \text{ and } n = q \\ M_{mn} & \text{otherwise} \end{cases}
\end{array}} \tag{ii}$$

Figure 2: (i) State transition system describing the reinforced game for agent α_i (ii) Transition rules

that the current M-state, and thus also the current Q-state, are fully observable to α_i. We also assume that each Q-state has the same action set B_i. A multiplicative reinforcement of magnitude r_i will be applied to M_{pq} following every occurrence of an action-combination outcome $(\beta_{1p}, \beta_{2q}) \in \eta_i$, resulting in reinforced payoff matrix M' as in Fig. 2 (ii). Whilst it would be possible to consider other types of reinforcements (e.g. an additive rule, or some form of convergent series), we chose to employ a multiplicative factor as the simplest case. We also note that our multiplicative factor has the desirable property of inducing proportional payoff increments.

We assume that reinforced agents learn how to act within the game according to a model-free Temporal Difference (TD)-based algorithm [35]. TD learning is a form of trial-and-error learning with roots in psychology and operant conditioning, where reward predictions are adjusted immediately following environmental feedback in order to improve subsequent predictions. It has long been argued that the properties of the TD error signal are reflected in the brain's dopamine system, with phasic firing patterns encapsulating a model-free reward-prediction error (we refer to [40] for a good overview).

Two prominent examples of TD-based algorithms for control are Q-learning and its on-policy variant SARSA, and evidence from animal studies has supported a biological basis for the prediction update mechanisms used in both [31] [27]. Here we employ the former on top of the underlying transition system as a simple model of how the choices that the agent makes will adapt over time, allowing us to capture an anticipation of future reward for deviating from the initial NE for different types of agents with differing discount factors. The Q function $Q : E \times B_i \rightarrow \mathbb{R}$ calculates a Q value for each action of the reinforced agent α_i associated with a particular Q-state: following the choice of action $\beta_{ij} \in B_i$ in the current Q-state $[M]$, we update the Q values according to the conventional single-agent update rule:

$$\begin{aligned} Q([M], \beta_{ij}) \leftarrow\ & Q([M], \beta_{ij}) + \ell(D(M, \beta_{ij}) \\ & + \delta_i \max_{\beta_{iq}} Q(s, \beta_{iq}) - Q([M], \beta_{ij})) \end{aligned}$$

where $s \in \{[M],\ [M']\}$ (i.e. the state may change, as according to Fig. 2), and $0 \leq \delta_i < 1$ is α_i's discount factor. The learning rate $0 < \ell \leq 1$ is set according to $\ell([M], \beta_{ij}) = (n([M], \beta_{ij}))^{-1}$, where $n([M], \beta_{ij})$ equals the number of times action β_{ij} has been chosen in Q-state $[M]$, so that initially $\ell([M], \beta_{ij}) = 1$ and decreases with each subsequent selection of action β_{ij} in Q-state $[M]$. The reward α_i receives for choosing action β_{ij} in state $(M, [M])$ is $D(M, \beta_{ij})$. Here, $D(M, \beta_{ij})$ is either a reinforced or non-reinforced payoff. In particular, if action-combination outcome $(\beta_{ij}, \beta_{pq}) \in \eta_i$ has just occurred then $D(M, \beta_{ij}) = r_i M_{jq}$. Alternatively, if $(\beta_{ij}, \beta_{pq}) \notin \eta_i$ has just occurred then $D(M, \beta_{ij}) = M_{jq}$. We employ a simple softmax action selection rule (based on Luce's choice axiom [24]), according to the following probability mass function:

$$P\left(\beta_{ij} | [M]\right) = k_i^{Q([M], \beta_{ij})} \ / \ \sum_j k_i^{Q([M], \beta_{ij})}$$

with exploration parameter $k_i > 1$ for α_i. Thus, reinforced agents choose their actions according to a path-dependent, non-stationary stochastic process.

To illustrate, consider as an example the game in Fig. 1, where α_1 is a reactive agent and α_2 is a reinforced agent. Following N rounds of play, where the outcome (β_{11}, β_{21}) has occurred j times, (β_{12}, β_{21}) has occurred m times, and (β_{11}, β_{22}) and (β_{12}, β_{22}) have occurred a total of $N - (j + m) \geq 0$ times, then given reinforcement set $\eta_2 = \{\ (\beta_{11}, \beta_{21}),\ (\beta_{12}, \beta_{21})\ \}$, α_2's payoff matrix will have been reinforced to $V(j, m)$ (Fig. 3), where the ordinal equivalence of this reinforced payoff matrix $[V(j, m)]$ is α_2's current Q-state in its state transition system. The payoff elements V_{12} and V_{22} have not been reinforced since $(\beta_{11}, \beta_{22}) \notin \eta_2$ and $(\beta_{12}, \beta_{22}) \notin \eta_2$.

$$V(j,m) = \begin{pmatrix} {r_2}^j V_{11} & V_{12} \\ {r_2}^m V_{21} & V_{22} \end{pmatrix}$$

Figure 3: Example of a reinforced payoff matrix (M-state). The densely-ranked ordinal equivalence of this payoff matrix forms the corresponding reinforced agent's current Q-state.

3.3 Alternative State Representations

One can consider alternative state representations. For example, the reinforced agent could use as its state representation its payoff matrix alone. In this case, the system would be a non-stationary, infinite transitional system with states given by $(j, m) \in \mathbb{N}^2$. In such a system, with $r_i > 0$ a constant, a state can never be re-visited, and thus such a representation could arguably just as well be modelled by a one-state framework such as a non-stationary multi-armed bandit [21].

For a such a single-state representation, our framework would also converge and would do so at a faster rate than for the formulation defined above. However, the focus of this paper is on dynamic, internally-generated state representations, for which the consider the simplest case consisting of a high-level preference ranking over outcomes. We discuss the framework within the context of richer state formulations (e.g. also incorporating external perception and memory representations) in Section 6.

4 Convergence

In Section 4.1 we give a convergence proof for the two-agent, two-action game given in Fig. 1, for a single reinforced agent and a single reactive agent. In Section 4.2, we provide a convergence proof for the case whereby both agents are reinforced agents, and each learn independently by treating the opposing agent as part of the environment.

4.1 Reinforcing A Single Agent

For reactive agent α_1 and reinforced agent α_2, we will show that the convergence criterion is:

$$\begin{gathered} U_{11} > U_{21} \quad \text{and} \\ \eta = \{NE_{target}\} \cup \zeta \quad \text{with} \quad \zeta \subseteq \{(\beta_{12}, \beta_{21})\} \end{gathered} \tag{C1}$$

i.e. the target NE outcome must be reinforced, and the deviation from the initial NE by α_2 can optionally also additionally be reinforced. If $U_{11} < U_{21}$ then by definition a new NE cannot be generated at $NE_{target} = (\beta_{11}, \beta_{21})$, since α_1's payoff matrix does not change. If $U_{11} = U_{21}$ then the dynamics will depend on how α_1 discriminates between outcomes with equal payoffs, although any new NE generated at NE_{target} will not be a strict NE. As an example, a coordinated outcome can be generated as a new NE in the battle of the sexes game using single agent reinforcement. We consider another example (an attachment game) experimentally in Section 5.

Lemma 4.1. *Suppose a sequence q_n of real numbers satisfies: $q_n \leq q_{n-1} + \frac{1}{n}(a + \delta q_{n-1} - q_{n-1})$ where $a \in \mathbb{R}$ and $\delta \in [0, 1)$ are constants. Then,* $\limsup_{n\to\infty} q_n \leq a/(1-\delta)$

Proof. Let $x_n := \frac{a}{1-\delta} - q_n$. Then:

$$x_{n-1} - x_n \leq ((1-\delta)/n)\, x_{n-1}$$

and thus:

$$\begin{aligned} x_n &\geq x_{n-1}\,(1 - ((1-\delta)/n)) \\ &\geq x_0\,(1 - ((1-\delta)/1))\,(1 - ((1-\delta)/2)) \ldots (1 - ((1-\delta)/n)) \end{aligned}$$

i.e. $\liminf_{n\to\infty} x_n \geq x_0 \liminf_{n\to\infty} \prod_{m=1}^{n} (1 - ((1-\delta)/m))$. Since $(1-\delta)\sum_{n=1}^{\infty} \frac{1}{n}$ diverges, it follows from [1, p.192-193] (Theorem 6) that the product diverges to zero. Hence, because

$$\liminf_{n\to\infty} -q_n = -\limsup_{n\to\infty} q_n$$

we have that $\limsup_{n\to\infty} q_n \leq a/(1-\delta)$. □

Corollary 4.1. *If $q_n = q_{n-1} + \frac{1}{n}(a + \delta q_{n-1} - q_{n-1})$ then $\lim q_n = a/(1-\delta)$*

Proof. We have $\limsup_{n\to\infty} q_n \leq a/(1-\delta)$ and by putting $p_n := -q_n$ we also get

$$\limsup_{n\to\infty} -p_n \leq a/(1-\delta)$$

or $\liminf_{n\to\infty} q_n \geq a/(1-\delta)$, from which the result follows. □

Suppose now a coupled pair of sequences $(p_m, q_n) \in \mathbb{R}^2$ for $m, n \in \mathbb{N}$ and some initial value for (p_0, q_0) is defined recursively by

$$p_m = p_{m-1} + \frac{1}{m}(a + \delta \max(p_{m-1}, q_n) - p_{m-1})$$

$$q_n = q_{n-1} + \frac{1}{n}(b + \delta \max(p_m, q_{n-1}) - q_{n-1})$$

where $0 \leq \delta < 1$ and $a \geq b$. At each point in time, either p_m is updated according to the above relation and m is incremented to $m+1$, or alternatively q_n is updated according to the above relation and n is incremented to $n+1$.

Lemma 4.2. *Assume that the sequences p_m and q_n are each updated infinitely many times. Then $(p_m, q_n) \to (a/(1-\delta), b - a + (a/(1-\delta))$*

Proof. Consider the doubly parametrised family of non-deterministic dynamical systems

$$(f^{(m)}, g^{(n)}) : \mathbb{R}^2 \to \mathbb{R}^2$$

$$f^{(m)}(p, q) = p + \frac{1}{m}(a + \delta \max(p, q) - p)$$

$$g^{(n)}(p, q) = q + \frac{1}{n}(b + \delta \max(p, q) - q)$$

for $m, n \in \mathbb{N}$. At each point in time, either $f^{(m)}$ or $g^{(n)}$ is chosen to act on (p, q). If $f^{(m)}$ is chosen then m is incremented by one whereas if $g^{(n)}$ is chosen then n is incremented by one.

For each $m, n \in \mathbb{N}$, this dynamical system has a unique fixed point at $(a/(1-\delta), b - a + (a/(1-\delta))$, which can be checked by solving:

$$p = p + \frac{1}{m}(a + \delta \max(p, q) - p)$$

$$q = q + \frac{1}{n}(b + \delta \max(p, q) - q)$$

In fact, by eliminating $\max(p, q)$, we obtain $p - a = q - b$, i.e., $p = q + (a - b)$ which implies $p \geq q$ since $a \geq b$. This gives using the first equation $a + \delta p - p = 0$ or $p = a/(1-\delta)$ and $q = -(a-b) + a/(1-\delta)$.

Now we change coordinates by letting $x = p - a/(1-\delta)$ and $y = q + a - b - a/(1-\delta)$ to obtain a new system $\mathbb{R}^2 \to \mathbb{R}^2$ given by:

$$(x, y) \mapsto x + \frac{1}{m}(\delta \max(x, y - (a-b)) - x)$$

$$(x, y) \mapsto y + \frac{1}{n}(\delta \max(x, y - (a - b)) - y)$$

The unique fixed point of the system is at the origin $(x, y) = (0, 0)$.

We consider two cases: (i) $x \geq y - c$ and (ii) $x < y - c$, where $c := a - b \geq 0$. Let $S_1 = \{(x, y) \in \mathbb{R}^2 : x \geq y - c\}$ and $S_2 = \{(x, y) \in \mathbb{R}^2 : x < y - c\}$. For (i), we obtain the following two maps to capture the non-deterministic dynamics and for convenience we drop the explicit reference to parameters m and n to define $F, G : S_1 \to \mathbb{R}^2$ with components F_1, F_2 and G_1, G_2 respectively:

$$F_1(x, y) = (1 - (1 - \delta)/m)x, \quad F_2(x, y) = y$$

$$G_1(x, y) = x, \quad G_2(x, y) = \delta x/n + (1 - 1/n)y$$

For (ii), we obtain the maps $H, T : S_2 \to \mathbb{R}^2$ with:

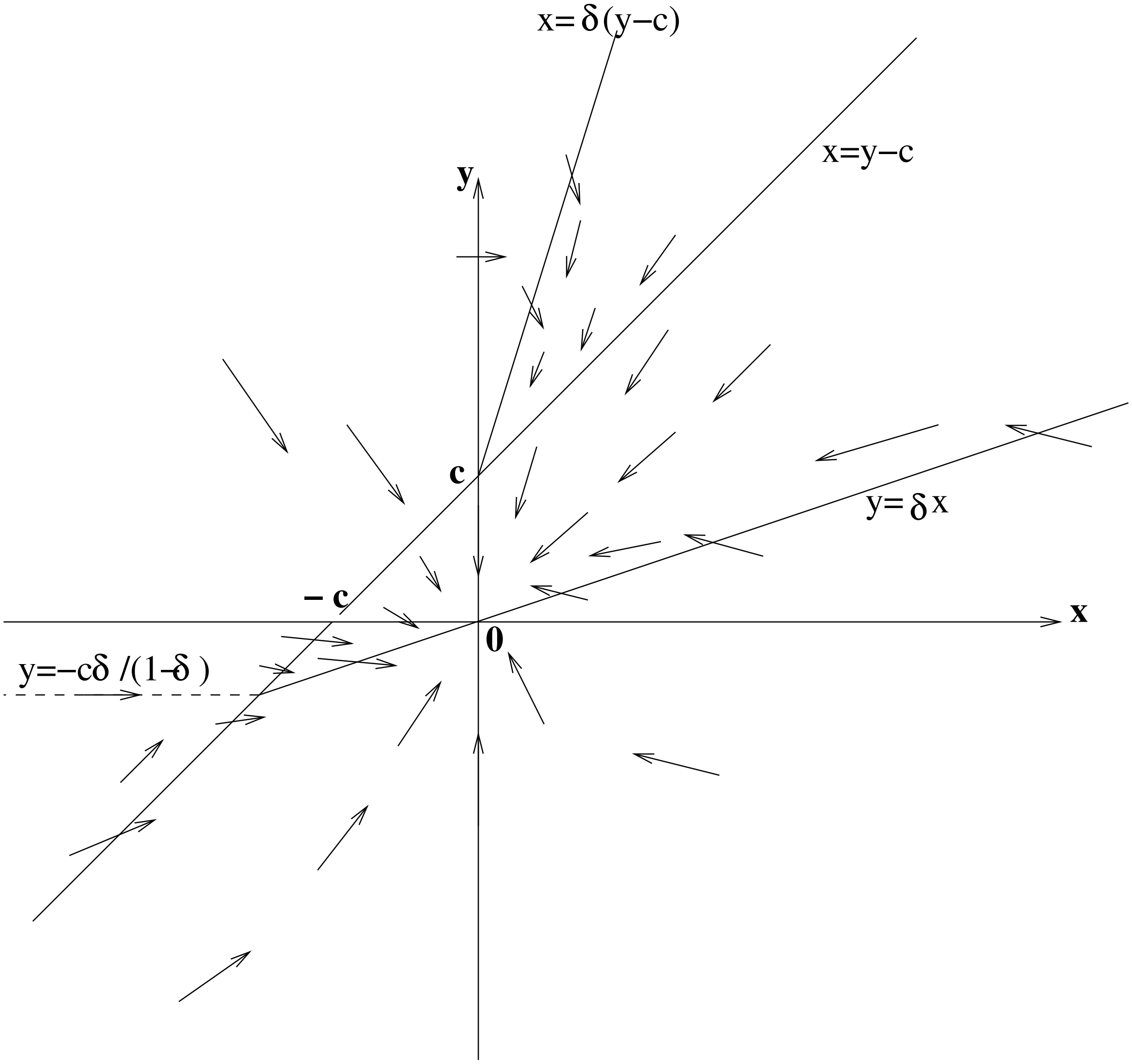
$$H_1(x, y) = (1 - 1/m)x + \delta y/m - \delta c/m, \quad H_2(x, y) = y$$

$$T_1(x, y) = x, \quad T_2(x, y) = (1 - (1 - \delta)/n)y - \delta c/n.$$

Given a point $(x, y) \in R^2$, the dynamics is now defined as follows. If $x \geq y - c$, then either the map F or G is selected to provide the next pair of elements for (x, y) and then either m or n is respectively incremented, whereas if $x < y - c$, then either H or T is selected to obtain the next pair (x, y) and then again either m or n is respectively incremented.

A simple calculation shows that

$$\begin{array}{rcl} x - F_1(x, y) & = & \frac{1}{m}(1 - \delta)x \\ x - H_1(x, y) & = & \frac{1}{m}(x - \delta(y - c)) \\ y - G_2(x, y) & = & \frac{1}{n}(-\delta x + y) \\ y - T_2(x, y) & = & \frac{1}{n}((1 - \delta)y + \delta c) \end{array}$$



From the above, the dynamics of the non-deterministic system can be determined from its phase portrait as in Figure 1. In the first quadrant, $x \geq 0$ and $y \geq 0$, the region bounded by the lines $x = \delta(y - c)$, $y = \delta x$ and the y axis, is invariant under the dynamics, any orbit entering into this region would stay in the region and converge to the origin $(0, 0)$. Similarly, in the third quadrant, i.e., for $x \leq 0$ and $y \leq 0$, the region bounded by the y axis, and the lines $y = \delta x$ and $x = y - c$ is an invariant region with every orbit converging to the origin. Orbits in the other regions will either enter into these two invariant regions first or will converge to the origin directly.

To prove the above assertions, we first show that the region R_1 bounded by the lines $x \geq 0$, $y = \delta x$ and the $x = \delta(y - c)$ axis, is invariant under the dynamics:

1. Suppose $(x, y) \in R_1 \cap S_1$, i.e., $x \geq 0$, $y \geq \delta x$, and $x \geq y - c$ (and thus $x \geq \delta(y - c)$).

    - To show $F_1(x, y) \geq \delta(F_2(x, y) - c)$, we consider two different cases. If $y \geq c$ then $F_1(x, y) - \delta(F_2(x, y) - c) = x(1 - \frac{1}{m}(1 - \delta)) - \delta(y - c) \geq (y - c)(1 - \delta)(1 - \frac{1}{m}) \geq 0$. If $y < c$, then $F_1(x, y) - \delta(F_2(x, y) - c) = x(1 - \frac{1}{m}(1 - \delta)) - \delta(y - c) \geq 0$ as both terms are non-negative.
    - To show $F_2(x, y) \geq \delta F_1(x, y)$, note that $F_2(x, y) = y \geq \delta x \geq \delta(1 - (1 - \delta)/m)x = \delta F_1(x, y)$.
    - To show $G_1(x, y) \geq \delta(G_2(x, y) - c)$, we calculate using $-\delta x \geq -y$: $G_1(x, y) - \delta(G_2(x, y) - c) = x - \delta(\delta x/n + (1 - 1/n)y - c) \geq \delta(y - c) - \delta(y/n + y - y/n - c) = \delta(y - c) - \delta(y/n - c) = \delta y(1 - 1/n) \geq 0$.
    - To show $G_2((x, y) \geq \delta G_1(x, y)$, we calculate: $G_2((x, y) - \delta G_1(x, y) = y(1 - \frac{1}{n}) + \delta x/n - \delta x = (y - \delta x)(1 - 1/n) \geq 0$.
    - We also have $F_1(x, y) = (1 - (1 - \delta)/m)x \geq 0$ and $G_1(x, y) = x \geq 0$.

2. Suppose $(x, y) \in R_1 \cap S_2$, i.e., $x, y \geq 0$, $x \geq \delta(y - c)$ and $x < y - c$ (and thus $y \geq \delta x$).

    - To show $H_1(x, y) \geq \delta(H_2(x, y) - c)$ we calculate: $H_1(x, y) - \delta(H_2(x, y) - c) = (1 - 1/m)x + \delta y/m - \delta c/m - \delta(y - c) \geq (1 - \frac{1}{m})\delta(y - c) + \delta y/m - \delta c/m - \delta(y - c) = (y - c)(1 - 1/m)(1 - \delta) \geq 0$, since $y \geq c$ in $R_1 \cap S_2$.
    - To show $H_2(x, y) \geq \delta H_1(x, y)$ we compute using $-x \geq -y + c$: $H_2(x, y) - \delta H_1(x, y) = y - \delta((1 - 1/m)x + \delta y/m - \delta c/m) = y - \delta(1 - 1/m)x - \delta(\delta y/m - \delta c/m) \geq y + \delta(1 - 1/m)(-y + c) - \delta^2(y - c)/m = y(1 - \delta(1 - 1/m) - \delta^2/m) + \delta(1 - 1/m)c + \delta^2 c/m = y(1 - \delta(1 - (1 - \delta)/m)) + \delta(1 - (1 - \delta)/m)c \geq 0$.
    - To show $T_1(x, y) \geq \delta(T_2(x, y) - c)$ we use $x \geq y + c$ to calculate: $T_1(x, y) - \delta(T_2(x, y) - c) = x - \delta((1 - (1 - \delta)/n)y - (\delta c/n) - c) \geq y + c - \delta(1 - (1 - \delta)/n)y + \delta^2 c/n + \delta c = y(1 - \delta + \delta((1 - \delta)/n)) + (\delta c/n) + \delta c = y(1 - \delta(1 - ((1 - \delta)/n))) + \delta c(1 + (\delta/n)) \geq 0$.

- To show $T_2(x,y) \geq \delta T_1(x,y)$ we calculate using $y \geq x+c$: $T_2(x,y) - \delta T_1(x,y) = (1-(1-\delta)/n)y - \delta c/n - \delta x \geq (x+c)((1-(1-\delta)/n)) - \delta c/n - \delta x = (1-1/n)(x(1-\delta)+c) \geq 0$.
- We also have $H_1(x,y) = (1-1/m)x + \delta y/m - \delta c/m \geq (1-\frac{1}{m})\delta(y-c) + \delta y/m - \delta c/m = (y-c)(1-(1-\delta)/m) \geq 0$ and also $T_1(x,y) = x \geq 0$.

We also note that for $(x,y) \in R_1 \cap S_1$, we have

$$\begin{aligned}
F_1(x,y) - x &= -\tfrac{1}{m}(1-\delta)x \leq 0 \\
F_2(x,y) - y &= 0 \\
G_1(x,y) - x &= 0 \\
G_2(x,y) - y &= -\tfrac{1}{n}(-\delta x + y) \leq 0
\end{aligned}$$

For $(x,y) \in R_1 \cap S_2$, we similarly have:

$$\begin{aligned}
H_1(x,y) - x &= -\tfrac{1}{m}(x - \delta(y-c)) \leq 0 \\
H_2(x,y) - y &= 0 \\
T_1(x,y) - x &= 0 \\
T_2(x,y) - y &= -\tfrac{1}{n}((1-\delta)y + \delta c) \leq 0
\end{aligned}$$

It follows that orbits in $(x,y) \in R_1$ move in the south-west direction and thus converge to the unique fixed point $(0,0)$.

Now we show that the region R_2 bounded by the y axis, $x = y - c$ and $y = \delta x$ is invariant under the dynamics. Suppose $(x,y) \in R_2$, i.e. $x, y \leq 0$, $y \leq \delta x$ and $x \geq y - c$ with $y < c$.

- We have that $F_1(x,y) = x(1-((1-\delta)/m)) \leq 0$ and $G_1(x,y) = x \leq 0$
- We also have that $F_2(x,y) = y \leq 0$ and $G_2(x,y) = \delta(x/n) + y(1-(1/n)) \leq 0$
- To show $F_2(x,y) \leq \delta F_1(x,y)$ we compute $\delta F_1(x,y) - F_2(x,y) = \delta x(1+((\delta-1)/m)) - y \geq \delta(y-c)(1+((\delta-1)/m)) - y = y(\delta(1+((\delta-1)/m)) - 1) - c(\delta(1+((1-\delta)/m))) \geq y(\delta(1+((\delta-1)/m)) - 1) - y(\delta(1+((1-\delta)/m))) = -y \geq 0$
- To show $G_2(x,y) \leq \delta G_1(x,y)$ we calculate: $\delta G_1(x,y) - G_2(x,y) = \delta x - (\delta(x/n) + y(1-(1/n))) = \delta x(1-(1/n)) - y(1-(1/n)) \geq \delta(y-c)(1-(1/n)) - y(1-(1/n)) \geq 0$

- To show $F_1 \geq F_2 - c$ we compute: $F_1(x,y) - F_2(x,y) + c = x(1 - ((1-\delta)/m)) - y + c \geq (y/\delta)(1 - ((1-\delta)/m)) + c - y = -y((1-m-\delta)/\delta m) + c - y \geq 0$
- To show $G_1(x,y) \geq G_2(x,y) - c$ we calculate: $G_1(x,y) - G_2(x,y) + c = x - ((\delta x)/n) - y(1-(1/n)) + c = x(1-(\delta/n)) - y(1-(1/n)) + c \geq x(1-(\delta/n)) - c(1-(1/n)) + c = x(1-(\delta/n)) + (c/n) \geq (y/\delta)(1 - (\delta/n)) + (c/n) = -y((\delta - n)/(\delta n)) + (c/n) \geq -c((\delta - n)/(\delta n)) + (c/n) = cn/(n\delta) \geq 0$

For $(x,y) \in R_2$, we have

$$\begin{aligned} F_1(x,y) - x &= -\tfrac{1}{m}(1-\delta)x \geq 0 \\ F_2(x,y) - y &= 0 \\ G_1(x,y) - x &= 0 \\ G_2(x,y) - y &= -\tfrac{1}{n}(-\delta x + y) \geq 0 \end{aligned}$$

It follows that orbits in $(x,y) \in R_2$ move in the north-east direction and thus converge to the unique fixed point $(0,0)$.

□

Corollary 4.2. *Suppose we have the non-deterministic, coupled recurrence relations:*

$$p_m = p_{m-1} + \frac{1}{m}(a + \delta \max(p_{m-1}, q_n) - p_{m-1})$$

$$q_n = q_{n-1} + \frac{1}{n}(\{b, b'\} + \delta \max(p_m, q_{n-1}) - q_{n-1})$$

where $\{b, b'\}$ means that either b or b' is chosen. Then $\limsup p_m \leq \max(p, p')$ *and* $\liminf p_m \geq \min(p, p')$*, where p and p' are the limits of p_m in the deterministic system when only b, respectively only b', are chosen. Similarly,* $\limsup q_n \leq \max(q, q')$ *and* $\liminf q_n \geq \min(q, q')$*.*

Proof. Suppose $b > b'$. Denote by $p_m(b)$ and $q_n(b)$ the pair of deterministic sequences that are updated as p_m and q_n except that the value b is always chosen in the recursive relation for q_n. Similarly, denote by $p_m(b')$ and $q_n(b')$ the pair of deterministic sequences that are updated as p_m and q_n except that the value b' is always chosen in the recursive relation for q_n. By our assumption, it follows that $p_m(b) \geq p_m \geq p_m(b')$ and $q_n(b) \geq q_n \geq q_n(b')$ from which the result follows. □

Corollary 4.3. *Now also suppose that either a or a' can be chosen in p_m, i.e.:*

$$p_m = p_{m-1} + \frac{1}{m}(\{a, a'\} + \delta \max(p_{m-1}, q_n) - p_{m-1})$$

$$q_n = q_{n-1} + \frac{1}{n}(\{b, b'\} + \delta \max(p_m, q_{n-1}) - q_{n-1})$$

Then $\limsup p_m \leq \max(p(a,b), p(a',b), p(a,b'), p(a'b'))$ *and* $\liminf p_m \geq \min(p(a,b), p(a',b), p(a,b'), p(a'b'))$*, where $p(a,b)$ is the limit of p_m in the deterministic system when only a and b are chosen. Similarly,* $\limsup q_n \leq \max(q(a,b), q(a',b), q(a,b'), q(a'b'))$ *and* $\liminf q_n \geq \min(q(a,b), q(a',b), q(a,b'), q(a'b'))$*.*

Proof. Suppose $a \geq a'$ and $b \geq b'$. Denote by $p_m(a,b)$ and $q_n(a,b)$ the pair of deterministic sequences that are updated as p_m and q_n except that the value a is always chosen in p_m, and the value b is always chosen in q_n. By our assumption we have $p_m(a,b) \geq p_m \geq p_m(a',b')$ and $q_n(a,b) \geq q_n \geq q_n(a',b')$, from which the result follows. □

Lemma 4.3. *If condition* (C1) *holds for the reinforced agent α_2's reinforcement set η, then at any arbitrary point in the reinforced game, the reinforced agent α_2 will almost never choose its actions according to the infinite, alternating sequence $(\beta_{22}, \beta_{21}, \beta_{22}, ...)$.*

Proof. Suppose that α_2 is in state $(V, [V])$, where β_{21} has been chosen a total of $n_{21} \geq 0$ times in Q-state $[V]$, and β_{22} a total of $n_{22} \geq 0$ times in Q-state $[V]$, such that the Q values are $Q_{n_{21}}([V], \beta_{21}) \geq 1$ for β_{21} and $Q_{n_{22}}([V], \beta_{22}) \geq 1$ for β_{22}. Since reactive agent α_1 has a deterministic action selection rule, we can state the outcome sequence for alternative β_{22}, β_{21} action choices by α_2:

$$(\beta_{11} \text{ or } \beta_{12}, \beta_{22}), (\beta_{12}, \beta_{21}), (\beta_{11}, \beta_{22}), (\beta_{12}, \beta_{21}), (\beta_{11}, \beta_{22}), ...$$

where the first choice of β_{11} or β_{12} by α_1 depends on whether α_2's previous action selection was β_{21} or β_{22}. For the case where α_2's reinforcement set is $\eta = \{NE_{target} = (\beta_{11}, \beta_{21})\}$, no reinforcements and thus no state transitions will occur as a result of this outcome sequence. Setting $Q_i([V], \beta_{21}) = a_i$ and $Q_j([V], \beta_{22}) = b_j$ in order to simplify the notation, then the infinite probability product is given by:

$$\begin{aligned}
&\prod_{\substack{i=n_{21}\\j=n_{22}}}^{\infty} \frac{k^{b_j}}{k^{b_j}+k^{a_i}} \frac{k^{a_i}}{k^{b_{j+1}}+k^{a_i}} \\
&= \prod_{\substack{i=n_{21}\\j=n_{22}}}^{\infty} \frac{1}{k^{b_{j+1}-a_i}+1+k^{a_i-b_j}+k^{b_{j+1}-b_j}}
\end{aligned} \tag{1}$$

Since $\lim_{j\to\infty} b_j$ exists (Lemma 4.2), it follows that for the sequence $c_j := b_{j+1} - b_j$, with $j \geq n_{22}$, we have $\lim_{j\to\infty} c_j = 0$. Since any convergent sequence is bounded there exists $t \geq 0$ so that $|c_j| \leq t$ for all $j \geq n_{22}$. Thus, each term in the product is less than $\frac{1}{1+k^{-t}} < 1$ and thus the infinite product is zero. As such, at any point in the reinforced game, the probability of the reinforced agent choosing its actions according to the infinite alternating sequence $(\beta_{22},\, \beta_{21}, \beta_{22}, ...)$ is zero.

Consider now the reinforcement set $\eta = \{(\beta_{11}, \beta_{21}), (\beta_{12}, \beta_{21})\}$: for the reinforced agent successively choosing $\beta_{22}, \beta_{21}, \beta_{22}, ...$, then starting with the initial payoff matrix, each of the (β_{12}, β_{21}) outcomes will result in a reinforcement on V_{21}. There is some finite number of reinforcements n on (β_{12}, β_{21}) required such that α_2 will transition to some state $(V^*, [V^*])$ in which $[V^*]_{21}$ is strictly a maximum. This state is achieved when $r^n V_{21} > \max\{V_{11}, V_{12}, V_{22}\}$, i.e. following $n = \lfloor log_r(\max\{V_{11}, V_{12}, V_{22}\}/V_{21})\rfloor + 1$ reinforcements on (β_{12}, β_{21}). Indeed, at any point in the reinforced game, for α_2 successively choosing β_{22}, β_{21}, β_{22}, ..., there will be some finite number of reinforcements on (β_{12}, β_{21}) required in order to transition to this Q-state $[V^*]$. From $[V^*]$, any number of further reinforcements on (β_{12}, β_{21}) will not result in a change to the Q-state, and we can see that the infinite product in (1) is again zero. □

Theorem 4.1. *If condition (*C1*) holds on the payoff matrix for the reactive agent α_1, and on the reinforced outcomes η for the reinforced agent α_2, then the reinforced game will converge to the new target NE almost surely.*

Proof. If condition (C1) holds, then in order for NE_{target} to emerge as a new strict NE we require n reinforcements on agent α_2's initial payoff V_{11}, such that $r^n V_{11} > V_{12}$. Therefore it follows that we require $n = \lfloor log_r(V_{12}/V_{11})\rfloor + 1$ reinforcements on V_{11} (n provides a lower bound on the number of reinforcements on (β_{11}, β_{21}) required for NE_{target} to emerge from any arbitrary point in the reinforced game).

Since the reactive agent α_1 is playing the BRTLM iterated strategy, we require agent α_2 to choose action $\beta_{21} \in NE_{target}$ two consecutive times in order for each reinforcement on NE_{target} to occur: n is therefore a lower bound on the number of consecutive choices of $\beta_{21} \in NE_{target}$ by agent α_2 required at some arbitrary point in the reinforced game as a sufficient condition for convergence. We have shown that, at any point in the reinforced game, the reinforced agent will almost never choose its actions according to the infinite sequence $(\beta_{22}, \beta_{21}, \beta_{22}, ...)$ (Lemma 4.3). We also have that, at any point in the reinforced game, the probability of some finite number $i' > 1$ of consecutive selections of β_{22} is clearly less than a single selection of β_{22}. Therefore it follows that:

$$\forall i_z \in \mathbb{N}_{>1}, n_{21} \geq 0, n_{22} \geq 0, [V] \in E :$$
$$\prod_{z=1}^{\infty} \left(\prod_{i=1}^{i_z} \frac{k^{Q_{n_{22}+i}([V],\beta_{22})}}{k^{Q_{n_{21}+z}([V],\beta_{21})} + k^{Q_{n_{22}+i}([V],\beta_{22})}} \right)$$
$$\cdot \frac{k^{Q_{n_{21}+z}([V],\beta_{21})}}{k^{Q_{n_{21}+z}([V],\beta_{21})} + k^{Q_{n_{22}+i_z}([V],\beta_{22})}} = 0$$

i.e., at any point in the reinforced game, the probability of the reinforced agent α_2 choosing its actions according to an infinite sequence consisting of some finite number of β_{22} selections, followed by a single β_{21}, followed by some finite number of β_{22} selections, etc, is 0. Thus, at any point in the reinforced game, agent α_2 will almost surely eventually select action β_{21} two consecutive times. □

4.2 Reinforcing Both Agents

For some games condition C1 does not hold on the initial payoff matrix (for example, for the Prisoner's Dilemma [5] and Snowdrift [22] games, where the desirable new NE is at the coordinated cooperation outcome). For such games we can instead reinforce both agents in order to almost surely guarantee the emergence of the new desirable NE.

Assume again that we start with the game in Fig. 1 with an initial NE in pure actions at (β_{12}, β_{22}), but now both agents are reinforced agents. Agent α_i is reinforced by factor $r_i > 1$, has exploration parameter $k_i > 1$, discount factor $0 \leq \delta_i < 1$ and reinforcement set η_i. Agents α_1 and α_2 have states $(U, [U]), (V, [V]) \in \mathbb{R}^{+2\times2} \times E$ respectively. We will show that the convergence criterion for the generation of a new strict NE in pure actions at (β_{11}, β_{21}) is:

$$
\begin{aligned}
\eta_1 &= \{NE_{target}\} \cup \zeta_1 \text{ with } \zeta_1 \subseteq \{(\beta_{11}, \beta_{22})\} \\
\eta_2 &= \{NE_{target}\} \cup \zeta_2 \text{ with } \zeta_2 \subseteq \{(\beta_{12}, \beta_{21})\}
\end{aligned}
\tag{C2}
$$

i.e. the target NE outcome must be reinforced for both agents, and the independent deviation from the initial NE by each individual agent can optionally also additionally be reinforced for that respective agent.

Suppose now one of the two sequences p_m and q_n given in Corollary (4.2), say q_n, is updated only a finite number of times from which point only p_m is updated. Then we have the recursive relation:

$$p_m = p_{m-1} + \frac{1}{m}(a + \delta \max(p_{m-1}, d) - p_{m-1})$$

where d is the final value of q_n after which it is not updated any more.

Lemma 4.4. *Suppose*

$$p_m = p_{m-1} + \frac{1}{m}(a + \delta \max(p_{m-1}, d) - p_{m-1}).$$

Then $\lim_{m\to\infty} p_m = a/(1-\delta)$ *if* $a \geq d(1-\delta)$ *and* $\lim_{m\to\infty} p_m = a + \delta d$ *if* $a < d(1-\delta)$.

Proof. We have the following two affine maps to capture the dynamics of the recursive relation: $f : [d, \infty) \to \mathbb{R}$ and $g : (-\infty, d)) \to \mathbb{R}$ given by

$$
\begin{aligned}
f(x) &= x + \frac{1}{m}(a + \delta x - x) = x(1 - (1-\delta)/m) + a/m \\
g(x) &= x + \frac{1}{m}(a + \delta d - x) = x(1 - 1/m) + (a + \delta d)/m
\end{aligned}
$$

It is easily checked that if $a \geq (1-\delta)d$ then f has its unique fixed point at $x = a/(1-\delta)$ which is the unique fixed point of the system; whereas if $a < (1-\delta)d$ then g has its unique fixed point at $x = a + \delta d$ which is the unique fixed point of the system. Suppose $a \geq (1-\delta)d$. Then, $x \geq d$ implies $f(x) - d = x(1-(1-\delta)/m) + a/m - d \geq (-d(1-\delta)+a)/m \geq 0$. If however, $x < d$, then $d < a + \delta d$ and g will be increasing in $(-\infty, d)$. Thus, there would be an integer m such that $p_m \geq d$. Since, for $x \geq d$, the dynamics is determined by f only it follows, by Corollary (1) in the main paper, that $\lim_{m\to\infty} p_m = a/(1-\delta)$. If however, $a < (1-\delta)d$, then a similar argument shows that $\lim_{m\to\infty} p_m = a + \delta d$. □

Corollary 4.4. *Suppose*

$$p_m = p_{m-1} + \frac{1}{m}(\{a, a'\} + \delta \max(p_{m-1}, d) - p_{m-1}).$$

Then $\limsup_{m\to\infty} p_m \leq \max(p, p')$ *and* $\liminf_{m\to\infty} p_m \geq \min(p, p')$ *where* p*, respectively* p'*, are the limits of* p_m*, given by Lemma* 4.4*, when only* a*, respectively only* a'*, is chosen.*

Proof. Suppose $a > a'$. As in Corollary 4.2, let $p_m(a)$, respectively $p_m(a')$, denote the sequences for which only a, respectively only a', is chosen. We then have $p_m(a) \geq p_m \geq p_m(a')$ from which the result follows. □

Lemma 4.5. *Suppose we have the following stochastic, coupled recurrence relation:*

$$p_m = p_{m-1} + \frac{1}{m}(a + \delta \max(p_{m-1}, q_n) - p_{m-1})$$

$$q_n = q_{n-1} + \frac{1}{n}(\{b, b'\} + \delta \max(p_m, q_{n-1}) - q_{n-1})$$

for $a, b, b' \in \mathbb{R}$ *and* $\delta \in [0, 1)$ *constant, and* $\{b, b'\}$ *means that either* b *or* b' *is chosen. Then* $\limsup_{m\to\infty} p_m$*,* $\liminf_{m\to\infty} p_m$*,* $\limsup_{n\to\infty} q_n$ *and* $\liminf_{n\to\infty} q_n$ *exist. This is also true in the case where there exists a finite* m, n *beyond which only* q_n *is updated.*

Proof. The proof follows from Corollary (4.2) and Corollary (4.4). □

Lemma 4.6. *From any arbitrary point in the reinforced game, both reinforced agents will almost surely eventually choose their target equilibrium actions.*

Proof. Consider the case whereby α_1 is in state $(U, [U])$ and has chosen β_{11} a total of $n_{11} \geq 0$ times in Q-state $[U]$, and β_{12} a total of $n_{12} \geq 0$ times in Q-state $[U]$, and that from this point on it chooses its actions according to an infinite sequence consisting only of β_{12}. Agent α_2 will choose either β_{21} or β_{22} probabilistically, and therefore the outcome sequence will be $(\beta_{12}, \beta_{21}$ or $\beta_{22})$, $(\beta_{12}, \beta_{21}$ or $\beta_{22})$, $(\beta_{12}, \beta_{21}$ or $\beta_{22}), \ldots$.

Since (β_{12}, β_{21}), $(\beta_{12}, \beta_{22}) \notin \eta_1$, α_1's state will not change. The Q value for β_{12} following the m^{th} selection of β_{12} in this infinite sequence will be defined according to the following stochastic recurrence relation:

$$\begin{aligned} Q_{n_{12}+m}([U], \beta_{12}) = Q_{n_{12}+m-1}([U], \beta_{12}) + \frac{1}{n_{12}+m}(&D_{n_{12}+m} \\ &+ \delta_1 \max(Q_{n_{11}}([U], \beta_{11}), Q_{n_{12}+m-1}([U], \beta_{12})) \\ &- Q_{n_{12}+m-1}([U], \beta_{12})) \end{aligned}$$

where $D_{n_{12}+m}$ is a discrete random variable yielding a reward of either U_{21} or U_{22} for α_1. By Lemma (4.5) we know that the sequence $Q_{n_{12}+m}([U], \beta_{12})$ is bounded above, and so α_1 will almost never choose β_{12} an infinite number of consecutive times:

$$\prod_{i=1}^{\infty} \frac{k_1^{Q_{n_{12}+i}([U],\beta_{12})}}{k_1^{Q_{n_{11}}([U],\beta_{11})} + k_1^{Q_{n_{12}+i}([U],\beta_{12})}} = 0$$

Therefore, from any arbitrary point in the reinforced game, α_1 will almost surely eventually choose β_{11}. A similar argument yields that α_2 will almost surely eventually choose β_{21}. □

Theorem 4.2. *If condition (*C2*) holds on the reinforced outcomes for agents α_1 and α_2, then the reinforced game will converge to the new target NE almost surely.*

Proof. Starting from the initial game in Fig. 1, agent α_1 requires m reinforcements on $NE_{target} = (\beta_{11}, \beta_{21})$ such that ${r_1}^m U_{11} > U_{21}$, and α_2 requires n reinforcements on NE_{target} such that ${r_2}^n V_{11} > V_{12}$. Therefore, at any point in the reinforced game, a lower bound on the number of outcomes (β_{11}, β_{21}) required for convergence is:

$$\max\left(\lfloor log_{r_1}(U_{21}/U_{11})\rfloor, \lfloor log_{r_2}(V_{12}/V_{11})\rfloor\right) + 1 \tag{2}$$

Consider first the reinforcement sets $\eta_1 = \eta_2 = \{(\beta_{11}, \beta_{21})\}$. We know that, at any arbitrary point in the reinforced game, α_1 will almost surely eventually choose β_{11}, and α_2 will almost surely eventually choose β_{21} (Lemma 4.6). What we need to show is that, at any arbitrary point in the reinforced game, the outcome (β_{11}, β_{21}) will almost surely eventually occur, and that it will thus occur the required finite number of times required for convergence (Equation (2)).

We proceed with a proof by contradiction. Consider the case whereby, from some arbitrary point in the reinforced game, (β_{11}, β_{21}) never occurs. Then α_1's state $(U, [U])$ will never change, and neither will α_2's state $(V, [V])$, since no reinforcements will occur. If α_1 has chosen β_{11} and β_{12} n_{11} and n_{12} times respectively, and α_2 has chosen β_{21} and β_{22} n_{21} and n_{22} times respectively, the joint probability of outcome (β_{11}, β_{21}) is:

$$\frac{k_1^{Q_{n_{11}}([U],\beta_{11})}}{k_1^{Q_{n_{11}}([U],\beta_{11})} + k_1^{Q_{n_{12}}([U],\beta_{12})}} \frac{k_2^{Q_{n_{21}}([V],\beta_{21})}}{k_2^{Q_{n_{21}}([V],\beta_{21})} + k_2^{Q_{n_{22}}([V],\beta_{22})}}$$

We know that both of the sequences $Q_{n_{11}+m}([U],\beta_{11})$ and $Q_{n_{21}+m}([U],\beta_{21})$ are bounded below, and $Q_{n_{12}+m}([U],\beta_{12})$ and $Q_{n_{22}+m}([U],\beta_{22})$ are bounded above (Lemma 4.5), and thus it follows that the joint probability of action-combination outcome (β_{11},β_{21}) is always greater than some minimum positive number. This contradicts the assumption that (β_{11},β_{21}) never occurs, since for an action-combination outcome to surely never occur its probability must always be 0.

Consider now the reinforcement sets $\eta_1 = \{(\beta_{11},\beta_{21}),(\beta_{11},\beta_{22})\}$, $\eta_2 = \{(\beta_{11},\beta_{21}),(\beta_{12},\beta_{21})\}$. We will show that, from any arbitrary point in the reinforced game, the outcome (β_{11},β_{21}) will almost surely eventually occur.

We again proceed with a proof by contradiction: consider the case whereby, from some arbitrary point in the reinforced game, (β_{11},β_{21}) never occurs. We know that α_1 will almost surely eventually choose β_{11}, and that α_2 will almost surely eventually choose β_{21} (Lemma 4.6). Due to our assumption that (β_{11},β_{21}) never occurs, these choices will not coincide but instead correspond to outcomes (β_{11},β_{22}) and (β_{12},β_{21}) respectively. Following m selections of β_{11} by α_1 (and therefore following m occurrences of (β_{11},β_{22})), where $m = \lfloor log_{r_1}(\max(U_{11},U_{21},U_{22})/U_{12})\rfloor + 1$, α_1 will be in state $(U^*,[U^*])$ in which $[U^*]_{12}$ is strictly a maximum. For β_{11} being chosen n_{11} times in Q-state $[U^*]$, β_{12} chosen n_{12} times, the Q value for β_{11} is:

$$\begin{gathered}Q_{n_{11}}([U^*],\beta_{11}) = Q_{n_{11}-1}([U^*],\beta_{11}) + \frac{1}{n_{11}}(r^{n_{11}}\ {U^*}_{12} \\ +\delta_1 \max(Q_{n_{11}-1}([U^*],\beta_{11}), Q_{n_{12}}([U^*],\beta_{12}))- \\ Q_{n_{11}-1}([U^*],\beta_{11}))\end{gathered}$$

Since $Q_{n_{11}}([U^*],\beta_{11}) \geq {U^*}_{12}\sum_{j=1}^{n_{11}}\frac{r^j}{j} \to \infty$, as $n_{11}\to\infty$, the probability of α_1 selecting β_{11} in $[U^*]$ will be 1. A similar argument yields that, over an infinite horizon, the probability of α_2 selecting β_{21} in some state $(V^*,[V^*])$ with $[V^*]_{21}$ strictly a maximum will be 1, implying that the joint probability of outcome (β_{11},β_{21}) will also be 1. This is a contradiction, since for an action-combination outcome to surely never occur it must always have probability 0, and implies that (β_{11},β_{21}) will almost surely eventually occur, thus guaranteeing convergence.

Finally, consider the case $\eta_1 = \{(\beta_{11},\beta_{21}),(\beta_{11},\beta_{22})\}$, $\eta_2 = \{(\beta_{11},\beta_{21})\}$. We again assume that, from some arbitrary point in the reinforced game, outcome (β_{11},β_{21}) never occurs. The proof follows from the previous two cases, since over an infinite horizon α_1 will come to be in state $(U^*,[U^*])$ with probability of selecting β_{11} equal to 1, and α_2 will remain in state

$(V, [V])$ with a probability of selecting β_{21} strictly greater than some minimum positive number. □

5 Simulation Results

As discussed previously, our framework can be applied to classical cooperation games such as the prisoner's dilemma and snowdrift in order to induce cooperative outcomes. Here, we apply our framework to the game in Fig. 4, which has been proposed as a model of avoidant attachment between a child and parent (ordinal type IIA [6]). Attachment theory, a dominant paradigm in psychology, outlines the need for every infant to develop an emotionally supportive, dependant relationship with a primary caregiver, and the central tenet is that the type of attachment that emerges has a significant and lasting impact on future psychological well-being. Work in computational modelling of attachments includes an agent-based simulation explaining attachment styles as adaptations to care-giving styles [30], the study of caregiver behaviour in human-robot attachment interactions [17], and work on the capacity of Hopfield strong attractors to model attachment schemata [12] [13]. Another recent study presented a neural-cognitive architecture in an attempt to explain adaptive infant attachment behaviour and physiology in terms of approach/avoid tendencies mediated by the OFC and fear circuitry in the amygdala [9].

The strange situation [2] is a protocol consisting of a series of separation and reunion episodes, designed to elicit attachment types between a child and parent; the game in Fig. 4 captures a single interaction on the final reunion episode. The parent's Ignore action dominates, and the child's stress will increase if they go for attention but are ignored by the parent. Thus, the child is always better off choosing Don't Go, and the game has an avoidant NE in pure actions at (Don't Go, Ignore).

Evidence suggests that radical shifts in the way a parent interacts with their child can result in a change in attachment style [37]; here we are particularly interested in modelling the transitional case from insecure to secure attachment. The child assumes the role of a reactive agent and the parent is a reinforced agent. The external agent can conceptually be considered as a mediator (e.g. a psychotherapist), whose aim is to encourage the emergence of a pure NE at the (Go, Attend) outcome, corresponding to the socially desirable secure form of attachment [15]. Alternatively, the external agent could be representative of a consistent, internal cognitive reappraisal process within the parent agent.

		Parent Agent	
		Attend	Ignore
Child Agent	Go	4,2	2,3
	Don't Go	3,1	**3,4**

Figure 4: An attachment game with an avoidant NE (child agent payoff is given first)

		Parent Agent	
		Attend	Ignore
Child Agent	Go	**4,4**	2,2
	Don't Go	3,1	**3,3**

Figure 5: An attachment game with both secure and avoidant NE (child agent payoff is given first)

Simulations of iterated games were run such that each iterated game consisted of 10,000 rounds, and the individual simulations were repeated independently 200 times. The average number of rounds before the emergence of the secure attachment NE (given by the game with ordinal type IIB2a [6] in Fig. 5) was calculated, where we also required the outcome to be played 5 consecutive times during play for termination. This process was repeated for various combinations of the discount factor δ, for reinforcement rates $r \in \{1.1, 1.3, 1.5\}$, for exploration parameters $k \in \{1.5, 2\}$, and for the reinforcement sets $\eta = \{(Go, Attend), (Don't\,Go, Attend)\}$ and $\eta = \{(Go, Attend)\}$. After generation of NE_{target}, we performed a reversal by reverting to the initial payoff matrix (i.e. initial M-state, Q-state pair). The game was judged to have decayed when the parent's Q value for Ignore was greater than for Attend, and when the $NE_{initial}$ outcome occurred 5 consecutive times during play.

The results are charted in Fig. 6 (for $k = 1.5$), and Fig. 7 (for $k = 2$). A solid line indicates that $\eta = \{(Go, Attend)\}$ was used, and a dashed line that the reinforcement set $\eta = \{(Go, Attend),$ $(Don't\,Go, Attend)\}$ was used. The top charts in each figure is for to the generation phase, and the bottom for the decay phase, with the horizontal axis giving the discount factor δ, and the vertical axis is the number of rounds for the generation or decay of NE_{target}.

As the size of the reinforcement r on the parent agent's payoff matrix is increased, the average number of rounds required before a stable, secure attachment style emerges decreases. As would be intuitively expected under our model, the results also show us that secure attachment relationships are more quick to emerge when $\eta = \{(Go, Attend), (Don't\,Go, Attend)\}$ (i.e. when any outcome resulting from the parent agent selecting the 'Attend' action is encouraged, represented by a dashed line) than when $\eta = \{(Go, Attend)\}$ (i.e. when only the $(Go, Attend)$ outcome is encouraged, represented by a solid line). However, we note that this effect becomes less pronounced as the size of the reinforcement r increases.

We also observe that, in general, parent agents with lower discount factors (i.e. those who prefer immediate rewards) see the emergence of a secure attachment style more quickly than those who favour future rewards, although the effect was far more pronounced at the higher exploration parameter of $k = 2$. Indeed, only 1.5% of parent agents with exploration factor $k = 2$ and a large discount factor $\delta = 0.8$ converged to a stable, secure attachment style within 10,000 rounds of play, whereas the convergence rate was 100% for $k = 1.5$. As the exploration parameter k increases, the prob-

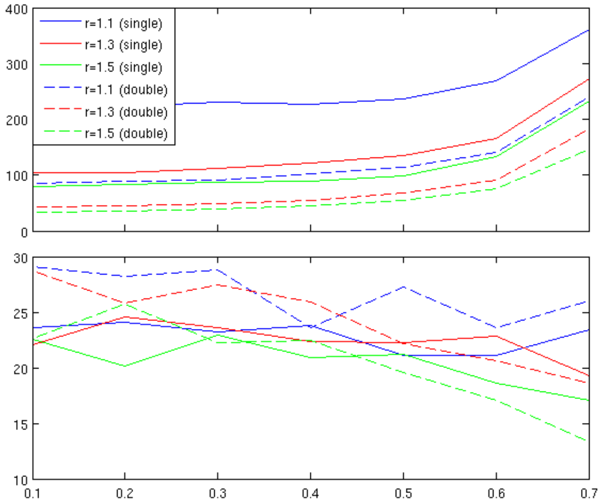


Figure 6: Average number of rounds required for a secure attachment NE to evolve (top), and average number of rounds following reversal for decay to the initial insecure NE (bottom), for $k = 1.5$. Horizontal axis is the discount factor (δ), and the vertical axis is the average number of rounds, for exploration parameter $k = 1.1$. Solid lines are results for various reinforcement rates $r \in \{1.1, 1.3, 1.5\}$ with the reinforcement set $\eta = \{(Go, Attend)\}$, and dashed lines for $\eta = \{(Go, Attend), (Don't\,Go, Attend)\}$. In all cases the child played BRTLM.

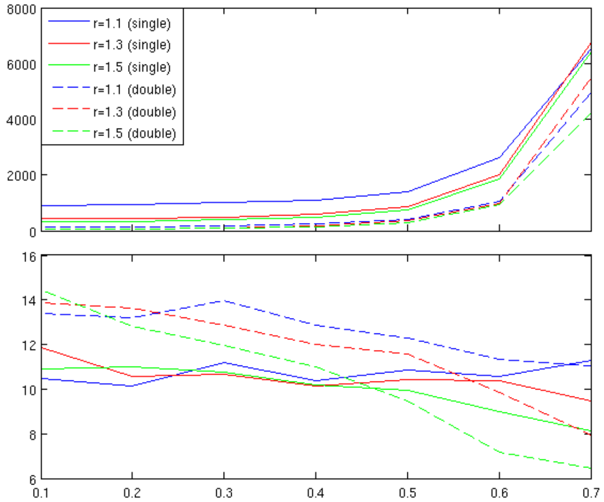


Figure 7: Average number of rounds required for a secure attachment NE to evolve (top), and average number of rounds following reversal for decay to the initial insecure NE (bottom). Horizontal axis is the discount factor (δ), and the vertical axis is the average number of rounds, for exploration parameter $k = 2$. Solid lines are results for various reinforcement rates $r \in \{1.1, 1.3, 1.5\}$ with the reinforcement set $\eta = \{(Go, Attend)\}$, and dashed lines for $\eta = \{(Go, Attend), (Don't\,Go, Attend)\}$. In all cases the child played BRTLM.

ability of selecting those actions with low Q values becomes smaller, and so one way of looking at the exploration parameter is to say that larger values of k are representative of a parent agent who has more 'embedded' behaviour. Increasing k decreases the limit of the probability of 'Attend' being chosen for the first time in any state, and it appears as though there is some threshold for δ for which the initial non-transitional Q value updates are so large that the probability of exploration becomes too small to result in a change in attachment style within some reasonable time frame. Generation was on average slower than decay for all cases, although the effect was far less pronounced for $k = 1.5$ compared to $k = 2$.

6 Conclusion

We have presented a conceptual framework for generating a new, strict NE at some predetermined pure action profile. Under the assumption that the extra reward is internally generated, we can consider changes in payoff to be the result of a (re)appraisal mechanism, and thus the new NE to persist even if the external agent ceases to exert an influence over the agents within the game. Alternatively, if we view the additional payoff as coming directly from the environment, a reversal to the original matrix will lead to a gradual decay away from the new equilibrium. We have shown how the emergence of this new NE is almost surely guaranteed for the two-agent, two-action case, both when only a single agent is reinforced, and when both agents are reinforced but learn independently without opponent modelling. We hypothesise that our new framework provides a conceptual model for how humans or agents engaged in a game can undergo learning under the influence of an external agent, who acts as a reinforcer in order to resolve conflict. Finally, we applied the framework to a game representing avoidant attachment between a parent and child, which led to an evolution in the game structure towards one with a NE representative of the socially desirable secure form of attachment.

The framework presented in this paper has been generalised to the n-agent case with almost-sure convergence still guaranteed (see Appendix). Future work should consider games with initial mixed strategy equilibria, and more than two actions per agent, plus stochastic reinforcement on the part of the external agent.

As discussed previously, under the intrinsic/extrinsic reward distinction, additional reward in our framework can be related to an internally-driven appraisal/re-appraisal process that is triggered by an additional environmental sensory cue (relating perhaps, for example, to a targeted manipulation of the control or valance dimensions of the internal appraisal mechanism in [33]). Such a process is essentially reflected to some forms of human psychotherapy. Our conceptual framework and results provide motivation for the development of internal appraisal mechanisms following its principles. Such mechanisms will provide a robust way to adapt multi-agent behaviour in a dynamic (but gradual) on-line manner, in which state representations are internally generated and dynamic in nature, without restricting or disabling intrinsic reward-appraisal mechanisms.

Although we have discussed some alternatives, we have only considered in detail a particular internal state representation consisting of a preference ranking over outcomes, and future work should consider alternative and richer state representations. Internally-generated, memory-based state representations are common in tackling partially observable environments. For example, in [29], a free energy-based reinforcement learning agent uses both perceptual input and the output from a recurrent neural network memory to form its internal state representation, and successfully uses this representation to learn an optimal policy for a partially observable environment with high-dimensional perceptual signals. Future work should consider richer state representations incorporating task-relevant perceptual and memory information.

Appendices

A Convergence Proofs For n-Agents, 2-Actions

We have a set of n agents $A = \{\alpha_1, \alpha_2, ..., \alpha_n\}$, where each agent $\alpha_i \in A$ has action set $B_i = \{\beta_{i1}, \beta_{i2}\}$. A subset of the agents $R \subseteq A$ are reinforced agents, and the remaining agents $A \setminus R$ are reactive agents. Action-combination outcomes $(\beta_{1a_1}, ..., \beta_{na_n})$ for the agents are determined by the vertices of the n-dimensional unit cube $(a_1, ..., a_n) \in {\mathbb{N}_2^+}^n$ with $a_q \in \mathbb{N}_2^+$, $q \in \mathbb{N}_n^+$ and $\mathbb{N}_n^+ = \{1, 2, ..., n\}$. The payoff for outcome $(\beta_{1a_1}, ..., \beta_{na_n})$ for α_i is given by $U_i(\beta_{1a_1}, ..., \beta_{na_n}) \in \mathbb{R}^+$.

A.1 Reinforcing a single agent

For $|R| = 1$, $|A \setminus R| = n-1 \geq 1$ (i.e. having a single reinforced agent α_i, and at least one reactive agent), then we will show that a sufficient convergence criterion is:

$$\eta_i = \{NE_{target}\} \cup \zeta_i \text{ with } \zeta_i \subseteq \{(\beta_{1a_1}, ..., \beta_{i1}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\} \text{ and} \tag{C1}$$

$$\forall \alpha_j \in A \setminus R : U_j(\beta_{11}, ..., \beta_{j1}, ..., \beta_{n1}) > U_j(\beta_{11}, ..., \beta_{j2}, ..., \beta_{n1})$$

i.e., for the reinforced agent, we reinforce the target equilibrium outcome and also optionally any outcome which results from the reinforced agent deviating from the initial NE by choosing their target NE action.

If $\exists \alpha_j \in A \setminus R : U_j(\beta_{11}, ..., \beta_{j1}, ..., \beta_{n1}) < U_j(\beta_{11}, ..., \beta_{j2}, ..., \beta_{n1})$ then by definition a new NE cannot be generated at $NE_{target} = (\beta_{11}, ..., \beta_{j1}, ..., \beta_{n1})$, since α_j's payoff matrix does not change. If $\exists \alpha_j \in A \backslash R : U_j(\beta_{11}, ..., \beta_{j1}, ..., \beta_{n1}) = U_j(\beta_{11}, ..., \beta_{j2}, ..., \beta_{n1})$ then the dynamics will depend on how α_j discriminates between outcomes with equal payoffs, although any new NE generated at NE_{target} will not be a strict NE. We will show below that convergence is almost surely guaranteed if condition (C1) holds.

Lemma A.1. *If condition (*C1*) holds for the reinforced agent α_i's reinforcement set η_i, then at any arbitrary point in the reinforced game, the reinforced agent α_i will almost never choose their actions according to the infinite, alternating sequence $(\beta_{i2}, \beta_{i1}, \beta_{i2}, ...)$.*

Proof. Suppose that α_i is in state $(U_i, [U_i])$, where β_{i1} has been chosen a total of $n_{i1} \geq 0$ times in Q-state $[U_i]$, and β_{i2} a total of $n_{i2} \geq 0$ times in Q-state $[U_i]$, such that the Q values are $Q_{n_{i1}}([U_i], \beta_{i1}) \geq 1$ for β_{i1} and $Q_{n_{i2}}([U_i], \beta_{i2}) \geq 1$ for β_{i2}. Since reactive agents $\alpha_j \in A \backslash R$ have deterministic action selection rules, we can state the outcome sequence for alternative β_{i2}, β_{i1} action choices by α_i:

$$\begin{gathered}(\beta_{11} \text{ or } \beta_{12}, ..., \beta_{i2}, ..., \beta_{n1} \text{ or } \beta_{n2}), (\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2}), \\ (\beta_{11}, ..., \beta_{i2}, ..., \beta_{n1}), (\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2}), (\beta_{11}, ..., \beta_{i2}, ..., \beta_{n1}), ...\end{gathered}$$

where the first choice of β_{j1} or β_{j2} by $\alpha_j \in A \setminus R$ depends on whether α_i's previous action selection was β_{i1} or β_{i2}. For the case where α_i's reinforcement set is $\eta_i = \{NE_{target} = (\beta_{11}, ..., \beta_{i1}, ..., \beta_{n1})\}$, no reinforcements and thus no state transitions will occur as a result of this outcome sequence. Setting $Q_p([U_i], \beta_{i1}) = a_p$ and $Q_q([U_i], \beta_{i2}) = b_q$ in order to simplify the notation, then the infinite probability product is given by:

$$\begin{aligned}&\prod_{\substack{i=n_{21}\\ j=n_{22}}}^{\infty} \frac{k^{b_j}}{k^{b_j} + k^{a_i}} \frac{k^{a_i}}{k^{b_{j+1}} + k^{a_i}} \\ &= \prod_{\substack{i=n_{21}\\ j=n_{22}}}^{\infty} \frac{1}{k^{b_{j+1}-a_i} + 1 + k^{a_i - b_j} + k^{b_{j+1}-b_j}}\end{aligned} \tag{3}$$

Since $\lim_{j\to\infty} b_j$ exists (Lemma 2 in main paper), it follows that for the sequence $c_j := b_{j+1} - b_j$, with $j \geq n_{22}$, we have $\lim_{j\to\infty} c_j = 0$. Since any convergent sequence is bounded there exists $t \geq 0$ so that $|c_j| \leq t$ for all $j \geq n_{22}$. Thus, each term in the product is less than $\frac{1}{1+k^{-t}} < 1$ and thus the infinite product is zero. As such, at any point in the reinforced game, the probability of the reinforced agent choosing its actions according to the infinite alternating sequence $(\beta_{i2}, \beta_{i1}, \beta_{i2}, ...)$ is zero.

Consider now the reinforcement sets such that $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2}) \in \zeta_i$. For the reinforced agent successively choosing $\beta_{i2}, \beta_{i1}, \beta_{i2}, ...$, then starting with the initial payoff matrix, each of the $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$ outcomes will result in a reinforcement on $U_i(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$. There is some finite number of reinforcements n on $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$ required such that α_2 will transition to some state $(U_i{}^*, [U_i{}^*])$ in which $[U_i{}^*](\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$ is strictly a maximum, i.e. state $(U_i{}^*, [U_i{}^*])$ for which:

$$\begin{gathered}\forall u \in \{[U_i{}^*](\beta_{1a_1}, ..., \beta_{ia_i}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\} \setminus \{[U_i{}^*](\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})\} : \\ [U_i{}^*](\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2}) > u\end{gathered}$$

This state is achieved when ${r_i}^n U_i(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2}) > \max(\{U_i(\beta_{1a_1}, ..., \beta_{ia_i}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\} \ \setminus \ \{U_i(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})\})$, i.e. following:

$$n = \left\lfloor log_{r_i} \left(\frac{\max(\{U_i(\beta_{1a_1}, ..., \beta_{ia_i}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\} \ \setminus \ \{U_i(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})\})}{U_i(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})} \right) \right\rfloor + 1$$

reinforcements on $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$. Indeed, at any point in the reinforced game, then for α_i successively choosing β_{i2}, β_{i1}, β_{i2}, ..., there will be some finite number of reinforcements on $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$ required in order to transition to this Q-state $[{U_i}^*]$. From $[{U_i}^*]$, any number of further reinforcements on $(\beta_{12}, ..., \beta_{i1}, ..., \beta_{n2})$ will not result in a change to the Q-state, and we can see that the infinite product in (Eq. 3) is again zero. The proof for any other reinforcement set η_i is equivalent to the case when $\eta_i = \{NE_{target}\}$, since the outcome sequence will not result in any reinforcements and thus α_i's state will not change.

□

Theorem A.1. *If condition (*C1*) holds on the payoff matrix for the reactive agents $\alpha_j \in A \backslash R$, and on the reinforced outcomes η_i for the reinforced agent α_i, then the reinforced game will converge to the new target NE almost surely.*

Proof. If condition (C1) holds, then in order for NE_{target} to emerge as a new strict NE we require n reinforcements on agent α_i's initial payoff $U_i(\beta_{11}, ..., \beta_{i1}, ..., \beta_{n1})$, such that ${r_i}^n U_i(\beta_{11}, ..., \beta_{i1}, ..., \beta_{n1}) > U_i(\beta_{11}, ..., \beta_{i2}, ..., \beta_{n1})$. Therefore it follows that we require:

$$n = \left\lfloor log_{r_i} \left(\frac{U_i(\beta_{11}, ..., \beta_{i2}, ..., \beta_{n1})}{U_i(\beta_{11}, ..., \beta_{11}, ..., \beta_{11})} \right) \right\rfloor + 1$$

reinforcements on $U_i(\beta_{11}, ..., \beta_{11}, ..., \beta_{11})$ (n provides a lower bound on the number of reinforcements on $(\beta_{11}, ..., \beta_{i1}, ..., \beta_{n1})$ required for NE_{target} to emerge from any arbitrary point in the reinforced game). Since the reactive agents $\alpha_j \in A \setminus R$ are playing the BRTLM iterated strategy, we require agent $\alpha_i \in R$ to choose action $\beta_{i1} \in NE_{target}$ two consecutive times in order for each reinforcement on NE_{target} to occur: n is therefore a lower bound on the number of consecutive choices of $\beta_{i1} \in NE_{target}$ by agent α_i required at some arbitrary point in the reinforced game as a sufficient condition for convergence. We have shown that, at any point in the reinforced game, the reinforced agent will almost never choose their actions according to the infinite sequence $(\beta_{i2}, \beta_{i1}, \beta_{i2}, ...)$. We also have that, at any point in the reinforced game, the probability of some finite number $p' > 1$ of consecutive selections of β_{i2} is clearly less than a single selection of β_{i2}, i.e.:

$$\forall p' \in \mathbb{N}_{>1}, n_{i1} \geq 0, n_{i2} \geq 0, [U_i] \in E :$$
$$\prod_{p=n_{i2}}^{p'} \frac{k^{Q_p([U_i],\beta_{i2})}}{k^{Q_{n_{i1}}([U_i],\beta_{i1})} + k^{Q_p([U_i],\beta_{i2})}} < \frac{k^{Q_{n_{i2}}([U_i],\beta_{i2})}}{k^{Q_{n_{i1}}([U_i],\beta_{i1})} + k^{Q_{n_{i2}}([U_i],\beta_{i2})}}$$

Therefore it follows that:

$$\forall p_z \in \mathbb{N}_{>1}, n_{i1} \geq 0, n_{i2} \geq 0, [U_i] \in E :$$
$$\prod_{z=1}^{\infty} \left(\prod_{p=1}^{p_z} \frac{k^{Q_{n_{i2}+p}([U_i],\beta_{i2})}}{k^{Q_{n_{i1}+z}([U_i],\beta_{i1})} + k^{Q_{n_{i2}+p}([U_i],\beta_{i2})}} \right) \cdot \frac{k^{Q_{n_{i1}+z}([U_i],\beta_{i1})}}{k^{Q_{n_{i1}+z}([U_i],\beta_{i1})} + k^{Q_{n_{i2}+p_z}([U_i],\beta_{i2})}} = 0$$

i.e., at any point in the reinforced game, the probability of the reinforced agent α_i choosing their actions according to an infinite sequence consisting of some finite number of β_{i2} selections, followed by a single β_{i1}, followed by some finite number of β_{i2} selections, etc, is 0. Thus, at any point in the reinforced game, agent α_i will almost surely (i.e with probability 1) eventually select action β_{i1} two consecutive times.

□

A.2 Reinforcing all agents

For $|R| = n > 1$ (i.e. having more than one reinforced agent and no reactive agents), then we will show that a sufficient convergence criterion is:

$$\begin{gathered} \forall \alpha_i \in A : \eta_i = \{NE_{target}\} \cup \zeta_i \text{ with} \\ (\zeta_i = \{U_i(\beta_{1a_1}, ..., \beta_{i1}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\} \text{ or } \zeta_i = \{\}) \end{gathered} \tag{C2}$$

Theorem A.2. *If condition (*C2*) holds on the reinforced outcomes for all agents α_i, then the reinforced game will converge to the new target NE almost surely.*

Proof. In order to show almost sure convergence, we need to show that, at any arbitrary point in the reinforced game, NE_{target} will almost surely eventually occur.

Consider first the reinforcement sets $\forall \alpha_i \in A : \eta_i = \{NE_{target}\}$. We consider each reinforced agent $\alpha_i \in R = A$ individually. At some arbitrary point in the game, α_i is in state $(U_i, [U_i])$ and has chosen β_{i1} a total of $n_{i1} \geq 0$ times in Q-state $[U_i]$, and β_{i2} a total of $n_{i2} \geq 0$ times in $[U_i]$ Assume that, from this point on, NE_{target} never occurs, and we proceed with a proof by contradiction. The joint probability of NE_{target} is:

$$\prod_{i=1}^{n} \frac{{k_i}^{Q_{n_{i1}}([U_i],\beta_{i1})}}{{k_i}^{Q_{n_{i1}}([U_i],\beta_{i1})} + {k_i}^{Q_{n_{i2}}([U_i],\beta_{i2})}} = \prod_{i=1}^{n} \frac{1}{1 + {k_i}^{Q_{n_{i2}}([U_i],\beta_{i2}) - Q_{n_{i1}}([U_i],\beta_{i1})}}$$

If β_{i1} is chosen at the next time-step, then its Q value will be defined according to the following stochastic recurrence relation:

$$Q_{n_{i1}+1}([U_i], \beta_{i1}) = Q_{n_{i1}}([U_i], \beta_{i1})$$
$$+\frac{1}{n_{i1}+1}\left(x_{n_{i1}+1} + \delta_i \max(Q_{n_{i1}}([U_i], \beta_{i1}), Q_{n_{i2}}([U_i], \beta_{i2})) - Q_{n_{i1}}([U_i], \beta_{i1})\right)$$

with $x_{n_{i1}+1}$ a discrete random variable for time-step $n_{i1}+1$ yielding a reward of:

$$U_i(\beta_{1a_1}, ..., \beta_{i1}, ..., \beta_{na_n}) \in X_i \setminus \{NE_{target}\}$$

where:

$$X_i = \{U_i(\beta_{1a_1}, ..., \beta_{i1}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\}$$

Alternatively, if β_{i2} is chosen at the next time-step, then its Q value will become:

$$Q_{n_{i2}+1}([U_i], \beta_{i2}) = Q_{n_{i2}}([U_i], \beta_{i2})$$
$$+\frac{1}{n_{i2}+1}\left(y_{n_{i2}+1} + \delta_i \max(Q_{n_{i1}}([U_i], \beta_{i1}), Q_{n_{i2}}([U_i], \beta_{i2})) - Q_{n_{i2}}([U_i], \beta_{i2})\right)$$

with $y_{n_{i2}+1}$ a discrete random variable yielding a reward of:

$$U_i(\beta_{1a_1}, ..., \beta_{i2}, ..., \beta_{na_n}) \in \{U_i(\beta_{1a_1}, ..., \beta_{i2}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\}$$

From Corollary (4.3) we know that the sequence $Q_{n_{i1}}([U_i], \beta_{i1})$ is bounded below, and the sequence $Q_{n_{i2}}([U_i], \beta_{i2})$ is bounded above. Thus, it follows that the joint probability of NE_{target} is always greater than some minimum positive number. This contradicts the assumption that NE_{target} never occurs, since for an outcome to surely never occur it must always have probability zero.

Consider now the reinforcement sets $\forall \alpha_i \in A : \eta_i = \{U_i(\beta_{1a_1}, ..., \beta_{i1}, ..., \beta_{na_n}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\}$. Again, we assume that α_i is in arbitrary state $(U_i, [U_i])$. Given that our reinforcement sets now reinforce all outcomes involving β_{i1} for α_i, we have that $Q_{n_{i1}}([U_i], \beta_{i1})$ is unbounded above:

$$\lim_{n_{i1} \to \infty} Q_{n_{i1}}([U_i], \beta_{i1})$$
$$= \lim_{n_{i1} \to \infty} (Q_{n_{i1}-1}([U_i], \beta_{i1})$$
$$+ \frac{1}{n_{i1}} ({r_i}^{n_{i1}} x_{n_{i1}} + \delta_i \max(Q_{n_{i1}-1}([U_i], \beta_{i1}), Q_{n_{i2}}([U_i], \beta_{i2})) - Q_{n_{i1}-1}([U_i], \beta_{i1}))) = \infty$$

since $\forall n_{i1} : Q_{n_{i1}}([U_i], \beta_{i1}) \geq x_{n_{i1}} \sum_{j=1}^{n_{i1}} \frac{{r_i}^j}{j}$ and $\sum_{j=1}^{n_{i1}} \frac{{r_i}^j}{j} \to \infty$ as $n_{i1} \to \infty$. Because the sequence $\frac{{r_i}^j}{j}$ monotonically increases in j, it follows that $\forall n_{i1} : Q_{n_{i1}}([U_i], \beta_{i1}) \geq \min(X_i) \frac{r^0}{1} \geq 1$. Thus, the joint probability of NE_{target} is again always greater than some minimum positive number.

Finally, consider the reinforcement sets:

$$\forall i \in [1, m], 1 \leq m < n : \eta_i = \{NE_{target}\} \text{ and}$$
$$\forall j \in (m, n] : \eta_j = \{U_j(\beta_{1a_1}, ..., \beta_{j1}, ..., \beta_{na_1}) | a_q \in \mathbb{N}_2^+, q \in \mathbb{N}_n^+\}$$

The proof follows trivially from the previous two cases, since for all agents we have that β_{i1} is bounded below and β_{i2} is bounded above, so that the joint probability of NE_{target} is always greater than some minimum positive number.

□

References

[1] L. Ahlfors. *Complex analysis: an introduction to the theory of analytic functions of one complex variable.* International series in pure and applied mathematics. McGraw-Hill, 1979.

[2] M. Ainsworth, M. Blehar, E. Waters, and S. Wall. *Patterns of attachment: A psychological study of the strange situation.* Lawrence Erlbaum, 1978.

[3] S. Albrecht and S. Ramamoorthy. Comparative evaluation of mal algorithms in a diverse set of ad hoc team problems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, pages 349–356, Richland, SC, 2012.

[4] I. Arel. The threat of a reward-driven adversarial artificial general intelligence. In *Singularity Hypotheses*, pages 43–60. Springer, 2012.

[5] R. Axelrod. The evolution of cooperation: revised edition. 2006.

[6] L. Buono, R. Chau, G. Lewis, N. Madras, M. Pugh, L. Rossi, and T. Witelski. Mathematical models of mother/child attachment. *Problem proposed by L. Atkinson, J. Hunter and B. Lancee at the Fields-MITACS Industrial Problem Solving Workshop August 2006.*

[7] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(2):156–172, 2008.

[8] N. Chentanez, A. G. Barto, and S. P. Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2004.

[9] D. Cittern and A. Edalat. An arousal-based neural model of infant attachment. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on*, pages 57–64, Dec 2014.

[10] S. Devlin and D. Kudenko. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 225–232. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[11] D. Dewey. Reinforcement learning and the reward engineering principle. In *2014 AAAI Spring Symposium Series*, 2014.

[12] A. Edalat. Capacity of strong attractor patterns to model behavioural and cognitive prototypes. In *Advances in Neural Information Processing Systems*, pages 2661–2669, 2013.

[13] A. Edalat and F. Mancinelli. Strong attractors of hopfield neural networks to model attachment types and behavioural patterns. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–10. IEEE, 2013.

[14] H. Farber. An analysis of final-offer arbitration. *Journal of conflict resolution*, 24(4):683–705, 1980.

[15] R. Fearon, M. Bakermans-Kranenburg, M. Van IJzendoorn, A. Lapsley, and G. Roisman. The significance of insecure attachment and disorganization in the development of childrens externalizing behavior: A meta-analytic study. *Child Development*, 81(2):435–456, 2010.

[16] F. Forgo. Measuring the power of soft correlated equilibrium in 2-facility simple non-increasing linear congestion games. *Central European Journal of Operations Research*, pages 1–17, 2012.

[17] A. Hiolle, L. Cañamero, M. Davila-Ross, and K. Bard. Eliciting caregiving behavior in dyadic human-robot attachment-like interactions. *ACM Trans. Interact. Intell. Syst.*, 2(1):3:1–3:24, Mar. 2012.

[18] R. Howard. Dynamic programming and markov processes. 1960.

[19] J. Hu and M. Wellman. Nash q-learning for general-sum stochastic games. *The Journal of Machine Learning Research*, 4:1039–1069, 2003.

[20] M. Keramati and B. Gutkin. Collecting reward to defend homeostasis: A homeostatic reinforcement learning theory. *bioRxiv*, 2014.

[21] D. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913–922, 2008.

[22] R. Kümmerli, C. Colliard, N. Fiechter, B. Petitpierre, F. Russier, and L. Keller. Human cooperation in social dilemmas: comparing the snowdrift game with the prisoner's dilemma. *Proceedings of the Royal Society B: Biological Sciences*, 274(1628):2965–2970, 2007.

[23] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, volume 94, pages 157–163, 1994.

[24] R. Luce. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215–233, 1977.

[25] D. Monderer and M. Tennenholtz. Strong mediated equilibrium. *Artificial Intelligence*, 173(1):180–195, 2009.

[26] K. Moriyama et al. Cooperation-eliciting prisoner's dilemma payoffs for reinforcement learning agents. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014.

[27] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman. Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience*, 9(8):1057–1063, 2006.

[28] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[29] M. Otsuka, J. Yoshimoto, and K. Doya. Free-energy-based reinforcement learning in a partially observable environment. In *ESANN*, 2010.

[30] D. Petters. *Designing agents to understand infants*. PhD thesis, Faculty of Science of the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY School of Computer Science, The University of Birmingham, 2006.

[31] M. Roesch, D. Calu, and G. Schoenbaum. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, 10(12):1615–1624, 2007.

[32] E. T. Rolls. *Emotion and decision-making explained*. Oxford University Press, 2013.

[33] P. Sequeira, F. S. Melo, and A. Paiva. Emotion-based intrinsic motivation for reinforcement learning agents. In *Affective Computing and Intelligent Interaction*, pages 326–336. Springer, 2011.

[34] L. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095, 1953.

[35] R. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[36] M. Tennenholtz. Game-theoretic recommendations: Some progress in an uphill battle. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '08, pages 10–16, Richland, SC, 2008.

[37] E. Waters, S. Merrick, D. Treboux, J. Crowell, and L. Albersheim. Attachment security in infancy and early adulthood: a twenty-year longitudinal study. *Child development*, 71(3):684–689, 2000.

[38] C. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[39] E. Wiewiora, G. Cottrell, and C. Elkan. Principled methods for advising reinforcement learning agents. In *ICML*, pages 792–799, 2003.

[40] F. Wörgötter and B. Porr. Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Computation*, 17(2):245–319, 2005.

[41] C. Yu, M. Zhang, and F. Ren. Emotional multiagent reinforcement learning in social dilemmas. In *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, pages 372–387. Springer, 2013.

[42] H. Zhang, D. C. Parkes, and Y. Chen. Policy teaching through reward function learning. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 295–304. ACM, 2009.