Journal of Artificial Intelligence Research 79 (2024) 895-969 Submitted 12/2023; published 03/2024

# Learning to Resolve Social Dilemmas: A Survey

**Shaheen Fatima** S.S.FATIMA@LBORO.AC.UK
**Nicholas R Jennings** N.R.JENNINGS@LBORO.AC.UK
*Loughborough University, UK*

**Michael Wooldridge** MJW@CS.OX.AC.UK
*Oxford University, UK*

## Abstract

*Social dilemmas* are situations of inter-dependent decision making in which individual rationality can lead to outcomes with poor social qualities. The ubiquity of social dilemmas in social, biological, and computational systems has generated substantial research across these diverse disciplines into the study of mechanisms for avoiding deficient outcomes by promoting and maintaining mutual cooperation. Much of this research is focused on studying how individuals faced with a dilemma can learn to cooperate by adapting their behaviours according to their past experience. In particular, three types of learning approaches have been studied: *evolutionary game-theoretic learning*, *reinforcement learning*, and *best-response learning.* This article is a comprehensive integrated survey of these learning approaches in the context of dilemma games. We formally introduce dilemma games and their inherent challenges. We then outline the three learning approaches and, for each approach, provide a survey of the solutions proposed for dilemma resolution. Finally, we provide a comparative summary and discuss directions in which further research is needed.

## 1. Introduction

The tension between individual interests and societal welfare is a fundamental problem in systems comprised of self-interested individuals. Such systems are riddled with *social dilemmas* (Van Lange et al., 2013) in which individually rational decisions lead to outcomes in which everyone is worse off than with another outcome. Dilemmas are widespread: many critical real-world problems such as multi-national conservation of natural resources, overfishing and overgrazing of common property, and deforestation represent social dilemmas (Colyvan et al., 2011). Dilemmas also arise in numerous other contexts such as traffic networks (Bonnefon et al., 2016), public health management (Bauch & Earn, 2004), organizational citizenship behaviours (Fu et al., 2011), Internet congestion (Huberman & Lukose, 1997), cybersecurity analysis (Schoenherr & Thomson, 2020), and pricing algorithms (Calvano et al., 2020).

The ubiquity of social dilemma problems creates the need to understand the factors that promote cooperative behaviour (i.e., behaviour directed toward socially desirable outcomes), and those that inhibit it, so that dilemmas can be resolved by promoting and sustaining cooperation, thereby avoiding outcomes of poor social quality. This understanding is crucial not only for human societies, but also for distributed AI systems – our focus in this article is on the latter. Further, when the individuals in the system are organised as a network, dilemma situations are also useful for understanding how cooperative and selfish behaviours spread in networks. This knowledge is again greatly useful in the design of both human and artificial societies. Besides, a better understanding of dilemmas is also necessary for addressing the issue of *tacit collusion* (Calvano et al., 2019) between pricing algorithms.

Given its significance and its inherent complexity, the problem of social dilemmas has generated significant research into its resolution. This research is spread across a wide range of scientific disciplines: computer science (Rogers et al., 2007; Han et al., 2013; Leibo et al., 2017; Peysakhovich & Lerer, 2018b; Noordman & Vreeswijk, 2019; Zhang et al., 2022, 2022), physical science (Szabó & Fath, 2007; Perc & Szolnoki, 2008; Droz et al., 2009; Matsuzawa et al., 2016; Huang et al., 2018; Zhang et al., 2019; Jusup et al., 2022), biological science (Masuda & Ohtsuki, 2009; Bitsch et al., 2018; Eimontaite et al., 2019; Mantas et al., 2022; Fu et al., 2010; Ifti et al., 2004; Killingback et al., 2010), and social science (Rapoport & Mowshowitz, 1966; Smale, 1980; Rubinstein, 1986; Selten & Stoecker, 1986; Raub, 1988; Kollock, 1998; Lopez et al., 2022). A primary objective of this research is to understand what factors drive individual behaviours toward socially desirable outcomes, and what are the ways in which such behaviours can be established and sustained in a society. In all of the literature on this topic, consideration is given to the repeated play of a dilemma game. The repeated play of a game creates opportunity for the participating agents to learn and adapt on the basis of their experience in previous episodes (Erev & Roth, 2007; Shoham et al., 2007; Brafman & Tennenholtz, 2004, 2002; Fudenberg & Kreps, 1993). However, effective adaptation is a challenge: it requires agents to avoid those actions that incite retaliation as well as exploitation, and choose actions so as to shape other's behaviours toward mutually beneficial outcomes. It is important to address this challenge in order to build AI systems in which the agents learn to autonomously resolve dilemmas.

The means by which an agent learns to play a game depends on the agent's degree of rationality (Luce & Raiffa, 1989). Variations in the degree of rationality have led to a host of learning models which can broadly be divided into three categories: *evolutionary learning* (a population-level learning analogous to evolutionary inheritance), *reinforcement learning* (an individual-level trial-and-error type of behavioural learning), and *best-reply learning* (an individual-level epistemic learning in which each agent anticipates the future action of others based on their observation of past plays, and uses this information to decide what to do). These three primary forms of learning have been combined in various ways resulting in hybrid learning models.

In this article, we survey the literature on these three main learning-theoretic approaches[1] for dilemma resolution. Learning in games is a complex phenomenon that depends on the interplay between a multitude of factors: on the parameters of the learning model, but also on the parameters of the dilemma game itself. Over half a century of research has gone into understanding this phenomenon since the first investigations by Flood (1958) into the applicability of game theoretical models to human learning in games. In this long history, there have been numerous studies with each choosing a small subset of parameters and focusing attention only on those parameters. These different individual choices have often led to apparently contradictory results, prompting further investigations and resulting in new insights. However, these insights are limited to narrow instances of the problem and our understanding of the general dilemma resolution problem remains incomplete.

[1] We will, as needed, make references to relevant experimental literature concerning dilemma games played by human subjects, but our primary focus is on the mentioned learning approaches.

### 1.1 Significance of This Survey

The existing literature on learning in dilemmas is vast and spread across many diverse disciplines. This article provides the first comprehensive and integrated survey on learning in dilemma games by bringing the fragmented results together and makes the following contributions:

1. Overviews the various learning methods and identifies, for each learning method, the key parameters.
2. For each learning method, provides a taxonomized description of dilemma solutions.
3. Enlists the cooperation indices proposed for the prediction of cooperation in dilemmas, provides an abstraction of the key aspects that underlie these indices, and gives a comparative summary of their performances.
4. Identifies gaps in the existing research and highlights future challenges for the field.

In particular, this article addresses the following research questions:

**RQ1:** How has the existing literature modelled learning in dilemma games?
This question is addressed in Sections 3.1 to 3.3 for evolutionary learning, in Sections 4.1 to 4.4 for reinforcement learning, and in Sections 5.1 to 5.5 for best-reply learning.

**RQ2:** What mechanisms has research on learning found to enhance cooperation in dilemmas?
This question is addressed in Section 3.4 for evolutionary learning, in Section 4.5 for reinforcement learning, and in Section 5.6 for best-reply learning.

**RQ3:** What indices have been proposed to predict cooperation in a dilemma in terms of the payoff matrix for the dilemma and how do their performances compare?
This question is addressed in Section 6.

**RQ4:** What are the main observations that result from this survey, and what are the key problems that are still open for future research?
This question is addressed in Section 7.

### 1.2 Related Surveys

The existing reviews on social dilemmas have focused on their sociological aspects (Kollock, 1998), their psychological aspects (Dawes, 1980; Van Lange et al., 2013) aspects, or their inherent challenges (Perc et al., 2019). In contrast, we take an AI perspective (Dafoe et al., 2021) and focus specifically on the three main learning approaches that existing literature has found to resolve dilemmas. The review (Perc et al., 2013) is targeted only at evolutionary learning. Dal Bó and Fréchette (2018) reviewed the experimental literature on learning in infinitely repeated Prisoner's Dilemma games. This is a very focused survey pertaining specifically to experimental literature while we provide a broad survey of the different learning approaches for dilemma resolution. (Gotts et al., 2003) reviewed the literature on agent-based simulation of social dilemmas focusing mainly on evolutionary approach. However, they did not consider details pertaining to the variety of different

| | | Col | |
|---|---|---|---|
| | | $C$ | $D$ |
| Row | $C$ | $R, R$ | $S, T$ |
| | $D$ | $T, S$ | $P, P$ |

Table 1: General payoff matrix for a two-agent social dilemma game. The first (second) entry in each pair is Row's (Col's) payoff.

learning approaches. Further, this article is over two decades old and much new research has been published since then. There are many surveys that focused specifically on a chosen approach for learning in multi-agent systems (for example, (Bloembergen et al., 2015; Adami et al., 2016; Newton, 2018) on evolutionary games, (Nowak et al., 2010; Shakarian et al., 2012; Allen & Nowak, 2014; Díaz & Mitsche, 2021) on evolutionary games on graphs, (Szabó & Fath, 2007; Busoniu et al., 2008; Nguyen et al., 2020; Da Silva & Costa, 2019; Hernandez-Leal et al., 2019; Gronauer & Diepold, 2022; Matsuo et al., 2022; Khetarpal et al., 2022) on reinforcement learning, (Hernandez-Leal et al., 2017) on reinforcement learning and opponent modeling). These surveys focus on learning per-se rather than on their use in dilemmas. The key challenge that arises in multi-agent learning, not necessarily involving dilemmas, is the non-stationarity of the environment. But this challenge is compounded when the environment is an inherently complex social dilemma. Learning mechanisms that are effective for other environments may not solve the dilemma problem. Therefore our aim is to survey multi-agent learning methods focusing specifically on dilemmas.

The rest of the paper is organised as follows. Section 2 introduces the structure of social dilemma games and motivates the need for introducing learning in dilemma games. Sections 3, 4, and 5 provide a survey of the evolutionary, reinforcement, and best-reply learning models respectively. Section 6 is a synopsis of cooperation indices proposed for predicting cooperation in dilemmas played by humans in laboratory conditions. Section 7 provides a summary and highlights avenues for further research.

## 2. Social Dilemmas

A game (Von Neumann & Morgenstern, 1947) is a system of payoffs that depend on the combination of choices made by the players. A dilemma is a game that satisfies certain constraints. In a typical $2 \times 2$ dilemma, each player makes one of two choices (see Table 1): cooperate (denoted $C$) or defect ($D$). These two choices lead to four possible outcomes, each with an associated payoff. $R$ (reward) and $P$ (punishment) are the payoffs for mutual cooperation and defection respectively, whereas $S$ (sucker) and $T$ (temptation) are the payoffs for cooperation by one player and defection by the other.

A dilemma has been defined in different ways. Initially, Dawes (1974, 1975, 1980) defined a *social dilemma* as a game in which

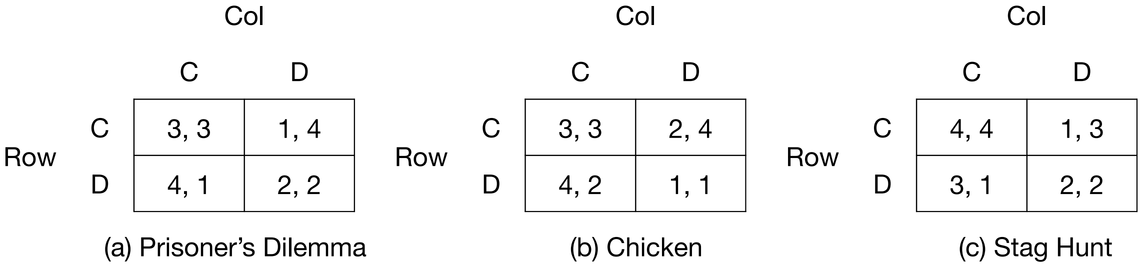
1. each player has available a dominant strategy, i.e., a strategy that is better regardless of the opponent's choice, and

2. the collective choice of dominant strategies results in a Pareto deficient outcome.

As per this definition, a dilemma is a game in which a Pareto deficient outcome is the only individually rational course of action, as in the well-known Prisoner's Dilemma (PD) (Chammah, 1965). Subsequently, Liebrand (1983) proposed a less restrictive definition; he did not consider Dawe's dominant strategy requirement as crucial for considering a situation a dilemma. Unlike Dawes's definition, Liebrand's criterion does not exclude the possibility that rational behaviour may also result in Pareto-efficient outcomes. Much of the literature has adopted Liebrand's definition of a dilemma, and this also what we do in this article. For $2 \times 2$ games, this definition is given considering that each player has a strict preference ordering over the four possible outcomes. The definition uses the notion of a *most-threatening* strategy. A strategy is called most-threatening if a rational player *Row* prefers player *Col* not to choose that strategy irrespective of *Row*'s choice. In such a case, *Col* has a most-threatening strategy. It is required that, for both players, a strategy, say $D$, be most-threatening. That is, each player has a choice between $C$ and the most-threatening strategy $D$. Then a dilemma is characterised by the following properties Liebrand (1983):

1. $C$ is not a dominant strategy for any player, and

2. both players are better off if both choose $C$ than if both choose $D$.



| | | Col | |
|---|---|---|---|
| | | C | D |
| Row | C | 3, 3 | 1, 4 |
| | D | 4, 1 | 2, 2 |

(a) Prisoner's Dilemma

| | | Col | |
|---|---|---|---|
| | | C | D |
| Row | C | 3, 3 | 2, 4 |
| | D | 4, 2 | 1, 1 |

(b) Chicken

| | | Col | |
|---|---|---|---|
| | | C | D |
| Row | C | 4, 4 | 1, 3 |
| | D | 3, 1 | 2, 2 |

(c) Stag Hunt

Figure 1: Example dilemmas

For the games shown in Figure 1, alternative $D$ is each player's most threatening strategy. Further, $C$ is not a dominant strategy for any player, and both players are better off if both choose $C$ than if both choose $D$. Thus, each game shown in Figure 1 is a dilemma as per the definition. In general, there are many games that satisfy the above listed properties.

**Definition 2.1.** A two-person social dilemma is any ordering of the payoffs $P$, $R$, $S$, and $T$ that satisfies the properties listed above.

Between all possible $2 \times 2$ symmetric games with a strict preference ordering for each player, the dilemma properties are satisfied by only three classes of games: $T > R > S > P$, $R > T > P > S$, and $T > R > P > S$. The following are some well-known examples of these three classes of games:

- $T > R > S > P$: The Chicken game (Rapoport & Chammah, 1966; Kahn, 2017) is an example of this class. For this game, Cooperate and Chicken are synonymous, and Defect and Tough are synonymous. This game is also known by other names such as the Hawk-Dove game and the Snowdrift (SD) game (Sigmund, 2010).

| | | Col | |
|---|---|---|---|
| | | $C$ | $D$ |
| Row | $C$ | $b-c, b-c$ | $-c, b$ |
| | $D$ | $b, -c$ | $0, 0$ |

Table 2: Payoff matrix for a Donation Game.

- $R > T \geq P > S$: The Stag Hunt game (for this game, Cooperate and Stag are synonymous, and Defect and Hare are synonymous), also known as *assurance game* or *common interest game*, is an example of this class (Skyrms, 2004).
- $T > R > P > S$: The Prisoner's Dilemma game and the Donations game (Chammah, 1965; Sigmund, 2010; Axelrod, 1990) are some examples of this class.

These orderings create tension between individual and collective interests. The tension is apparent when the combination of individually preferred choices result in the outcome that both players would prefer to avoid: mutual defection. This outcome is Pareto deficient in all social dilemmas as there is always the mutual cooperation outcome that is preferred by everyone since $R > P$. The problem is that mutual cooperation is Pareto optimal, yet may be undermined by the temptation to cheat (if $T > R$), or the fear of being cheated ($P > S$), or both. In the Chicken game (Rapoport & Chammah, 1966), the problem is greed but not fear. In the Stag Hunt (Belloc et al., 2019), the problem is fear but not greed. In the Prisoner's Dilemma (Chammah, 1965), there is both fear and greed making it arguably the most challenging problem, which is perhaps the reason why a major share of the literature is devoted to PDs[2]. To make the PD game less complicated by avoiding the possibility of multiple forms of tacit collusion[3], Chammah (1965) introduced the additional constraint $2R > S + T$.

Often, a specific form of Prisoner's Dilemma called the *donation game* is studied in the literature. In a donation game, the payoff matrix is given in terms of only two parameters: a *cost* $c$ and *benefit* $b$. In a two-player donation game, each player can either Cooperate or Defect. A cooperator incurs a cost $c$ in order to give a benefit of $b$ (where $b > c$) to their co-player. A defector does not incur any cost. The payoff matrix for this game (Sigmund, 2010) is as shown in Table 2.

In this article, we will focus on two-person[4] dilemma games. This is because the literature on learning in two-person dilemmas is still growing and is much more than the literature

[2] The PD has a long history and has been much studied for the light it may shed on the evolution of altruistic or cooperative behaviour (Axelrod, 1997; Burguillo-Rial, 2009; Salazar et al., 2011; Santos et al., 2019, 2020; Rodriguez-Soto et al., 2020).

[3] Rapoport and Chammah (Chammah, 1965) note two forms of tacit collusions: the tacit agreement to play CC in a single-shot PD game, and the alternation between CD and DC in repeated plays of the game. The constraint $2R > S + T$ dictates that players prefer mutual cooperation over an equal probability of unilateral cooperation and defection. Under this constraint, alternating between cooperation and defection cannot be more profitable than joint cooperation, so the possibility for players to collude in repeated plays of the game is eliminated.

[4] Note that although we focus on two-player interactions, we nevertheless consider pairwise interactions between the agents that comprise a large population.

on $n$-person dilemmas. One reason for the continued interest in two-person dilemmas is that the conditions for sustaining cooperative behaviour in a game depend on the number of players playing the game, and cooperative behaviour that can be sustained in a two-person case may break down in a large group, as illustrated by Myerson (1997) for a repeated Prisoner's Dilemma game. Another reason could be that it is easier to analyze and implement learning in two-player settings; in a two-player repeated dilemma game, it is possible for an individual to try to shape the other's behaviour by choosing their own actions suitably. For example, one could reward (punish) the other's previous choice by choosing to cooperate (defect). But when there are multiple players, the effect of a single player's actions may be less discernible.

Given a dilemma game, it can be subject to classical game theoretic analysis. In classical game theory (Von Neumann & Morgenstern, 1947), the solution to a game is given in terms of some notion of equilibrium (such as a dominant strategy or Nash) hypothesized to result from reasoning and introspection by the players who have common knowledge about the rules of the game, the rationality of the players, and the players' payoffs. What behaviour is rational crucially depends on the number of times the game is played.

For one-shot play, dominant strategy equilibrium predicts mutual defection in the PD game. Nash equilibrium predicts unilateral defection in the Chicken game, and either mutual cooperation or mutual defection in the Stag Hunt game. Repeated plays of a dilemma game give rise to multiple Nash equilibria in the super-game. When there are multiple equilibria, there is a loss of predictive power; the question of exactly which equilibrium arises is not addressed by classical game theory. Implicit in equilibrium notions is the assumption that the players somehow figure out what equilibrium to play. But how can all the players expect the same equilibrium? If expectations are not coordinated, play need not correspond to any equilibrium at all. Although some coordination procedures have been proposed (Harsanyi & Selten, 1988), how such a procedure becomes common knowledge is left unexplained. Further, there is no way to know the equilibrium dynamics; how play can arrive at a certain equilibrium, or how it might be possible to move between equilibrium points.

Another limitation of classical game theory is that there is an underlying assumption about the players' cognitive capabilities; players having well-defined preferences and making rational choices consistent with those preferences. Such forward-looking calculative rationality may not be possible for bounded rational players. There is also the common knowledge assumption which when relaxed results in much weaker conclusions (Börgers, 1994; Dekel & Fudenberg, 1990).

Yet another concern is that game theoretic prediction does not always match with behaviours observed in the real-world (Henrich et al., 2001). While theory predicts defection[5], there are examples of people cooperating in dilemma situations. This concern led to the development of new notions of game-theoretic equilibrium (Capraro et al., 2013) for certain social dilemmas and, although shown to be statistically precise at predicting human behaviour, are still based on complete information and rationality assumptions.

The aforementioned limitations led to a shift in attention toward *boundedly rational* learning models. Learning is any change in behaviour owing to experience (Bush &

[5] For the finitely repeated PD game with complete information, theory predicts defection at each stage, although cooperation can be explained under incomplete information (Basu, 1977; Kreps et al., 1982; Andreoni & Miller, 1993).

Mosteller, 1955). When a game is played repeatedly, it is possible for the individual players to gain experience and thereby learn and adapt. The dynamics that arises when all the players learn is then studied (Tuyls & Stone, 2017). Attention is focused on the long term behaviour of the players and how this relates to various game-theoretic equilibrium concepts. Learning models of bounded rationality can broadly be divided into two categories: those that treat learning as a population phenomenon derived from the genetic inheritance of individual member strategies, and those that treat learning as an individual cognitive process. *Evolutionary game theory* belongs to the former category, while *reinforcement learning* and *best-reply learning* belong to the latter category. The following three subsections describe how these three learning approaches have been utilised in the context of dilemma games.

## 3. Learning in Evolutionary Games

The limitations of classical game theory led to its expansion to *evolutionary game theory*[6] (EGT). Smith and Price (Smith & Price, 1973, 1973; Smith, 1982) founded EGT on Darwin's (Darwin, 1871, 1909; Wright, 1929) and Fisher's (1930) theory of natural selection, in an attempt to study the behaviour of large populations of agents who repeatedly engage in strategic interactions, that is, interactions in which each agent's outcome depends on his own choice but also on the distribution of others' choices. The goal was to uncover the crucial link between micro behaviour of individuals and the aggregate behaviour of the population. However, they did not attribute rationality to the individuals, rather they maintained that *natural selection* weeded out poorly adapted individuals. They defined evolutionary dynamics in terms of two fundamental mechanisms: *natural selection* and *mutation*. Since then, attention has been focused on another dynamic model known as the replicator dynamics. Section 3.1 is an overview of the evolutionary game model and Section 3.2 that of replicator dynamics.

### 3.1 Evolutionary Learning Model: Evolutionary Games

EGT (Nowak, 2006a; Sandholm, 2010; Phelps & Wooldridge, 2013; Tanimoto, 2015) provides a fundamentally different view on strategy selection to that proposed by classical game theory (Colman, 1982; Tuyls & Parsons, 2007). An evolutionary game is a population game in which each individual has the means to replicate by making copies of itself. The individuals play games against each other. Let $N$ be the set of all individuals in the population. No individual overtly reasons or makes explicit decisions about which strategy to play; a strategy is simply an individual's genetically determined behaviour.

In an evolutionary game, there is a set $N$ of individuals. The agents in $N$ *interact* in a pairwise fashion: each agent plays a game $G$ against its co-player. Between all strategies that are possible for $G$, an individual plays a certain strategy with different individuals using potentially many different strategies. Each interaction results in a certain payoff to each participating individual. An individual's *fitness*[7] depends on how it's strategy *interacts* with the other strategies in the population and is typically measured in terms of the *cumulative*

[6] See (Weibull, 1997; Vincent & Brown, 2005) for a description of the biological foundations of the theory and the resulting concepts from an economic perspective.

[7] Analogous to payoff function in a classical game, there is a fitness function in an evolutionary game.

*payoff* that results to the individual from all their interactions. An individual's fitness determines their offspring; the fitter the individual, the more numerous their offspring.

In more detail, each individual in the population plays a strategy from a given set $S = \{s_1, \ldots, s_n\}$ of strategies. An individual playing strategy $s_i$ is said to be of type $i$. In a population, different strategies occur with different frequencies. An individual's fitness depends on its type and also on the relative frequencies of the other types in the population. Two abstract evolutionary forces, *natural selection* and *mutation*, act on the population. The players in the population generally inherit strategies through selection and occasionally acquire a novel strategy as a mutation. These two forces account for the evolutionary change of individuals in the population through the generations.

Selection, the primary evolutionary force, favors the fitter individuals over others for *replication* and brings about changes to the relative frequencies of the strategies in $S$ from generation to generation. Depending on their fitness, some strategies can thus persist in the population while others get eliminated in the long run. Natural selection also serves as an agent of optimization; it favors traits that lead to individuals behaving as if they were maximizing their evolutionary fitness. It may not be that the individuals are consciously trying to maximize their fitness, just that natural selection will lead to individuals that appear to be intending to do this.

The interactions, together with the evolutionary forces, shape the composition of the population in any generation, with the composition changing from generation to generation. The resulting dynamics and the equilibria which can arise become the object of study.

The game $G$ (referred to as the *inner game* or the *stage game*) could be any classical game such as a Prisoner's Dilemma or a Stag Hunt. Each individual $i \in N$ interacts in a pair-wise fashion with each one of its co-player $j \in C_i \subseteq N - \{i\}$ by playing $G$ with them. Each one of these interactions results in a certain payoff to each player. Then $i$'s fitness is given by the payoff he accumulated in all the $C_i$ interactions.

The dynamics of an evolutionary game is given by an *outer game* (also referred to as *supergame*). The outer game describes how the strategy frequencies change during the process of evolution, i.e., the manner in which strategies spread within the population via inheritance and fitness. In more detail, the stage game $G$ is played over a series of discrete time periods. In each time period $t = 1, 2, \cdots$, the individuals of a population are randomly paired to play the stage game $G$ once. If the proportion of individuals of type $j$ is $p_j$ at a particular time, the state of the population is $\sigma = (p_1, \cdots, p_n)$ where each $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. Let $\pi_{ij}$ denote the payoff that results to an individual of type $i$ from its interaction with a co-player of type $j$. The payoff to a player of type $i$ when the state of the population is $\sigma$ is then given by

$$\pi_{i\sigma} = \sum_{j=1}^{n} \pi_{ij}\ p_j, \tag{1}$$

which is the player's expected payoff before being assigned a particular partner. An individual's fitness is a function of their payoff in the game with their co-player.

Once the stage game is played by the individuals, an evolutionary process of reproduction is initiated which results in a change to $\sigma$, the strategy frequencies. The evolutionary process is modelled by a stochastic process such as a Moran process (Moran, 1962). The outer game of an evolutionary game specifies how individuals are chosen for reproduction as well as how the birth and death of individuals comes about.

Given a stage game and an outer game, the dynamics of strategy frequency becomes the object of study. A stable solution to an evolutionary game is given using the notion of *evolutionarily stable strategy* (ESS) (Smith, 1982). An ESS is a Nash equilibrium satisfying an additional stability property. This stability property is interpreted as ensuring that if an ESS is established in a population, and if a small proportion of the population adopts some mutant behaviour, then the process of selection arising out of differing rates of reproduction will eliminate the mutant. Once an ESS becomes established in a population, it should therefore be able to withstand the pressures of mutation and selection.

### 3.2 Evolutionary Learning Model: Replicator Dynamics

In an evolutionary game, there are two basic elements: *a selection mechanism* and a *mutation mechanism* (recall that an ESS is a strategy that is resistant to small mutations). In contrast, the *replicator dynamics* is an evolutionary model that is focused only on the dynamics of natural selection. The replicator dynamic was first proposed by Taylor and Jonker (Taylor & Jonker, 1978) to model the dynamics of an evolutionary game. The growth rates of the individual strategies, i.e., the replicators, are proportional to their fitnesses (where the fitness of an individual is a function of its payoff in the stage game) and are given by the *replicator equation* (RE) (Samuelson, 1997) which is a system of differential equations describing how the relative frequencies of strategies in a population change over time $t$ as a consequence of selection. In more detail, the RE is given as follows. Consider an evolutionary game with $n$ strategies $s_1 \cdots s_n$. Let the payoffs for the stage game be given by an $n \times n$ matrix whose entries, $a_{ij}$, denote the payoff for strategy $s_i$ versus strategy $s_j$. If the relative frequency of strategy $s_i$ is given by $x_i$ where $\sum_{i=1}^{n} x_i = 1$, then the fitness of strategy $s_i$ is given by $f_i = \sum_{j=1}^{n} x_j a_{ij}$ and the average fitness of the population by $\phi = \sum_{i=1}^{n} x_i f_i$. The fundamental replicator equation is given by

$$\dot{x}_i = x_i(f_i - \phi) \quad \text{for} \quad 1 \leq i \leq n. \tag{2}$$

The dot in $\dot{x}$ indicates the derivative of $x$ with respect to $t$, i.e., $\dot{x} = dx/dt$. Equation 2 describes the evolutionary game dynamics (frequency dependent selection) in the deterministic limit of an infinitely large, well-mixed population. The fundamental RE given by Equation 2 describes pure selection dynamics, mutation is not considered.

When used for the study of dilemma games, the basic RE has been generalised in various ways: by considering populations with varying interaction rates (Taylor & Nowak, 2006), and by considering structured populations (Ohtsuki & Nowak, 2006b).

### 3.3 Evolutionary Learning: Key Parameters

The population dynamics in an evolutionary game depends on the game parameters. These are summarized in Table 3. Three types of parameters may be distinguished: the exogenous stage game parameters, the parameters of the outer game, and the parameters pertaining to the individuals in the population:

1. **The type of stage game**: The stage game may be *deterministic* (Hauert et al., 2002; Santos et al., 2006b; Wang et al., 2015; Han et al., 2017; Perc et al., 2017)

| **Stage game (SG) parameters** | | |
|---|---|---|
| Payoff matrix | Deterministic or non-deterministic | |
| | Symmetric or asymmetric | |
| **Outer game parameters** | | |
| Population size | Finite or infinite | |
| Population type | Unstructured | Random |
| | | Well-mixed |
| | Structured | Lattice |
| | | Graph |
| Boundary conditions | Constant or Periodic | |
| Partner links | Static/dynamic | |
| | Exogenous/endogenous | |
| Partner restructuring | Synchronous/asynchronous | |
| Evolutionary dynamics (strategy update) | Moran process | Birth-death update |
| | | Death-birth update |
| | | Imitation update |
| | Wright-Fisher process | |
| | Fermi update | |
| Type of evolution | Evolution of stage game strategies alone | |
| | Co-evolution of SG strategies and graph topology | |
| **Individual player parameters** | | |
| Cognitive traits | Memory bound | |
| | Ability to recognize other individuals | |
| Strategic traits | Aspiration level | |
| | Ability to choose co-players | |
| Psychological traits | Emotion, mood | |

Table 3: Evolutionary games: A summary of key parameters

where matrix entries are deterministic variables, or *random* (Fudenberg & Harris, 1992; Galla & Farmer, 2013; Gross et al., 2009; Duong et al., 2020). Most of the literature considered symmetric payoff matrix.

2. **Population size**: The number of individuals in the population may be finite (Taylor et al., 2004) or infinite (Gokhale & Traulsen, 2010).

3. **Population type**: The population may be *unstructured* or *structured.* In an unstructured population, any two random individuals can be chosen as co-players for the stage game. A special class of unstructured populations is the *well-mixed* population in which any two individuals interact with the same probability. It is a *mean-field approximation* (Tembine et al., 2012) of a general population structure. In a structured population, the individuals occupy the vertices of a spatial lattice or a graph (Allen & Nowak, 2014). Interactions between two individuals are allowed if they are connected by an edge. Structured interactions are modeled using evolutionary graph

theory (Lieberman et al., 2005; May, 2006; Casasnovas, 2012). Evolutionary games have a long history of being studied on lattices (Nowak & May, 1992; Killingback & Doebeli, 1996; Nakamaru et al., 1998; Van Baalen & Rand, 1998; Irwin & Taylor, 2001; Szabó & Hauert, 2002; Hauert & Doebeli, 2004; Ifti et al., 2004; Nakamaru & Iwasa, 2005; Jansen & Van Baalen, 2006; Traulsen et al., 2010), and more recently, on graphs (Nowak & May, 1992; Lieberman et al., 2005; Santos & Pacheco, 2005; Santos et al., 2005; Ohtsuki et al., 2006; Ohtsuki & Nowak, 2006a, 2006b; Santos et al., 2006c, 2006b; Ohtsuki et al., 2006).

Within this broad framework of structured populations, there are many possible ways of organising the underlying graph. These include the celebrated Watts-Strogatz (Watts & Strogatz, 1998) small-world (SW) networks and scale-free (SF) networks (Barabási & Albert, 1999). Apart from this, random graphs, regular graphs (each node is connected to $k$ neighbours), lattice graphs (each interior node is connected to $k$ neighbours), and cyclic graphs (each node is linked to two neighbours) have also been studied (Ohtsuki et al., 2006). Further, as in (Sun et al., 2018), each link in a graph may be weighted by the strength of the relationship between the two nodes.

While most of the evolutionary games literature has considered lattice or graph structured populations, some works (Tarnita et al., 2009; Nowak et al., 2010) have used sets to structure populations. In evolutionary set theory, the individuals of a population are distributed over sets. An individual can belong to several sets. Whether an individual cooperates with another depends on how many sets they have in common. The more the number of common sets between any two individuals, the greater the rate of interaction between them. In a set-structured population, attention is focused on investigating set membership conditions for cooperation to evolve in a population.

4. **Type of boundary condition**: For a structured population, boundary conditions (Kim et al., 2002) are needed if each individual is required to interact with the same number of co-players. For example, a lattice gives rise to two types of boundary conditions: *constant boundary condition* and *periodic boundary condition.* For the former, all individuals on the lattice interact only with each of their connected neighbor, but the boundary players ($0, \cdots, 9$ and $a, \cdots, f$ in the example of Fig 2) are excluded from calculations. For periodic boundary condition, all individuals on the lattice interact with each of their connected neighbor, and, in addition, the boundary players on a lattice edge interact with the corresponding row/column individual on the opposite edge (in Fig 2, individual 0's co-players will be its two connected neighbors and individuals 4 and 5).

5. **Partnership links**: For structured populations, the links between nodes may be set *exogenously* or *endogenously* by the individual players. Further, links may either remain *static* throughout the evolutionary process or may be *dynamically* created and severed during evolution (Pacheco et al., 2006). In a dynamic model, members of a population choose their partners, i.e., who they interact with and for how long. The key idea is that partner choices will be made such that mutually beneficial interactions endure longer than interactions in which one party exploits the other. Partnerships can form in various ways. In some treatments (Eshel & Cavalli-Sforza, 1982; Noë &

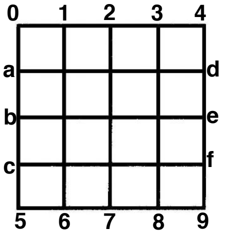


Figure 2: Boundary conditions

Hammerstein, 1994; Biely et al., 2007; Bala & Goyal, 2001; Ebel & Bornholdt, 2002; Eguíluz et al., 2005; Skyrms & Pemantle, 2009; Iyer & Killingback, 2020) individuals meet assortatively, by means of a selective partner choice[8]. In other treatments (Peck & Feldman, 1986; Hauert et al., 2002; Hauert & Szabo, 2003; Aktipis, 2004; Szabó & Hauert, 2002), individuals meet by means of volunteering participation. Another possibility is to allow individuals to *migrate* to empty sites of a network (Vainstein et al., 2007). For the case of dynamic networks, a key aspect is the timescale for link updates. Suppose links are updated every $\tau_L$ time units and that the timescale for evolutionary dynamics is $\tau_E$. The two timescales may be same (*synchronous*) or different (*asynchronous*), and the relation between them impacts on the evolutionary dynamics (Pacheco et al., 2006).

6. **The type of evolutionary dynamics**: Strategies are updated by a stochastic process. The main forms of stochastic processes studied are:

   (a) Moran process (Nowak et al., 2004; Taylor et al., 2004): At each time step, a random individual is chosen for reproduction proportional to its fitness; the offspring replaces a randomly chosen neighbor. The population size remains constant throughout the evolutionary process. The following are ways of implementing a Moran process:

      i. Birth-Death (BD) update rule (Ohtsuki & Nowak, 2006b): An individual is chosen from the population with a probability proportional to their fitness. Then the strategy of the chosen node replaces the strategy of one of its randomly chosen neighbours.

      ii. Death-Birth (DB) update rule (Ohtsuki & Nowak, 2006b): A random individual from the entire population is chosen to die; the neighbors compete for the empty site proportional to fitness.

      iii. Imitation (IM) update rule (Hofbauer & Sigmund, 2003; Ohtsuki & Nowak, 2006b): A player's strategy is changed to that of an individual chosen randomly from the entire population or from its own neighborhood, with a certain probability. The probability depends on the difference in the payoffs of the two players.

[8] See Section 3.4 for details on *selective partner choice*, *voluntary participation*, and *migration*.

(b) Wright-Fisher process (Imhof & Nowak, 2006): Each individual produces a number of offspring proportional to its fitness; the next generation is sampled from this pool of offspring. The total population size remains constant.

(c) Fermi rule: For this rule (Traulsen et al., 2007), with a certain probability, an individual's strategy is updated to one of its randomly chosen neighbor. Player $x$ will adopt $y$'s strategy with a probability given by the Fermi function:

$$p(s_x \leftarrow s_y) = \frac{1}{1 + e^{\beta(\pi_x - \pi_y)}} \tag{3}$$

where $\pi_x$ ($\pi_y$) is $x$'s ($y$'s) fitness, $\beta$ is a tunable parameter that represents the imitation strength or intensity of selection, i.e., how strongly the individuals base their decision to imitate on fitness comparison. Note that, in order to use the above rule, an individual must know the fitness of the other players. When this information about the fitness/payoff of the other players is unreliable, then using averaged payoffs (calculated by surveying an available neighborhood) can lead to increased cooperation as shown in (Szolnoki & Perc, 2021).

Stage game strategies alone may undergo evolution, or *co-evolve* with the interaction topology as in (Ashlock et al., 1996).

7. **Individual traits**: Individuals can vary in terms of their *cognitive* traits: the constraints on their memory (i.e, how far back they can recall the past), and their ability to recognize other individuals. There are also variations in their *strategic* ability to adjust their aspiration payoffs (see Section 3.4 for details) and their ability to choose co-players. Some models of evolutionary learning ascribe *psychological traits* such as emotion and mood to individuals. These traits are not part of evolutionary model per-se but are added on mostly to achieve cooperation through increased rationality.

### 3.4 Mechanisms for Incentivizing Coopertion

Since evolution by selection is based on competition between individuals, it rewards only selfish behaviour. Hence, mechanisms that extend elementary natural selection are needed for supporting cooperation. In the context of evolutionary games, a variety of incentivizing mechanisms, many of which use ideas from social theory (Coleman, 1994), have been shown to enhance cooperation. These mechanisms (summarized in Table 4) can be taxonomized into three main categories: *strategic*, *structural*, and *psychological*, although some may belong to multiple categories.

**Strategic:** Individuals strategically employ the following mechanisms:

1. **Kinship and green beards**: Hamilton (1963, 1964a, 1964b) proposed the theory of *kin selection* which hypothesises that social evolution can be understood as a process of *inclusive fitness* maximization. An individual's inclusive fitness is defined as a weighted sum of its personal fitness and the fitness of its genetically related individuals. Weights are given by a *coefficient of relatedness* (Wright, 1922; Ohtsuki, 2010). The core of kin-selection theory is *Hamilton's rule* which tells that natural selection

| | |
|---|---|
| Strategic<br>(Individual) | Kinship |
| | Reciprocity |
| | Group selection |
| | Altruism |
| | Reward/ punishment |
| | Aspiration level |
| Structural<br>(Interaction structure) | Fixed exogenously |
| | Evolves endogenously |
| Psychological<br>(Individual) | Emotion |
| | Mood |

Table 4: Evolutionary mechanisms for dilemma resolution: A taxonomized summary.

will favour an altruistic behaviour so long as the cost to the altruist is offset by a sufficient amount of benefit to sufficiently closely related recipients. Kin selection has been shown in (Nowak, 2006b; Nowak & Sarah, 2013) to result in the evolution of cooperation in dilemmas. Apart from genetic relationship, there are other ways to manifest relations. A *green beard* is a tag or a conspicuous feature that allows the bearer to recognize it in other individuals, and causes the bearer to behave differently towards other individuals depending on whether or not they possess the feature. In a green-beard model (Jansen & Van Baalen, 2006), an individual's decision to cooperate depends on the tags associated with the agents. Tags may be interpreted as imposing an abstract topology on the agents in which an agent's neighbourhood is defined by its tag and a threshold of similarity tolerance. Green beard models require individuals to have only a rudimentary ability to detect environmental signals and, no memory of past encounters is required.

2. **Reciprocity**: The theory of kin selection is useful for explaining cooperation among relatives. In contrast, reciprocity is a model for explaining cooperation between unrelated individuals engaged in repeated encounters. There are three types of reciprocity: *direct*, *indirect*, and *network*.

   Direct reciprocity: Trivers (1971) proposed a model of direct reciprocity for repeated encounters between the same two individuals: if one individual cooperates now, the other may cooperate later. Direct reciprocity can lead to the evolution of cooperation in a well-mixed population if the probability, $w$, of another encounter between the same two individuals exceeds the cost-to-benefit ratio of the altruistic act, i.e., $w > c/b$ (Nowak, 2006b; Nowak & Sarah, 2013).

   Reputation and indirect reciprocity: Direct reciprocity is possible only when the same two individuals encounter each other repeatedly. Indirect reciprocity, based on the idea 'I help you and somebody else helps me', is another mechanism useful for the evolution of cooperation in dilemmas (Nowak & Sigmund, 1998, 2005). In pairwise interactions, one individual typically acts as donor, the other as recipient. The donor can decide whether or not to cooperate. The interaction is observed by a subset of the population who might inform others. An individual's *reputation* is built from these observations (Ohtsuki & Iwasa, 2004; Panchanathan & Boyd, 2004; Nowak & Sigmund, 2005). When others reputations are available, they can be used to define one's individual behaviour (Santos et al., 2021). One could formulate simple heuristic strategies, for example, by having the probability of cooperating with an individual proportional to their reputation (Gross & De Dreu, 2019). Further, switching partners based on reputation can also lead to stable cooperation in networked Prisoner's Dilemma (Fu et al., 2008). A key consideration in reputation-based models is the timescale for updating reputations and that for strategy updates (Xu et al., 2019). Both reputation building and strategy learning are dynamic processes assumed to happen at different timescales. The individuals adopt a fixed behavioural strategy for a sufficiently long time that the reputation distribution stabilizes before they explore other strategies.

Indirect reciprocity was shown to lead to the evolution of cooperation (Hamilton & Taborsky, 2005; Nowak & Roch, 2007; Rankin & Taborsky, 2009; Alexander, 2017; Santos et al., 2018; Berger & Grüne, 2016; Paiva et al., 2018) and social norms (Brandt & Sigmund, 2004; Ohtsuki & Iwasa, 2004). However, this mechanism has substantial cognitive demands relative to direct reciprocity and kin selection where individuals cooperate with those who are related to them. Memory is needed to establish reputations and language is needed to gain and spread information about reputation.

Network reciprocity: This is useful for supporting the evolution of cooperation in a structured population. In a structured population, natural selection can result in cooperation although in a well-mixed population, it results in defection. This is because reciprocity can be induced in a structured population. The players are arranged on a spatially structured topology and interact only with their direct neighbors. In such a setup, if one individual imitates a neighbor's altruistic act, then the neighbor will subsequently experience reciprocity so that two altruistic neighbors help each other, and two defecting neighbors harm each other. This phenomenon of network reciprocity facilitates the spread of cooperation and retards the spread of defection (Nowak & May, 1992; Nowak, 2006b; Wang et al., 2013). For the Prisoner's Dilemma, the condition for network reciprocity to arise is $b/c > k$ (Ohtsuki et al., 2006) where $b/c$ is the benefit-to-cost ratio, and $k$ is the degree of the underlying static regular graph (the condition was generalized in (Allen et al., 2017) to non-regular graphs where $k$ is the average degree). It is noteworthy that, in (Ohtsuki et al., 2006), there is no direct or indirect reciprocity, the evolutionary dynamics is driven by a simple death-birth strategy update rule and this is sufficient for cooperation to evolve. Thus individuals do not need to have cognitive abilities such as those needed by other incentivizing mechanisms such as direct reciprocity (requires recalling previous interactions), indirect reciprocity (requires establishing and maintaining reputations), or kin selection (requires recognising kin). Network reciprocity has been confirmed in several other models (Lieberman et al., 2005; Santos et al., 2006b; Ohtsuki & Nowak, 2007; Nowak & Roch, 2007; Iwagami & Masuda, 2010; Van Doorn & Taborsky, 2011; Konno, 2011; Perc et al., 2013; Débarre et al., 2014; Dercole et al., 2019; Su et al., 2022).

3. **Group selection**: Natural selection is known to act at multiple levels: at the level of individuals and at the level of groups (Wright, 1945). Many researchers (Williams & Williams, 1957; Taylor & Wilson, 1988; Rogers, 1990; Rainey & Rainey, 2003; Wilson & Hölldobler, 2005; Keller, 1999; Michod, 2000) have used this insight to study the evolution of cooperation in populations organised into groups. A simple yet very effective model of *group selection*[9] was studied in (Traulsen & Nowak, 2006). In this model, the population subdivides into groups. Cooperators help others in their own group. Defectors do not help. Individuals reproduce proportional to their payoff and offspring are added to the same group. A group is split into two if it outgrows a certain limit. The splitting of a group is accompanied by the extinction of another group so as to constrain the population size. Only individuals reproduce but selection emerges on two levels. There is competition between groups; some groups grow faster and split more often. In particular, pure cooperator groups grow faster than pure defector groups, while in any mixed group, defectors reproduce faster than cooperators. Therefore, selection within groups favors defectors, while selection between groups favors cooperators. It was shown that cooperation evolves if the benefit-to-cost ratio $b/c$ of the altruistic act satisfies $b/c > 1 + (n/m)$ where $n$ is the maximum allowed group size, and $m$ is the maximum number of allowed groups.

4. **Altruism**: Kin selection can explain cooperation among related individuals. Different ideas are required to explain the emergence of cooperation among unrelated individuals. Trivers (1971) proposed the theory of *reciprocal altruism* to explain this phenomenon. He defined altruistic behaviour as one in which an individual (the altruist) benefits another individual, not closely related, despite having to incur a cost. The theory posits that that if there are opportunities for repeated interactions between the same individuals, then an individual who behaves altruistically only to those which reciprocate the altruistic act will be favored by natural selection. In general, such reciprocal altruism can evolve only if the same individuals meet repeatedly, if they are capable of memory and recognition, and if the benefits to the individual who is helped exceed the costs to the helper. There is strong evidence for reciprocal altruism in a number of animal systems. Inspired by this theory, numerous studies used it to demonstrate that it can lead to a gradual evolution of cooperation in dilemma games from an initially non-cooperative state and sustain cooperation thereafter (Stephens, 1996; Killingback & Doebeli, 2002; Skyrms, 2010).

5. **Reward/Punishment** Several field experiments with humans have shown that *altruistic punishment* (altruistic punishment means that individuals punish defectors, although the punishment is costly for them) of defectors is a key factor in the explanation of cooperation in the ultimatum game (Henrich et al., 2001), a public goods game (Fehr & Gächter, 2002), and social dilemmas (Gurerk et al., 2006). In computer simulations, punishments have also been shown to enhance cooperation when combined with indirect reciprocity (Sigmund et al., 2001), group selection (Boyd et al., 2003; Bowles & Gintis, 2004), and network reciprocity (Nakamaru & Iwasa, 2005). In the context of the Prisoner's Dilemma, Han et al. (2013) combined punishment

[9] See (Marshall, 2011) for the relationship between *group selection* and *kin selection*.

costs (which they refer to as 'apology') with prior commitment to show how it can lead to the evolution of cooperation if the punishment are costly enough. While many studies considered punishment in isolation, (Chen et al., 2015) showed how combining rewards with penalties can be a successful mechanism for promoting cooperation in a public goods game using replicator dynamics.

6. **Aspiration levels**: In aspiration-based models, individuals evaluate their achieved payoffs in relation to their *aspiration payoff.* Each individual has an aspiration payoff which may be calculated in different ways. Typically, this is done in terms of received payoffs and the number of neighbors. The players update their strategies with a probability that depends on the difference between their actual payoffs and their aspiration payoff. Chen and Wang (2008) studied the Prisoner's Dilemma on a population structured as Newman-Watts small-world network (Newman & Watts, 1999) for different average aspiration levels under stochastic rules for strategy updating. They found that the level of cooperation depends on the aspiration level and that cooperation is maximised for intermediate levels of aspiration. Similar works on the role of aspiration in cooperation include (Liu et al., 2016; Zeng et al., 2017; Liu et al., 2019; Zhang et al., 2019; Wang et al., 2021; You et al., 2021; Du et al., 2023).

**Structural:** Much of the early literature on evolutionary games focussed on the evolution of player strategies given a fixed and *exogenously* defined population structure. However, social networks are dynamic structures; over time, individuals create and terminate social ties, thereby altering the structure of the network in which they participate. Such interactions are modelled by extending evolutionary games to allow for the *co-evolution* (Perc & Szolnoki, 2010) of strategies and network structure. In a co-evolutionary system, (Cardillo et al., 2010; Rezaei & Kirley, 2012; Ranjbar-Sahraei et al., 2014; Tanimoto, 2017; Bandyopadhyay & Kar, 2018; Li et al., 2020) strategies evolve by *strategy update rules*, and the network topology evolves by means of *link update rules.* A weight is associated with each network link, and link update rules are used for updating weights, creating new links, and breaking existing links. Co-evolution leads to richer dynamics with more variable parameters and interesting inter-plays between them. In such systems, the timescale that separates the different aspects that co-evolve has a major influence on the dynamics of co-evolution (Pinheiro et al., 2016; Bara et al., 2022).

1. **Fixed exogenously**: The evolutionary dynamics on a static graph crucially depends on the underlying graph and the strategic game that the nodes of the graph play[10]. Different games and different graph structures give rise to a vast combination of cases to consider. The dynamics are, in general, very complex; even in an evolutionary game on a graph in which all nodes play the same strategy (called resident strategy), the problem of determining whether a given mutant strategy introduced at a random node in the graph can invade the entire population under frequency dependent selection is NP-hard and in PSPACE (Lieberman et al., 2005). The probability of a mutant

[10] See (Iyer & Killingback, 2016) for a comparative study of the effects of the structural properties of a graph, such as its average degree, variance in degree distribution, clustering coefficient, and assortativity coefficient, on the promotion of cooperative behaviour in all three classes of dilemma games, and (Szabó & Fath, 2007) for a comprehensive review of evolutionary PD on graphs.

invading a population and driving the resident strategy to extinction is called the *fixation probability.* Computing an approximate fixation probability is #P-hard and in PSPACE (Ibsen-Jensen et al., 2015). Further, it is an open problem to characterize the set of graphs that promote cooperation in dilemmas. Because there is no efficient algorithm for this problem, special cases such as regular graphs (Ohtsuki et al., 2006), complete graphs (Díaz et al., 2014), or some other exogenously defined structure (Nakamaru et al., 1997; Richter, 2019; Locodi & O'Riordan, 2021; Hsiao & Nau, 2022; Zhang et al., 2022) are considered for the study of evolution of cooperation in dilemma games. Under various conditions, structured populations (including set structured populations (Tarnita et al., 2009; Nowak et al., 2010)) have been shown to enhance cooperation in dilemmas relative to well-mixed populations; the extent to which cooperation improves depends on factors such as the stage-game payoffs and the degree distribution of the underlying graph (or the set membership distribution for set structured populations).

An important aspect relating to structured populations is the *contagion* or *herd* behaviour. Herding is a phenomenon in which individuals imitate group behaviours even if it sub-optimal for them to do so (Fowler & Christakis, 2010; Centola et al., 2005; Willer et al., 2009). In order to study the relation between contagion and cooperation, Masuda (2012) introduced *zealous cooperators* (a zealous cooperator is an individual who often mimics the cooperation of neighbors even if defection is more lucrative than cooperation) in the population. For Prisoner's Dilemma and Snowdrift games, they showed that a small fraction of zealous cooperators can reliably induce cooperation at the population level in a well-mixed population. However, for structured population, Matsuzawa et al. (2016) showed that zealous cooperators do not always enhance cooperation and that they may be counterproductive depending on the underlying graph topology.

2. **Evolves endogenously**: Structure can evolve in various ways: through *voluntary participation*, through *partner selection* by the players (see (Stanley et al., 1995; Ashlock et al., 1996) for partner selection), or through *player migration and mobility.*

   In voluntary participation, players can choose whether or not to participate in a game, with players typically making a decision on the basis of previous payoffs. This choice leads to three strategy types: cooperators, defectors, and loners. Voluntary participation has been shown to boost cooperation in dilemmas played on both unstructured and structured populations (Hauert et al., 2002; Szabó & Hauert, 2002).

   Partner selection allows individuals to decrease the risk associated with cooperation by having them choose and refuse their partners. Partners are typically chosen on the basis of the expected payoffs; a partner is chosen only if the previous payoff exceeds a certain threshold. This gives rise to interesting population dynamics; the entangled evolution of individual strategy and network structure constitutes a key mechanism for sustaining cooperation in dilemmas (Stanley et al., 1994; Ashlock et al., 1996; Santos et al., 2006a).

   In a population of mobile agents, the individuals occupy some nodes of a graph while some nodes remain unoccupied. The graph edges determine who interacts with whom.

> Individuals can migrate to vacant sites where migration is typically payoff driven. Immigration enhances cooperation in dilemmas (Vainstein et al., 2007; Helbing & Yu, 2008; Droz et al., 2009; Meloni et al., 2009; Jiang et al., 2010; Cheng et al., 2010; Yang et al., 2010; Cheng et al., 2010; Roca & Helbing, 2011; Lütz et al., 2021) relative to no-migration case and the level of cooperation depends on the threshold payoff that individuals use for deciding to migrate.
>
> Since the individuals in an evolutionary game imitate others, two types of individuals can be distinguished: *leaders* whose strategies are imitated and *followers* who imitate leaders. In time varying networks, leader/ follower roles also vary with time. It is therefore possible to study the dynamics of leadership. For example, in the context of PD games, Zimmermann et al. (2001) allowed individuals to create and terminate links and showed how this leads to global cooperation with the emergence of a leader who is a cooperator with the maximum number of connections. Others (Kim et al., 2002; Szolnoki & Szabó, 2007; Rong et al., 2019) have studied the evolution of leaders in the context of a variety of population structures.

**Psychological:** Individual psychological traits such as *emotion* and *mood* can effect their learning behaviour. Emotion, a mental state resulting from one's assessment of the world from their own viewpoint, is known to have a place in human motivation and behaviour (Frijda et al., 1986; Scherer & Moors, 2019; Boone & Buck, 2003). This knowledge lead to the investigation of the impact of player emotion on the evolution of cooperation in dilemmas. A novel form of imitation in which the players imitate the emotional profiles of other players was examined in (Szolnoki et al., 2011, 2013). Each player is assigned an *emotional profile*. Player $x$'s emotional profile is defined as a pair $(\alpha_x, \beta_x)$ where $\alpha_x$ is the probability that $x$ will cooperate with a player $y$ if $p_x \geq p_y$ (where $p_i$ denotes player $i$'s payoff) and $\beta_x$ is the probability that $x$ will cooperate with $y$ if $p_x < p_y$. Goodwill is associated with the former and envy with the latter. For Snowdrift, Stag-Hunt and Prisoner's Dilemma games, they found that players are much more likely to cooperate with less successful neighbors than they do with the more successful ones, suggesting that goodwill facilitates the evolution of cooperation while envy favors the evolution of defection.

Some works (Chen et al., 2021) considered a Prisoner's Dilemma played on a square lattice with four types of player emotion: joy, anger, regret, and neutral. Players' emotions were quantified, and accumulated over the most recent interactions. They assumed two types of players: competitive (who consider only their own payoffs) and non-competitive (who consider their own payoffs together with others'). They showed that the proportion of non-competitive individuals plays a key role in promoting overall cooperation. The probability that a player adjusts their strategy to that of a random neighbor is calculated in terms of the player's emotion and the difference between their own payoff and that of a random neighbor. They showed that the existence of non-competitive individuals promotes cooperation and that cooperation rate depends on the emotion accumulation length.

Apart from emotion, certain *mood* motivated strategies (Dercole et al., 2019) have been shown to give rise to network reciprocity and result in the fixation of cooperation in a graph structured Prisoner's Dilemma.

## 4. Reinforcement Learning

Reinforcement learning (RL) (Littman, 2015; Sutton & Barto, 2018) is learning what to do – how to map situations to actions – so as to maximize a numerical *reward function*. Learning is accomplished by trial-and-error; the learner is not told which actions to take, but is given evaluative feedback (in the form of rewards/payoffs) when it tries an action. The learner must try different actions and use the received feedback to discover which actions yield the most reward. RL is a stimulus-response model grounded in theories of learning from cognitive psychology. It is derived from the eminent psychologist Thorndike's law of effect (Thorndike, 1898, 1931; Postman, 1947; Thorndike, 2013; Estes, 1967) which states that if responses to environmental stimuli are *rewarded* (*punished*), they are likely to be repeated (avoided). What consitutes as reward or punishment depends on the learner's *aspirations*. With this law as the basis, reinforcement learning models originated in the mathematical psychology literature with the works of Estes (1950, 1967), Bush and Mosteller (1955), and Suppes and Atkinson (1960). Similar models were subsequently developed in other disciplines. In the economics literature, early pioneering work by Simon (1955) led to a huge follow up literature but the Roth and Erev model (Roth & Erev, 1995; Erev & Roth, 1998) has been very influential in the context of RL in games. In the computer science literature, there are various RL models such as Narendra and Thathachar's learning automata model (Narendra & Thathachar, 1974; Rubinstein, 1986; Thathachar & Sastry, 2003; Narendra & Thathachar, 2012) and Watkins and Dayan's (1992) Q-learning model (and its variants such as SARSA, temporal difference learning, actor-critic model (Sutton & Barto, 2018)) built on the theory of dynamic programming (Bellman & Dreyfus, 1962).

In all of these different RL models, the learning agent is bounded in terms of its rationality. Game-theoretic rationality is not needed but cognitive abilities are needed for trial-and-error learning (contrast this with the pre-dominantly inheritance base learning in evolutionary games that does not require such cognitive abilities). Thus, an RL agent must be able to observe the feedback for its previous actions and revise its strategies through *reinforcement rules* that favor the best and inhibit the worst actions. However, reinforcement learners do not need to know their opponent's strategies or payoffs. Despite being modest in terms of the learner's cognitive abilities, RL models have been used successfully for explaining many deviations from game-theoretic solutions that are empirically observed in laboratory experiments with human subjects (McAllister, 1991; Roth & Erev, 1995; Mookherjee & Sopher, 1997; Chen & Tang, 1998; Erev & Roth, 1998, 2002; Erev et al., 1999).

The early RL models were intended for single-agent learning. In these models, the agent is situated in an environment that it interacts with repeatedly by responding to environmental stimuli. The interactions take place over a series of discrete time periods with one stimulus-response interaction in each time period. In any time period, the learner uses their past experience to choose a response (i.e., action). Those actions that previously resulted in positive stimuli (i.e., satisfactory outcomes, rewards or payoffs) tend to be repeated in the future while those that resulted in unsatisfactory outcomes are avoided. An agent's *strategy* is defined by a probability distribution over the available actions and an action is chosen randomly as per the distribution. The probabilities are initialised arbitrarily and then updated iteratively through a process of learning.

The single agent RL models were subsequently extended to multi-agent strategic games (Shoham et al., 2003; Busoniu et al., 2010). Bendor et al. (2001a) provide a good overview of RL in repeated games. In multi-agent learning, each agent's environment includes all the other agents in the system. The players repeatedly play a *stage game* (which in our case is a dilemma game) in successive time periods. For the stage game, each player has a finite set of pure strategies to choose from and all players simultaneously and independently make this choice. After choices are made, each player receives a reward/ payoff which depends on the combination of choices of all the players and is given in terms of the payoff matrix for the stage game. Learning is stochastic; an agent learns a distribution over the possible actions such that the resulting choices maximise their reward accumulated over a series of time periods. Learning is formally modeled under the general framework of a *stochastic game* (SG) (Shapley, 1953) also referred to as a *Markov game* (MG). In a stochastic game, the process of learning begins with action probabilities initialised randomly and ends when the strategies converge. The speed and the point of convergence depend on the probability update rules employed by the agents. Probabilities can be updated in a variety of ways giving rise to a host of RL methods. Many of these are derived from three fundamental models: the Bush-Mosteller model, the Roth-Erev model, and the Q-learning model. The following sections provide an overview of these three methods.

### 4.1 The Bush-Mosteller Model

Bush and Mosteller (Bush & Mosteller, 1951, 1955; Mosteller, 1956) introduced a sequential probability learning model that was derived from Estes' statistical theory of learning (Estes, 1950). It was first introduced in the context of single-agent learning but was subsequently extended to multi-agent settings (Flood, 1952; Bales et al., 1952; Hays & Bush, 1954) including dilemma situations. It has been used to achieve cooperation in dilemmas and to explain observed data in laboratory studies with humans (Flache & Macy, 2002; Macy & Flache, 2002; Izquierdo et al., 2007; Izquierdo & Izquierdo, 2008).

We will overview (Izquierdo et al., 2007; Izquierdo & Izquierdo, 2008) as a representative extension of the BM model (Bush & Mosteller, 1955) to two-player dilemma games. The agents learn over a series of time periods by playing a PD game in each stage. The players decide what action to choose stochastically: each player's strategy is defined by a probability distribution over their actions. The probabilities are initialised arbitrarily and then updated in each time period as per an update rule.

In each stage, each player chooses an action according to their strategy and receives the corresponding payoff. Each player then updates their strategy as follows. Each player increases their probability of undertaking a certain action if it led to payoffs above their aspiration level, and decreases this probability otherwise. In any time period $t$, strategy updating is done in two steps. First, each player $i$ calculates her stimulus $s_i(y)$ (where $y$ denotes the pure strategy profile played in time period $t$, and $s_i(y)$ is a number in the interval[11] $[-1, 1]$) from the payoff $u_i(y)$ she received in that time period as follows:

$$s_i(y) = \frac{u_i(y) - A_i}{sup_{k \in Y} |u_i(k) - A_i|} \tag{4}$$

[11] A special case of the BM model where all stimuli are positive was originally considered in (Cross, 1973), and later analyzed in (Börgers & Sarin, 1997), where it was also related to the replicator dynamics.

where $Y$ denotes the space of all pure strategy profiles, and $A_i$ denotes $i$'s aspiration level. The players are thus assumed to know $sup_{k \in Y}|u_i(k) - A_i|$. Next, each player $i$ updates her probability $p_{i,y_i}(t+1)$, of undertaking the selected action $y_i$ at time $t+1$ as follows:

$$p_{i,y_i}(t+1) = \begin{cases} p_{i,y_i}(t) + l_i \times s_i(y) \times (1 - p_{i,y_i}(t)) & \text{if } s_i(y) \geq 0 \\ p_{i,y_i}(t) + l_i \times s_i(y) \times p_{i,y_i}(t) & \text{otherwise} \end{cases} \tag{5}$$

where $0 < l_i < 1$ denotes the learning rate. Thus, the higher the stimulus magnitude (or the learning rate), the larger the change in probability. The updated probability for the action not selected derives from the constraint that probabilities must add up to one. The next iteration then begins with an action being chosen as per the updated probabilities. When learning, players in the BM model use only information concerning their own past choices and payoffs. Information about the payoffs and choices of their co-player is not needed.

### 4.2 The Roth-Erev Model

The Roth and Erev (RE) model (Roth & Erev, 1995) originated from the works done by experimental psychologists Skinner (1938) and Herrnstein (1961, 1970). This is a stochastic model in which an agent's strategy is a probability distribution over the available actions. The probability of choosing an action is proportional to the total accumulated rewards from choosing it in the past. In the basic model of (Roth & Erev, 1995), the probabilities are calculated as follows. At time $t = 1$, (before any experience has been acquired) each player $i$ has an initial propensity to play his $k$th pure strategy, given by some number $q_{i,k}(1)$. If player $i$ plays his $k$th pure strategy at time $t$ and receives a payoff of $x$, then the propensity to play strategy $k$ is updated by setting

$$q_{i,k}(t+1) \quad = \quad q_{i,k}(t) + x \tag{6}$$

while for all other pure strategies $j$,

$$q_{i,j}(t+1) \quad = \quad q_{i,j}(t) \tag{7}$$

The probability $p_{i,k}(t)$ that player $i$ plays his $k$th pure strategy at time $t$ is

$$p_{i,k}(t+1) \quad = \quad q_{i,k}(t) / \sum q_{i,j}(t) \tag{8}$$

where the sum is over all of player $i$'s pure strategies $j$. There are similarities and differences between the RE and the BM models. Like the BM model, the RE model also obeys the *law of effect* (choices that have led to good outcomes in the past are more likely to be repeated in the future). However, in contrast to the RE model, the BM model allows for negative stimuli. Further, unlike the BM model, the RE model obeys the *law of practice* (learning curves tend to be steep initially, and then flatten).

The above described RE model was later extended (Erev & Roth, 1998) by introducing additional parameters for 'experimentation' and 'forgetting' (i.e., weighing recent stimuli more heavily than past ones) to make it more robust at explaining and predicting human behaviour observed in experiments.

### 4.3 The Q-Learning Model

The Q-learning model (Watkins & Dayan, 1992) and its variant SARSA (Modified Connectionist Q-learning) (Rummery & Niranjan, 1994) originated from the theory of dynamic programming (Bellman & Dreyfus, 1962). These models were intended for single agent systems and have the property of guaranteed convergence to optimal strategies. A key difference between single and multi-agent settings is that the environment for the former is stationary while for the latter it may be non-stationary (Tuyls & Weiss, 2012). As a result, optimality of learnt behaviours cannot be guaranteed if there are multiple agents. These single agent learning models were later extended to make them suitable for multiagent settings (Bloembergen et al., 2015; Busoniu et al., 2008).

At a high level, multi-agent Q-learning is modelled by a Markov game. In each time period, each player is in one of several possible states and chooses one of the available actions. For example, for the Prisoner's Dilemma, the actions available to any agent $i$ are $A_i = \{C, D\}$, and the rewards are given by the payoff matrix for the dilemma game. If an agent receives a payoff $r_t$ at time $t$, his discounted reward for the time period is $\delta^{t-1} \times r_t$ where $0 < \delta \leq 1$ is the discount factor. The next state that results depends on the combination of actions chosen by all the players. The goal for each agent is to learn a policy, i.e., a mapping from states to actions such that the expected sum of their discounted future rewards is maximized.

Q-learning works by calculating approximate numerical estimates of state-action values. For any time period $t$ and state $s$, a player's Q-value $Q(s, a)$ for action $a$ is the maximum discounted sum of future rewards the agent can expect to receive if it starts in $s$, chooses the action $a$, and then continues to follow an optimal policy. The initial Q-values are set arbitrarily for all the states and then iteratively updated as follows. Suppose that $r$ is the reward received from executing $a$ in $s$ at time $t$ and that the resulting new state is $s'$. Then $Q(s, a)$ is updated as follows:

$$Q(s,a) = Q(s,a) + \alpha\big(r + \delta \times \underset{x}{max}\ Q(s',x) - Q(s,a)\big). \tag{9}$$

Here, $\alpha$ is the agent's learning rate and $\delta$ is the discount factor. For calculating *max*, the payoff to the focal player is calculated under the assumption that the other players play certain strategies.

How an agent's state $s$ is defined depends on the number of agents playing the game and the agent interaction structure. For the two-player case, $s$ is typically defined in terms of the action the agent chose in the previous time period. Such a player can distinguish between two states: $s = C$, and $s = D$. Alternatively, state may be given by the combination of choices made by the player and their opponent, leading to $s \in \{(C, C), (C, D), (D, C), (D, D)\}$. A state may be more elaborate and include longer histories and payoffs (Vrancx et al., 2008; Bazzan et al., 2011).

A Q-learner chooses actions seeking to balance the exploration of new and seemingly sub-optimal actions, with the exploitation of those actions that are optimal as per their Q-values. Each action is chosen with a certain probability, with the probability for any action being determined in terms of the Q-value (as per Equation 9) for the action.

Most of the literature on RL in dilemmas has used one of these three models or some variant of it. Although all RL models follow the same underlying principle that actions resulting in satisfactory payoffs are more likely to be repeated, there are variations in terms of the model parameters. The following section lists the key parameters.

| **Stage game (SG) parameters** | |
|---|---|
| Payoff matrix | Deterministic or Non-deterministic |
| | Symmetric or Asymmetric |
| | Positive or negative payoffs |
| **Structural parameters** | |
| Types of learners | Homogeneous or heterogeneous |
| Partner links | Static or dynamic |
| | Exogenous or endogenous |
| Partnership duration | Frequency of restructuring partnerships |
| **Individual player parameters** | |
| Cognitive | Learning speed |
| | Memory bound |
| | Inertia |
| | Ability to reognize co-players |
| Strategic | Ability to choose co-players |
| | Aspiration level |
| Psychological | Emotion, mood |

Table 5: Reinforcement learning models: A summary of key parameters.

### 4.4 Reinforcement Learning: Key Parameters

The learning trajectory of a model is determined by its parameters. The various models differ in terms of their parameters. The key parameters of an RL model (summarized in Table 5) are as follows:

1. **Type of stage game**: The stage game payoffs could be *deterministic* or *non-deterministic*, *symmetric* or *asymmetric*. Much of the literature is focused on deterministic and symmetric games. Further, payoffs may be positive only or else allow both positive and negative payoffs. This difference leads to a difference in the type of RL stimulus. In RL, there are two main types of behaviours: *approach* and *avoidance*. Approach behaviour is the tendency to repeat the associated choices after receiving a *positive stimulus*; avoidance behaviour is the tendency to avoid the associated actions after receiving a *negative stimulus*. It is possible to allow for avoidance behaviour in addition to approach behaviour. Some models, such as (Cross, 1973), consider only

positive stimuli. Others, such as (Lahkar, 2017), allow for both positive and negative stimuli. Learning may converge to different points depending on the type of stimulus.

2. **Types of learners**: The population may be *homogeneous* (Stimpson & Goodrich, 2003; Crandall & Goodrich, 2005) or *heterogeneous* (Vassiliades et al., 2011)) in that the players use the same (different) learning strategies.

3. **Partner links**: The interaction structure defines an agent's co-players and is given by a lattice or a graph structure. The structure may be imposed *exogenously* or *evolve endogenously*. Further the links may be held *static* throughout the learning process (Bazzan et al., 2011; Ezaki & Masuda, 2017; Feehan & Fatima, 2022), or allowed to change *dynamically* during the course of play (Skyrms, 2004; Ezaki & Masuda, 2017). A key aspect related to dynamic links is the frequency of re-structuring, i.e., the number of times a matched pair will play a game before re-matching occurs.

4. **The speed of learning**: The speed of learning may be held constant or may be allowed to change during the process of learning. Some models such as the RE model (Roth & Erev, 1995; Erev & Roth, 1998) vary the learning rate, initially learning fast but gradually slowing down, thus obeying the law of practice. In other models such as the BM model (Bush & Mosteller, 1955) the learning rate does not change. This parameter is a key determinant of the time taken for learning to converge and also the point of convergence; see (Izquierdo & Izquierdo, 2008) for an account of the effect of learning rate on the evolution of cooperation in a PD game.

5. **The learner's ability to forget the past**: When taking experience into account, a learner may give equal weight to all experiences regardless of the time (the distant past or the recent past) the experience was gained. Alternatively, newer experience could be given a higher weight than old ones giving the learner the ability to forget the past (Erev & Roth, 1998; Rustichini, 1999; Beggs, 2005). How far back a player can remember depends on their *memory bound*.

6. **Inertia**: In some models such as (Izquierdo et al., 2008), the action that an agent chooses in any time period is given by the following rule. If the payoff of the action chosen in the previous time period is no less than their aspiration, then the same action is chosen again. Otherwise, some other action is chosen with probability $1 - p$ where $p$ indicates *inertia* toward the recently chosen action. In some models such as (Karandikar et al., 1998; Bendor et al., 2001a, 2001b), there is *inertia*, i.e., a positive bias toward most recently selected actions. Others such as (Bush & Mosteller, 1955; Macy & Flache, 2002) lack inertia. Learning strategies are typically formulated so as to balance *inertia* and *experimentation*, i.e., trying new actions regardless of the past payoff experience.

7. **Ability to recognize co-players**: Some RL models ascribe higher levels of rationality to the individuals. For example, the ability to recognize co-players. Defections become safe if players are anonymous. The identification of individuals is a possible means for discouraging defections (Ozaita et al., 2020).

8. **Ability to choose co-players**: In the context of dynamic networks, individuals must be able to strategically choose and refuse co-players (Skyrms, 2004).
9. **The aspiration level**: The payoffs received during RL are evaluated against an *aspiration payoff*. In some models such as (Roth & Erev, 1995) the aspiration payoffs are not explicit while in others such as the BM model (Bush & Mosteller, 1955) they are explicit. A learner's aspiration level may be fixed as in (Bendor et al., 1994) or varied endogenously as in (Karandikar et al., 1998; Bendor et al., 2001a, 2001b; Macy & Flache, 2002). Learning agents can thus respond to stimulus by i) adapting their behaviour, and ii) by adapting their aspirations, a process known as *habituation*. Aspirations may be varied in many different ways, and the level and adaptability of aspirations crucially determine the learning trajectory.
10. **Psychological traits**: The important role of psychology in human behaviour led to explorations of psychologically motivated RL models. In this context, several works (Gracia-Lázaro et al., 2012b; Feehan & Fatima, 2022) combined individual psychological traits such as *emotion* and *mood* with the rational decision making abilities of individuals to show how these traits impact on cooperation in dilemmas.

### 4.5 Mechanisms for Incentivizing Coopertion

In the context of games, RL can be used for various purposes. One of the uses is to learn to play some equilibrium. However, our focus is on the use of RL for dilemma resolution. Many studies (Sandholm & Crites, 1996; Leibo et al., 2017) using RL for dilemma games showed that, by itself, an RL model may be insufficient for dilemma resolution. In order to overcome this problem, a basic RL model is typically supported with some mechanisms for incentivising cooperation. With such mechanisms, many theoretical (Kim, 1999; Palomino & Vega-Redondo, 1999; Bendor et al., 2001a, 2001b; Izquierdo et al., 2007; Norman, 1968, 1972) and empirical (Macy, 1991; Crandall & Goodrich, 2005; Izquierdo et al., 2007; Izquierdo & Izquierdo, 2008; Masuda & Ohtsuki, 2009; Bazzan et al., 2011; Yu et al., 2015; Ezaki & Masuda, 2017; Ozaita et al., 2020) works have shown mutual cooperation

| | |
|---|---|
| Strategic (Individual) | Prosociality |
| | Sanctions |
| | Reputation |
| | Aspiration |
| | Reciprocation |
| Structural (Interaction structure) | Fixed exogenously |
| | Evolve endogenously |
| Psychological (Individual) | Emotion |
| | Mood |

Table 6: Reinforcement learning: A taxonimized summary of mechanisms.

as the outcome for dilemma games. Literature on this topic has utilized a variety of mechanisms which may be taxonomised into three main categories: *strategic*, *structural*, and *psychological*, although some may belong to multiple categories. These are summarized in Table 6.
**Strategic:** The following strategic mechanisms are employed at the level of individuals:

1. **Prosociality**: Prosociality, an individual's tendency to take into account the rewards of others for the appraisal of their own rewards, has the capacity to resolve dilemmas. For example, in the context of Stag-Hunt coordination games, Peysakhovich and Lerer (2017, 2018a) examined a deep RL strategy by adding an element of *prosociality* to it. Starting from randomly initialized policies, they compared performance in three different situations: both agents are selfish, one is selfish and the other prosocial, and both are prosocial. Coordination was best achieved when both agents are prosocial, but having just one prosocial agent can also help lead the agents to coordinate on Pareto-dominant outcomes. More recently, Fan et al. (2022) studied Q-learning for a PD played on a square lattice with periodic boundary conditions. By defining and individual's reward to include the rewards of their neighbours they showed that Q-learning can effectively promote cooperation. Their results generalised well to small-world and scale-free networks.

2. **Sanctions**: Cooperation may be achieved by means of sanctioning mechanisms for punishing non-cooperative behaviours. Sanctions may be administered in various ways by various sanctioning agencies. For example, payoffs may operate as sanctions in PD games. In this case, the sanctioning mechanism may be centralized (Babes et al., 2008; Grimm & Mengel, 2011), or a decentralized one (Kosfeld & Riedl, 2004) in which those individuals who cooperate punish free riders by decreasing their payoffs. For spatial PD games, individuals can be punished by making participation voluntary (this allows cooperators the opportunity to decline interaction with free riders), or by allowing individuals to create and terminate links with co-players (Macy, 1991). For all these mechanisms, the achievement of cooperation depends on the magnitude and severity of sanctions (Macy, 1991), the players initial propensities to cooperate, and the number of individuals in the population.

3. **Reputation**: A key obstacle for cooperative behaviour is the anonymity of players. Free riders can get away without retaliation if they cannot be identified. It is therefore crucial to be able to identify free riders. This can be realized by means of reputations or some form of identification marks. Phelps (2013) studied a donations game played by a structured population in which reputations are used by the players to manipulate the network connections to their strategic advantage and showed how the network evolves. Ozaita et al. (2020) investigated a Bush-Mosteller model together with *ethnic markers*[12] (a *marker* is an observable agent characteristic useful for identifying individuals) in *coordination dilemma games* in order to study the influence of markers on agent coordination. In addition to choosing a strategy for playing the coordination game, agents also choose partners on the basis of their markers. The authors showed

[12] Contrast this with (Macy & Flache, 2002) where aspirations are considered without ethnic markers.

that markers allow to resolve the coordination problem through RL provided the agent aspiration levels lie in a suitable range.

4. **Aspiration**: In any RL model, there are two distinct cognitive mechanisms that guide a learner toward better outcomes: *approach* which is driven by rewards, and *avoidance* which is driven by punishments. What is reward and what is punishment depends on a learner's aspiration level. Aspirations thus crucially shape the learning process, and in the context of dilemmas, whether behaviours converge and how quickly they converge to cooperation depends on the aspirations of the learners. Many studies have demonstrated this phenomenon. Karandikar et al. (1998) extended the fixed-aspiration model given in (Bendor et al., 1994) (which, in turn, has its origins in (Mosteller, 1956)) by varying aspirations endogenously. For a class of $2 \times 2$ games which includes PD, they showed analytically that, conditional on the speed at which aspirations are updated, both players ultimately cooperate most of the time. Other variants of the BM model (Bush & Mosteller, 1955), have been studied elsewhere (Izquierdo et al., 2007; Izquierdo & Izquierdo, 2008). Flache and Macy (2002) and Macy and Flache (2002) introduced a model which integrates the BM (Bush & Mosteller, 1955) and the RE (Roth & Erev, 1995; Erev & Roth, 1998) models, and used computer simulations to analyse the effect of interaction of the model parameters (the aspiration level, the learning rate, and a probability update parameter) on the cooperation rate for three social dilemmas, viz., Prisoner's Dilemma, Stag Hunt, and Chicken. Stimpson and Goodrich (2003) extended the model in (Karandikar et al., 1998) ((Karandikar et al., 1998) is for $2 \times 2$ games) to multi-agent games and showed how it leads to mutual cooperation in self play. Izquierdo et al. (2008) extended Macy and Flache's (2002) work on aspiration-based reinforcement learning for $2 \times 2$ dilemma games by providing analytical insights into the dynamics of the model. In particular, they analyzed the robustness of (Macy & Flache, 2002) to occasional mistakes made by players in choosing their actions (i.e. trembling hands) and showed how the inclusion of small quantities of randomness in the players' decisions can change the dynamics of the model dramatically.

5. **Reciprocation**: RL in which the learning is over repeated game strategies (rather than only stage game strategies) help enhance cooperation. Repeated game strategies facilitate *reciprocation*; they are typically conditioned on some history of choices made previously, rather than being guided merely by rewards. This approach was used in (Erev & Roth, 2002) for the RE model to enable agents to reciprocate, thereby resulting in more cooperation in PD and a better prediction of experimental data relative to an approach in which only stage game strategies are learnt. Other works have also shown that learning repeated strategies leads to more cooperation in dilemmas. For example, Crandall and Goodrich (2005, 2011) introduced an RL algorithm called M-Qubed (an acronym for Max or Minimax Q-learning) in which the agents learn to make compromises to reach mutually beneficially outcomes. In self-play, which may be viewed as a form of reciprocation, M-Qubed agents were shown to resolve dilemmas. The effectiveness of repeated game strategies was also confirmed in (Masuda & Ohtsuki, 2009) for PD game using a temporal-difference variant of the SARSA algorithm.

**Structural:** The following mechanisms are employed at the population level:

1. **Exogenously fixed structure**: For dilemmas that are played by a population, cooperation can be achieved by using an appropriate interaction structure. One possibility (Bazzan et al., 2011) is to have a hierarchical structure over the population with individuals who are above in the hierarchy using knowledge about the Q-values for the agents below to recommend to them actions directed toward reaching socially desirable outcomes. In this approach, the underlying assumption is that such a hierarchy exists and that knowledge about other agents' Q-values is available. In contrast to this approach, others have considered independent learners without any hierarchy. For example, Ezaki and Masuda (2017) considered a PD played on a static regular graph by learners using the BM model and observed network reciprocity under certain conditions that are given in terms of the relation between the benefit-to-cost ratio for the dilemma game and the degree of a node. Similar results were also confirmed in (Cassar, 2007; Grujić et al., 2010; Rand et al., 2011; Gracia-Lázaro et al., 2012b; Rand et al., 2014; Fan et al., 2022).

2. **Endogenously evolving structure**: Interaction structure may be imposed exogenously, or else allowed to evolve endogenously by having players choose their co-players in addition to choosing their stage game strategies. The partner selection approach is instrumental in resolving dilemmas. For the Prisoner's Dilemma, Skyrms (2004) studied reinforcement learning for partner selection; the player strategies are fixed but the interaction structure changes as a result of players being allowed to choose their co-players. The players use RL to learn who to choose as co-players. The number of interactions in a given duration is not the same for each individual. If dynamic aspirations are introduced, cooperators learn to visit only cooperators. Defectors learn to interact with defectors. The population of players gets segregated into two mutually exclusive classes, each of which interacts exclusively with itself. Partner selection has been used in many other works including (Macy, 1991; Phelps, 2013).

**Psychological:** Instead of viewing individuals as purely rational decision makers, a broader perspective can be taken by combining an individual's affect with their rationality to study learning dynamics. Despite the nascent state of research on the role of affect on decision making, there is growing evidence that emotion and mood are potent and predictable drivers of decision making (Schwarz, 2000; Lerner et al., 2015). This finding led to investigations on the existence of a link between affect and cooperation in dilemmas. For a Chicken dilemma, Hertel et al. (2000) studied the effect of mood on human cooperative choices under laboratory conditions. They concluded that, contrary to the presumed simple relation stating positive mood leads to higher cooperation than negative mood, positive mood led to quicker heuristic style decision making. On the other hand, negative mood produced a more time-consuming decision making possibly leading to more cooperation than positive mood. Yu et al. (2015) extended the Q-learning model by endowing agents with emotions and studied dilemma games played on different network structures. The individual agents have an internal model for emotion appraisal. Emotions are appraised in terms of individual and social welfare (the term 'social' is used to refer to the agent's neighbourhood in the network topology) and the relation between them. They studied different ways of appraising emotions together with different network topologies and showed how these differences can impact cooperation in two-person Prisoner's Dilemma, Stag Hunt, and Chicken games. Horita et al. (2017) examined whether *moody conditional coperation* (MCC) (Gracia-Lázaro et al., 2012b; Grujić et al., 2014; Cimini & Sánchez, 2014; Ezaki et al., 2016; Gutiérrez-Roig et al., 2014) observed in humans playing a repeated PD could be explained by RL. MCC is a behaviour rule under which a player's probability to choose an action depends on the amount of cooperation they observed in the previous round and their own previous action. More specifically, a player cooperates more when more of their neighbors cooperated in the previous round, and further, a player's probability to cooperate also depends on his mood, i.e., whether the player himself cooperated in the previous round. They showed that the BM model (Bush & Mosteller, 1955) and the RE model (Roth & Erev, 1995; Erev & Roth, 1998) account for the observed human behaviour roughly as accurately as the MCC model did. Some works (Collenette et al., 2017b, 2017a; Feehan & Fatima, 2022) introduced simulated mood (Marsella et al., 2010), in the decision-making process of reinforcement learners and showed how the addition of mood influences cooperation in a spatial Prisoner's Dilemma. In general, the literature on this topic is recent and scant, the existing results lack robustness as the studied models are very stylized with huge variations in terms of the methods used for affect communication, appraisal, and reasoning.

To sum up, RL is directed toward bounded rational behaviour: the players *satisfice* rather than *maximise* payoffs. Players choose their current actions on the basis of their past experience: those actions that resulted in satisfactory (satisfaction is judged against an aspiration level which they may acquire through social inheritance or experience) payoffs are more likely to be chosen relative to actions with unsatisfactory payoffs. In contrast to players in classical game theory, RL agents are informationally and cognitively less demanding. RL agents do not have a model of their environment, i.e., the strategic structure of the game. Yet, by combining aspirations, reciprocation, and forgiving strategies, myopic reinforcement learners can learn to play dominated strategies and enhance cooperation in dilemma games. Further, these models provide support to data observed in experiments with human subjects.

Although the basic principles of reinforcement learning and evolutionary game theory appear to be different, there are links between them. The principles that underlie strategy update rules are analogous; in both, evolution and RL, the probability that an individual uses a given strategy increases if the associated payoff is above some benchmark and decreases if below. In evolution, the benchmark is typically assumed to be the mean payoff for the population. In RL, the benchmark depends on an individual's aspirations. Börgers and Sarin (1997) gave a formal analogy between reinforcement learning at the individual level and biological evolution. They used the BM model (Bush & Mosteller, 1955) with aspiration level exogenously fixed at zero. The sets of stimuli in their learning model correspond to the populations of players in the biological model. The re-programming of stimuli in their learning model is the analog of the reproduction and death processes in the biological model. For two-player normal form games, they showed that their model converges to replicator dynamics. On a related note, Tuyls et al. (2003) and Tuyls et al. (2006) considered certain classes of two-player matrix games and derived a connection between the exploration-exploitation scheme of RL and the selection-mutation mechanisms of evolutionary game theory; exploration being analogous to mutation and exploitation to selection. Further, there are links between RL and replicator dynamics; Kaisers and Tuyls

(2010) showed empirical confirmation of the match between the learning trajectories of their frequency-adjusted Q-learning (FAQ-learning) algorithm and replicator dynamics for three $2 \times 2$ games, viz., Prisoner's Dilemma, Battle of Sexes, and Matching Pennies.

## 5. Best-Reply Learning

The cognitive ability of a reinforcement learner is limited to suitably adjusting their strategies according their own received feedback without any knowledge of their opponent's strategies or payoffs, i.e., reinforcement is based on the learner's personal past experience. In contrast, best responders have a relatively higher degree of rationality. Best responders are epistemic learners; they possess knowledge about the structure of the game and how the combination of others' actions and their own affect their payoffs (Walliser, 1998; Fudenberg et al., 1998). Owing to their high rationality[13], best-reply models have been heavily used in the literature to explain data gathered from laboratory experiments with humans playing dilemma games. To this end, the parameters of a best-reply model are fitted to experimental data. The fitted learning model can then be used to simulate the behaviours of players under various conditions of interest that could not easily be tested with human subjects, for example when the horizon is too long for subjects to play under laboratory conditions.

In more detail, best-response learning works as follows. A learner observes all past moves of their opponent and uses this information to form a *model* of the opponent. The model contains their beliefs about how the opponent acted in the past. Learning is an iterative process; the learner updates its model of the opponent, uses it to anticipate the opponent's next move and play a myopic best response to the anticipated move.

Since a learner's best response depends on their acquired beliefs, some form of belief revision rules are needed for updating beliefs as the process of learning unfolds and new experience is gained. On the basis of their beliefs, a learner extrapolates the future behaviours of the other players. Extrapolation can be done in a variety of ways giving rise to a host of best-response models described in Sections 5.1 to 5.4.

### 5.1 Best Reply Model: Cournot Dynamics

In Cournot learning (Cournot, 1838), a stage game is played repeatedly over a series of discrete time periods. A learner's opponent model simply contains the action that the opponent played in the previous time period. A learner assumes that the opponent will do the same thing they did in the previous time period. The learner then chooses a strategy that is their best response under this assumption. The dynamics that results when all players do this is called *Cournot dynamics*.

[13] Note that while best responders have a higher degree of rationality relative to reinforcement learners, both RL and best-reply models are individual learning models as opposed to the population learning phenomenon of evolutionary games.

### 5.2 Best-Reply Model: Fictitious Play

Fictitious play (Luce & Raiffa, 1957; Shapley, 1964; Cournot, 1838; Fudenberg et al., 1998; Berger, 2007) was introduced by Brown [14] (1951) and Robinson (1951). In fictitious play, a stage game is played repeatedly over a series of discrete time periods. In each round, each player can observe the action chosen by their opponent. Based on these observations, each player forms certain beliefs about their opponent in the form of an opponent model. In each round, beliefs are updated by considering the entire history up to a current round, and under the assumption that the opponent is playing a stationary mixed strategy. Beliefs are initialised arbitrarily and then updated in each round as follows. If $A$ is the set of the opponent's available actions for the stage game, and for each $a \in A$, $w(a)$ denotes the number of times that the opponent played action $a$ so far, then the agent assesses the probability of $a$ in the opponent's mixed strategy as

$$P(a) = \frac{w(a)}{\sum_{a' \in A} w(a')}.$$

Each player then chooses an action that is a best reply with respect to the updated probabilities breaking any ties randomly. The next iteration begins after an action is chosen by both agents. Beliefs thus evolve over time as a player gains experience. The learning process[15] was shown to converge in a large class of games (Berger, 2007) though not in all games (Shapley, 1964).

More generalised versions of this basic form have incorporated various additional features. One variation is to use different weights for different observations; in traditional fictitious play, all observations are weighted equally. Weights can be calculated in more general ways, possibly giving more importance to the recent plays (Fudenberg et al., 1998) or having time-varying weights (Crawford, 1995). Placing the entire weight on the most recent play results in Cournot dynamics. Another possible variation to the basic framework is to choose different tie-breaking rules in the event of multiple optimal actions. Yet another possible variation is to consider different initial beliefs. Another possibility is to vary the basic framework by allowing the players to choose actions that are suboptimal (within a bound) with respect to their beliefs (Fudenberg & Kreps, 1993). Other generalizations are *stochastic fictitious play* (Fudenberg et al., 1998) in which players randomize when they are nearly indifferent between choices thereby allowing a best reply to be a mixed strategy, *dynamic fictitious play* (Shamma & Arslan, 2005) in which players use the best response to a forecasted opponent strategy, and *moderated fictitious play* (MacKay, 1992; Rezek et al., 2008) in which the probabilities over opponent's actions are moderated for possible errors and uncertainties. Finally, some generalizations (Kaniovski & Young, 1995) have considered incomplete information and stochastic perturbations.

[14] Brown (Brown, 1951) introduced fictitious play as an iterative method of solving an iterated unperturbed game, and Robinson (Robinson, 1951) proved convergence for 2 × 2 games. Fictitious play has since been examined in many articles including (Kaniovski & Young, 1995; Fudenberg & Levine, 1995; Fudenberg et al., 1998; Benaım & Hirsch, 1999) as a learning rule for perturbed (in a perturbed game (Harsanyi, 1973), payoffs vary randomly around a mean that defines the unperturbed or classical game) games.

[15] *No-regret learning* (Jafari et al., 2001) is a type of best-response learning whose behaviour closely resembles fictitious play in that, for many games including dilemmas, learning converges to a game-theoretic equilibrium.

### 5.3 Best-Reply Model: Adaptive Play

In fictitious play, the agents choose an action that is optimal considering the *entire history* of their opponent's actions. In contrast, in *adaptive play*, the agents base their decisions on limited information about actions of other agents in the *recent past*, and they do not always optimize. Young (1993) showed that, for general games, adaptive play need not converge to a Nash equilibrium, either pure or mixed strategies. However, for repeated play of a certain restricted class of coordination games with multiple Nash equilibria, he showed that regardless of the initial choice of strategies, there exists a sequence of best replies that converges to a strict, pure strategy Nash equilibrium. It cannot be said in advance which equilibrium will prevail, since this depends on the learning process and on the initial state.

### 5.4 Hybrid Models

The best-reply and RL approaches model different aspects of human cognition. Best-reply models start with the premise that players keep track of the history of previous play by other players and form some belief about what others will do in the future based on their past observations. Then they tend to choose a best-response, a strategy that maximizes their expected payoffs given the beliefs they formed. In contrast, RL assumes that strategies are *reinforced* by their previous payoffs, and the propensity to choose a strategy depends on its stock of reinforcements. Players who learn by reinforcement do not generally have beliefs about what other players will do. The information used by each approach is quite different. Best-reply models do not reflect past successes (reinforcements) of chosen strategies. RL models do not reflect the history of how others played. Further, RL models were primarily used by psychologists while best-reply models by game-theorists. These differences prompted a comparison of their descriptive powers for explaining human behaviour. The results of such comparative studies have generally been inconclusive; with RL appearing to do better in constant-sum games (Mookherjee & Sopher, 1997) while fictitious play in coordination games (Ho & Weigelt, 1996). However, the different comparative studies vary widely in terms of model specifications and parameter estimation techniques. These differences led to the development of more general hybrid models that include several different models as special cases. The hybrid models are parameterized such that its special cases can be simulated by suitable adjusting those parameters. This facilitates a more systematic comparison of the various models.

Cheung and Friedman (1997) constructed a one-parameter hybrid model that includes Cournot and fictitious play as special cases. In this model, any player $i$ observes a portion of the current outcome of the stage game and forms a belief $s_i$. What is observed as outcome depends on the institutional arrangements of the played game, for example, the observed outcome could be $i$'s own payoff or perhaps the entire combination of payoffs of the others. The learning rule is given by:

$$s_i(t+1) = \frac{s_i(t) + \sum\limits_{u=1}^{t-1} \gamma_i^u s_i(t-u)}{1 + \sum\limits_{u=1}^{t-1} \gamma_i^u} \tag{10}$$

where $\gamma_i$ denotes player $i$'s discount factor. Setting $\gamma = 0$ in yields the Cournot learning rule, and setting $\gamma = 1$ yields fictitious play. The case $0 < \gamma < 1$ is adaptive learning where all observations influence the state but the more recent observations have greater weight. The case $\gamma > 1$ implies that older observations have greater weight. Such values would characterize a player who relies on first impressions. The case $\gamma < 0$ is counter-intuitive in that it implies that the influence of a given observation changes sign each period. For $2 \times 2$ games, they defined the probability that player $i$ chooses the first action as:

$$P(a_{i,t} = 1 \mid r_{i,t} \cdot \alpha_i, \beta_i) = F(\alpha_i + \beta_i r_{i,t}) \tag{11}$$

where $r_{i,t}$ is the expected advantage of the first action given $s_i(t)$ and the payoff matrix for the stage game, $\beta_i$ is $i$'s degree of responsiveness to $r_{i,t}$ and her own idiosyncractic tendency $\alpha_i$ to favor the first action, and $F(x)$ is a cumulative distribution function on $(-\infty, \infty)$ such as a logistic function $F(x) = (1 + e^{-x})^{-1}$.

Camerer and Hua Ho (1999) and Camerer (2003) introduced a more generalized learning model which includes reinforcement learning and fictitious play as special cases and hybridizes their key elements. They called this model *experience-weighted attraction* (EWA). The key idea that underlies EWA is the actual relation between RL and fictitious play: best-reply models do not reflect past reinforcements and reinforcement models do not reflect the history of how others played. In RL, if player 1 (player 2) picks strategy $s_1^j$ $(s_2^k)$, then player 1's strategy $s_1^j$ is reinforced according to the payoff $\pi_1(s_1^j, s_2^k)$ while unchosen strategies $s_1^h(h \neq j)$ are not reinforced at all. In EWA, an expanded notion of reinforcements which includes foregone payoffs is used: unchosen strategies are reinforced based on a multiple $\delta$ of some hypothetical payoff $\pi_1(s_1^h, s_2^k)$ they would have earned. The model weights hypothetical payoffs that unchosen strategies would have earned by $\delta$, and weights the payoff actually earned from a chosen strategy by $1 - \delta$ so that the total weight is 1. Action probabilities are then calculated in terms of these reinforcements.

The EWA learning model (Camerer & Hua Ho, 1999) was subsequently extended in various ways. Camerer et al. (2002) extended it to capture sophisticated learning and strategic teaching. In the extended model, the players develop multi-period rather then single-period forecasts of others' behaviours. This was shown to perform better than RL and other belief models at fitting and predicting data pertaining to human behaviours in several games including dilemmas (Ho et al., 2007; Zhu et al., 2012).

The EWA was extended in another way taking into account the growing evidence suggesting that social norms are successful in the provision and maintenance of cooperation in everyday life (Fehr & Fischbacher, 2004). Given the role of norms in the emergence of cooperation, Realpe-Gómez et al. (2018a) introduced a cognitive-inspired model, called Experience Weighted Attraction with Norm Psychology (EWAN), that incorporates some key features of norm psychology into the EWA model (Camerer & Hua Ho, 1999). The EWAN model was shown in (Realpe-Gómez et al., 2018b) to support human cooperation in large-scale PD games on square lattices.

More recently, Vazifedan and Izadi (2023) generalized the EWA model by combining it with cognitive hierarchical models for learning in games (Camerer et al., 2003) and showed that their model describes and predicts human behaviour better than some existing models.

| | |
|---|---|
| **Stage game (SG) parameters** | |
| Payoff matrix | Deterministic or Non-deterministic |
| | Symmetric or Asymmetric |
| | Cooperation index |
| Horizon | Finite or indefinite |
| Nature of equilibria | Cooperative or non-cooperative |
| **Structural parameters** | |
| Types of learners | Homogeneous or heterogeneous |
| Partner matching | Exogenous or endogenous |
| | Static or dynamic |
| Partnership duration | Frequency of restructuring partnerships |
| **Individual player parameters** | |
| Cognitive | Memory bound |
| | Ability to recognize co-players |
| | Foresight |
| Strategic | Ability to create/ break partnerships |
| Psychological | Emotion, mood |

Table 7: Best-reply models: A summary of key parameters.

### 5.5 Best-Reply Models: Key Parameters

The key parameters pertaining to best-reply learning are as follows (see Table 7 for a summary):

1. **Type of stage game**: The stage game payoffs could be *deterministic* or *non-deterministic*, *symmetric* or *asymmetric* (Ahn et al., 2007). A lack of determinism can result in lower cooperation rates, as demonstrated in (Poppe, 1980) for the case of probabilistic payoff matrices used in laboratory experiments with the Prisoner's Dilemma. Another aspect related to stage game is the *cooperation index.* Cooperation index, a function of the payoffs $R$, $T$, $P$, and $S$, is a useful predictor of cooperation in humans (see Section 6 for details).

2. **The horizon**: Game-theoretic predictions for repeated games crucially depend on whether a game is *finitely* or *infinitely* repeated: in a finitely repeated dilemma game, cooperation usually cannot occur (Luce & Raiffa, 1989), but in the infinite case, some equilibria generate cooperative choices while some generate individualistic choices (Roth & Murnighan, 1978). Given the significance of horizon[16] in theory, numerous studies (Engle-Warnick & Slonim, 2006; Camera & Casari, 2009; Dal Bó &

[16] Horizon is relevant only to best-reply models. In contrast, EGT and RL models do not have horizon.

Fréchette, 2011; Sherstyuk et al., 2013; Fréchette & Yuksel, 2017) have investigated the influence of horizon in laboratory experiments with humans and also in simulations with learning models to show how the horizon can impact on cooperation rates. Many different horizon rules have been used with much disagreement about their pros and cons. Broadly, these rules can be divided into three main categories:

- A known finite horizon rule: The stage game is repeated a finite number of times and the players know in advance how many times the game will be repeated, as in (Flood, 1952; Rapoport et al., 1965).
- An unknown horizon rule: The players are not informed about the number of times the stage game will be repeated. The players only know that there will be a certain minimum number of periods they will play, but the actual number of periods is unknown, as in (Fouraker & Siegel, 1963).
- A random stopping rule: In each time period, there is a non-zero probability $p$ that the game will be played again in the next time period (i.e., each period has probability $1 - p$ of being the last), and $p$ is made known to the players (Roth & Murnighan, 1978; Axelrod, 1980; Fréchette & Yuksel, 2017). This is known as *indifinitely* repeated game, and for experiments conducted in the laboratory, this is a very common way of implementing an infinite horizon since it is impossible to experimentally study infinite horizon games[17]. Another type of design (Sabater-Grande & Georgantzis, 2002; Cabral et al., 2014) has also been used to implement infinite horizon: a fixed number of rounds are played with certainty with payoffs that are exponentially discounted at rate $p$, after those rounds, a random termination rule with probability $p$ of continuation is used. Now, for an indefinitely repeated game with continuation probability $p$, the length $T$ of the repeated game is a random variable with expected value $\mathbb{E}(T) = \frac{1}{1-p}$ and standard deviation $\sqrt(\frac{p}{(1-p)^2})$. In standard theory, only the expected match length should matter for behaviour in indefinitely repeated dilemma games with continuation probability $p$, and match length realizations should be irrelevant for behaviour (Mengel et al., 2022). However, given that match length realizations will be small in practice, several studies have focussed on understanding whether match length realization influences behaviour. These investigations (Murnighan & Roth, 1983; Blonski et al., 2011; Dal Bó & Fréchette, 2018; Mengel et al., 2022) have demonstrated that the sequence of match length realizations has a substantial effect on cooperation in dilemma games. Further, in their experiments with humans, Normann and Wallace (2012) compared cooperation rates in the iterated PD game for three different horizon rules: known, unknown, and random, and showed that cooperation rates increased significantly with the expected length of the game. Numerous other studies (Battalio et al., 2001; Bó, 2005; Camera & Casari, 2009; Dal Bó & Fréchette, 2011; Fréchette & Yuksel, 2017; Dal Bó & Fréchette, 2018; Bernard et al., 2018; Embrey et al., 2018; Mengel et al., 2022) have confirmed the dependence of cooperation on the horizon.

[17] See (Dal Bó & Fréchette, 2018) for a recent survey on cooperation in infinitely repeated games.

3. **The nature of equilibrium**: In the infinitely repeated play (usually implemented with a certain continuation probability $p$) of the Prisoner's Dilemma some equilibria generate cooperative choices, while some generate individualistic choices. For example, (Roth & Murnighan, 1978) showed that there are cooperative equilibria which may generate cooperative choices by both players at every period if $p \geq (T - R)/(T - P)$, but no such equilibria exist if $p < (T - R)/(T - P)$. Several studies on laboratory experiments with PD games (Roth & Murnighan, 1978; Fréchette & Yuksel, 2017), showed that subjects made the cooperative choice more frequently when cooperation formed an equilibrium strategy.

4. **Types of learners**: The interactants in a dilemma may be *homogeneous* in that they employ the same learning algorithm or *heterogenous* with different individuals employing different learning algorithms.

5. **Partner matching**: A matching rule prescribes how the players in a population will be paired for playing a stage game. This can be prescribed *exogenously* or decided *endogenously* (Kandori, 1992; Hauk, 2001) by the players. Further, the links may be *static* or may vary *dynamically* during play.

6. **Partnership duration**: Once matching is done, the duration of play can be varied for the matched pairs. The frequency of restructuring partnerships effects the behaviours that individuals learn.

7. **Individual traits**: Individuals may vary in terms of their *cognitive abilities* such as their memory bound, their ability to recognize co-players, and the degree of foresight (i.e., looking ahead into the future to decide the current optimal action). Further, there may be variations in *strategic abilities* relating to partner choice, and how their *psychological traits* such as emotion and mood combine with their rationality.

| | |
|---|---|
| Strategic<br>(Individual) | Reward/ punishment |
| | Trust and reputation |
| | Communication |
| | Reciprocity |
| Structural<br>(Interaction structure) | Fixed exogenously |
| | Endogenously evolving |
| Psychological<br>(Individual) | Emotion |
| | Mood |
| | Intelligence |

Table 8: Best-reply models: A taxonomized summary of mechanisms.

### 5.6 Mechanisms for Incentivizing Coopertion

In the context of best-reply models, many different mechanisms have been shown to enhance cooperation. These can be categorised into *strategic*, *structural*, and *psychological* as shown in Table 8. Some of the mechanisms may belong to more than one category.

**Strategic:** Individuals strategically employ the following mechanisms:

1. **Reward/punishment**: In the context of computer-computer interactions, Baumann et al. (2020) introduced an external agent for promoting cooperation between agents learning to play dilemma games by actor-critic reinforcement learning. The external agent can observe the actions of the players and distribute (positive or negative) rewards to the players after observing their actions, so as to guide the learners to a socially desirable outcome. In addition to the reward from the external agent, the players also receive a reward as per the payoff matrix for the dilemma game. Empirical results showed that their model guides the learners to the socially preferred outcome of mutual cooperation in PD, Chicken, and Stag Hunt games.

2. **Trust/ reputation**: In an attempt to resolve the finitely repeated PD in the context of rational self-interested behaviour, Kreps et al. (1982) introduced incomplete information about one or both players' options, motivation or behaviour, and analytically showed how reputation effects due to information asymmetries can generate significant levels of cooperation in sequential equilibrium. Ahn et al. (2001) examined cooperative behaviour in one-shot PD games as a function of the payoff structure of the PD games and the history of prior play in a series of Stag Hunt coordination games. In their experiments with human subjects, they found that the history of prior play in coordination games is a good predictor of cooperation in the PD games. In particular, the importance of history was significantly more pronounced if players were matched repeatedly with the same person, due to trust and reputation effects, compared to random matching. In a similar vein, Ivanov et al. (2023) used a hybrid approach. They used RL together with a trusted mediator, who can collect information and act on behalf of the dilemma game players, to achieve cooperation in their computer simulations.

3. **Communication**: Several studies have shown that communication is a key determinant of cooperation. This finding was confirmed for spoken as well as written communication. For example, Steinfatt (1973) showed that spoken communication promoted cooperation in laboratory experiments with humans playing a repeated PD. Lindskold and Finch (1981) experimentally studied the repeated PD allowing the participants to communicate by sending hand written notes after playing some initial rounds. Many other studies (Balliet, 2010; Kagel, 2018) showed a mixed bag of results possible through communication, with the main message being that cooperation increases with the effectiveness of communication in inducing trust. Trust is induced not just by sending conciliatory notes but by demonstration of cooperative choices. Crandall et al. (2018) examined an RL approach that employs computing a variety of expert strategies optimized for a range of 2 × 2 games that include dilemmas. A meta-strategy is used by agents to select an expert to follow. Agents are allowed to

exchange non-binding signals with co-players. This model was shown to cooperate with people and also with a wide range of other learning algorithms. All the above works are based on the assumption that the environment in which a dilemma is played is noise-free, i,e., that the players' actions can never be mis-executed. Rogers et al. (2007) relaxed this assumption and introduced a method that uses communication between players for enhancing cooperation in a noisy IPD.

4. **Reciprocity**: The effectiveness of best-response models at solving a dilemma depends on who a best-responder is paired with. If the interactants use different learning models, cooperation is hard to achieve. However, in self-play, it is possible for best responders to reciprocate and thereby learn to cooperate. Several studies (Smale, 1980; Kendall et al., 2007; Banerjee & Sen, 2007; Han et al., 2011; Foerster et al., 2018; Willi et al., 2022) have confirmed the efficacy of best-response in self-play. Most of the literature on learning in repeated dilemma games has focused on having one single game played in each stage. Thus, learning takes place specifically for the chosen stage game. In contrast, LiCalzi and Mühlenbernd (2022) took a broader perspective and studied best-reply learning across similar games (Mengel, 2012). In this learning approach, the games are segregated into partitions, as distinguishing all games can be too costly (require too much reasoning resources), the higher the cardinality of the partition, the greater the reasoning cost. The learning approach allows agents to learn to categorize games such that they tend to play the same action for games placed in the same category. This category based learning was shown to fit empirical data better than competitor models from the literature.

**Structural:** The interaction structure can be fixed exogenously or varied endogenously to facilitate cooperation:

1. **Exogenous structure**: Several studies have shown how cooperation can emerge in dilemmas when the structure is defined exogenously. For the Prisoner's Dilemma, Johnson et al. (1998) analytically investigated a best-reply model based on fictitious play in which the players choose optimal strategies with probability less than one. The players of a population are matched randomly to play the PD game. The payoff matrix is given by $S = 0$, $T = R + 1$, and $P = 1$, i.e., $R$ is the only variable. Under the assumptions that all the players have a discount factor $\delta$, and all the players have access to global historical information, they concluded that cooperation emerges for sufficiently large $R$ and sufficiently small $\delta$. Airiau et al. (2014) used fictitious play (Fudenberg et al., 1998), Q-learning (Watkins & Dayan, 1992), and WoLF (Bowling & Veloso, 2002) to show the emergence of conventions in social dilemma games such as those that arise when two drivers arrive simultaneously at an intersection. The agent interactions are given by a fixed network topology. While Airiau et al. (2014) showed that an agent's own experience is sufficient for the emergence of conventions, prior work (Epstein, 2001; Kandori & Rob, 1995; Young, 1993) had shown the same result but required agents to have knowledge about non-local interactions between other agents. Duffy and Ochs (2009) showed that cooperation is supported in laboratory experiments with humans for PD games for certain matching rules. In their experiments, they compared the levels of cooperation for *random matching* and *fixed*

*matching* in PD games with a high continuation probability $p$ and showed that cooperation is supported in sequential equilibrium. They found that cooperation increases as subjects gain more experience under fixed matching but not under random matching, suggesting that random matching tends to suppress the inclination of subjects to treat all stage games in a given session as a single supergame. Several other studies (Camera & Casari, 2009; Bigoni et al., 2013) have also shown how cooperation can be sustained in dilemmas for exogenously defined structures.

2. **Endogenously evolving structure**: Kandori (1992) considered situations where each player carries a reputation, defectors are sanctioned, and the players change their partners dynamically over time. He showed analytically how a population of players can sustain cooperation in repeated PD games even when each individual knows nothing more than his personal experience, and how the population can realize any mutually beneficial outcome when each agent carries a reputation and reputations are revised in a systematic way. Hauk (2001) studied a choice-refusal mechanism for boundedly rational agents using an individual learning approach. The agents are capable of remembering the past behaviour of their opponents and adjusting their behaviour accordingly. The agents use an endogenously varying payoff tolerance level for each possible pairing. Partners are accepted or rejected on the basis of tolerance level. The model was shown to converge to stable cooperative behaviour.

**Psychological:** Certain psychological traits pertaining to individuals can influence cooperation rates:

1. **Emotion**: Nobel Laureate Herbert Simon (Simon, 1983) noted that, in order to understand human rationality, we have to understand what role emotion plays in it. Recent research (Frank, 1988; Van Kleef et al., 2010; De Melo et al., 2014) provides evidence that affect is influential in shaping human decision making. These findings led to new decision-making models (Lerner et al., 2015; De Melo et al., 2014; de Melo & Terada, 2019, 2020) that synthesize traditional rational choice inputs and emotional inputs. These new models have been used in experimental investigations of the effect of emotion on human behaviours in dilemma games. The emotion component is then used to disambiguate the intentions of one's counterpart. Results of these works support the correlation between an individual's expectation of other's behaviour based on their observed emotion and actual cooperation in dilemmas. Studies (de Melo & Terada, 2019) of human-human and human-machine interactions in dilemma games have shown that humans cooperated just as much with humans as with a machine's virtual face that expressed cooperative emotion (e.g., joy following cooperation as opposed to joy following exploitation). All these findings have implications for dilemma resolution; it is therefore critical to understand methods for reasoning with emotions and methods for conveying intent through emotion.

2. **Mood**: Grujić et al. (2010) conducted laboratory experiments with human subjects playing a lattice-structured PD game. They fitted an agent-based learning model to their experimental data and explained human behaviours as being one of three types: *almost always defect*, *almost always cooperate*, and *moody conditional cooperator* (MCC). An MCC's propensity to cooperate in any round depends on the number of

cooperative neighbors observed in the previous round, and their own mood modelled in terms of their own action in the previous round. These findings were confirmed in (Gracia-Lázaro et al., 2012b, 2012a; Grujić et al., 2014).

3. **Intelligence**: Several studies considered the relationship between human cognitive abilities (such as their information processing ability, memory, and emotional intelligence) and their behaviour in PD games. Pincus and Bixenstine (1977) investigated the effect of information about the PD payoff matrix on human behaviours in a laboratory setting. The study revealed the same payoff matrix in different formats to the participants. One format was the standard PD matrix while the other was a decomposed PD matrix such that the summed payoffs of the components were equal to the standard matrix. Their findings suggest that the effect on cooperation produced by the decomposed format is due to the revealing of information that is not readily grasped for the standard matrix. Pincus and Bixenstine (1979) showed that subjects above the median on abstract information-processing ability, quantitative ability, and verbal ability were more likely to achieve a cooperative resolution of the PD than those below the median on these variables. Sela and Herreiner (1999) examined the effect of memory; using fictitious play with bounded and unbounded recall for pure coordination games (a pure coordination game is one for which payoffs off the diagonal are zero), it was shown that players with unbounded recall coordinated almost surely against their own type as well as against players with bounded recall. For PD games, Fernández-Berrocal et al. (2014) showed that individuals with a high emotional intelligence score are not pre-disposed to cooperate but are able to respond flexibly to others' strategies in order to maximize long-term gains.

In short, best-reply models (especially the hybrid ones), owing to their higher level of rationality relative to the other learning approaches, are not only useful for dilemma resolution, but have been shown to effectively explain human behaviour in experiments pertaining to dilemma games. As such, these models can be particularly useful for the resolution of dilemmas arising in human-machine interactions.

## 6. Cooperation Indices

In the context of the learning methods studied in the previous sections, focus is on building support mechanisms for promoting cooperation in dilemmas. In contrast, the topic of *cooperation indices* is concerned with the understanding how the exogenous payoff structure of a dilemma game impacts human cooperative behaviour. Human behaviour does not match game-theoretic rational behaviour. People's decisions can be more socially oriented. When individuals have social preferences, they derive utility from the positive payoffs that other decision makers receive. Recall that a dilemma is not just one game but a class of games that satisfy certain properties. For example, PD is that class of games that satisfy $T > R > P > S$. Different PD games have different relative payoffs and when the decision makers hold heterogeneous social preferences, different games yield different behaviors. Not all games in a class are equal in terms of elicited behaviour; all else being equal, cooperation may be higher in certain games relative to others in the same class (Moisan et al., 2018). It is therefore important to understand how the exogenous parameters, i.e., the payoffs

$R$, $T$, $P$, and $S$ of a dilemma game relate to cooperative behaviour. To this end, several cooperation indices were proposed.

A cooperation index is some function of $R$, $T$, $P$, $S$, and the horizon, defined to be a predictor of cooperation. The idea of cooperation index is appealing because it is simple; cooperation can be predicted only in terms of these exogenous parameters. A variety of cooperation indices were suggested. These are listed in Table 9.

Rapoport et al. (1965) proposed two indices of cooperation : $r_1$ and $r_2$. Subsequently, numerous other indices were introduced. While most of these are defined in terms of payoffs alone, the index $g$ and the index $BAD$ (*basin of attraction for defection*) consider also the number of times a dilemma is repeated. These indices differ in terms of how they combine various aspects of a dilemma game. The key aspects may be abstracted as follows:

- The *risk* that a dilemma presents; i.e., the loss in unilaterally cooperating against a defector.
- The *temptation* that a dilemma presents; i.e., the gain in unilaterally defecting against a cooperator.
- The *efficiency*, i.e., how much can be gained by mutual co-operation as opposed to mutual defection.

Owing to differences in how these aspects are combined, the indices differ in terms of their usefulness for explaining human cooperation in dilemma games. These differences motivated a comparison of the indices in terms of their effectiveness as predictors of cooperation. The findings of this research may be summarized as follows.

Rapoport et al.'s (1965) preliminary analysis of $r_1$ and $r_2$, as well as the analysis by others (Steele & Tedeschi, 1967) supported the predictions of $r_1$ in repeated PD games. Wyer's (1969) comparative evaluation of Thibaut and Kelly's (Thibaut & Kelly, 1959) and Harris's indices (Harris, 1969) is more supportive of the former. Murnighan and Roth (1983) correlated various indices to cooperative choices made by human subjects playing indefinite horizon PD games and showed that, as $r_1$, $r_2$, $k_1$, $k_2$, $k_4$, and $g$ increase (or $r_4$, $e_1$, $e_2$ and $k_3$ decrease), the percentage of mutual cooperative choices decreases. Further, the cooperation rates for $p > g$ exceeded those for $p \leq g$. More recently, Mengel (2018) defined two indices *risk* and *temptation* (as listed in Table 9). Treating these as two separate dimensions (observe from Table 9 that this is in contrast to several other indices that combine these two aspects), she showed that risk better explains cooperation rates in experimental data obtained from humans playing one-shot PD, while temptation and several other indices are better for finitely repeated PD games in which there is no pre-play communication.

## 7. Summary and Avenues for Future Research

Given its inherent challenge and its breadth of applicability, the dilemma resolution problem has been subject of study across a range of scientific disciplines and the problem has been investigated from many different angles: three different models of boundedly rational learning, viz., evolutionary games, reinforcement learning, and best-reply learning, have been investigated for the study of the dilemma problem. All of this research has resulted in a substantial improvement in our understanding of the underlying challenges and led to the

| Index | References |
|---|---|
| $k_1 = (R+S-T-P)$<br>$k_2 = (R-S+T-P)$<br>$k_3 = (R-S-T+P)$<br>$k_4 = (R+T+P+S)$ | (Thibaut & Kelly, 1959; Wyer, 1969) |
| $r_1 = (R-P)/(T-S)$<br>$r_2 = (R-S)/(T-S)$ | (Rapoport et al., 1965; Rapoport, 1967) |
| $r_3 = (P-S)/(T-S)$<br>$r_4 = (T-R)/(T-S)$ | (Harris, 1969) |
| $e_1 = (T-R)/(R-P)$<br>$e_2 = (T-R)/(R-S)$<br>$g = (R-T+p(T-P))/(1-p)$ | (Roth & Murnighan, 1978; Murnighan & Roth, 1983) |
| $D_g = (T-R)$<br>$D_r = (P-S)$ | (Tanimoto & Sagara, 2007; Tanimoto, 2009; Wang et al., 2015; Arefin et al., 2020) |
| $BAD = (P-S)/((R-P) \times m + x)$<br>$x = 2P - S - T$ | (Dal Bó & Fréchette, 2011; Mengel, 2018; Dal Bó et al., 2021) |
| $D'_g = (T-R)/(R-P)$<br>$D'_r = (P-S)/(R-P)$ | (Wang et al., 2015) |
| $Risk = (T-R)/T$<br>$Temptation = (P-S)/P$ | (Mengel, 2018) |
| $K = \alpha + \beta \times e_1 + \gamma \times r_3$<br>$0 < \alpha < 1$, $-1 < \beta < 0$, $-1 < \gamma < 0$ | (Ahn et al., 2001) |

Table 9: A summary of indices. The symbol $p$ denotes continuation probability, and $m$ the number of times a dilemma game is repeated. The labels for the indices are adopted from their respective references.

development of a range of methods for dilemma resolution. This article is a consolidated survey of the methods useful for mitigating the problem. The key insights that emerge from bringing the different pieces of literature together may be abstracted as follows.

**I1:** The three basic learning models (i.e., without the addition of any cooperation enhancing mechanisms) vary widely in terms of their cognitive demands. Evolutionary models are the least and best-reply the most demanding while the requirements for reinforcement learning lie in between these two extremes. Despite this difference, the three models are strikingly similar in the sense that, on their own, they are mostly insufficient for dilemma resolution. This is particularly true for evolutionary games and reinforcement models owing to their lower rationality relative to best-reply models. It is therefore necessary to support these learning methods to varying extents with some supplementary mechanisms for incentivising cooperation.

**I2:** The support mechanisms impose additional cognitive demands over and above those required by the underlying learning method. These mechanisms primarily require the ability to remember history, recognise co-players, and forecast the future. These additional requirements diminish the pre-existing cognitive differences between the learning methods.

**I3:** Although evolutionary games, reinforcement learning, and best-reply learning differ in terms of the modelling approach (individual versus aggregate population level), interestingly, there are major overlaps in the support mechanisms. For example, prosociality, sanctions, reputation, reciprocation, and the ability to choose partners are useful for promoting cooperation in all the three learning models.

**I4:** While learning models enable a study of the dynamics of interaction between boundedly rational behaviours for a given payoff matrix, cooperation indices are useful for understanding how the exogenous payoff structure of a dilemma game impacts cooperative behaviour. These indices are very simple functions of the payoffs $R$, $T$, $P$, and $S$, yet they can be effective indicators of cooperation.

**I5:** Regardless of the learning model, the proliferation of cooperation is a complex phenomenon that depends on the interplay between numerous parameters such as the dilemma game, the interaction structure, and the strategic behaviour of the participating individuals. This complexity makes it hard to find efficient algorithmic solutions to the dilemma problem. Although it is known that certain interaction structures promote the evolution of cooperation more than others, characterizing the set of cooperation enhancing structures is computationally hard (Ohtsuki et al., 2006; Ibsen-Jensen et al., 2015). Because an efficient algorithm does not exist, the existing studies have focused on special cases which pertain to specifically chosen parameters.

For the aforementioned reason, the existing models used are highly stylized; they use a single stage game such as a PD with specifically chosen parameters such as the payoffs, the information available to the players, and the matching rule. A number of aspects therefore need further investigation. In general, a more rigorous conceptual and practical understanding of the dynamics of learning is needed to confirm the robustness and general validity of the existing findings. In particular, the following open problems stand out:

**P1:** There are three different approaches to learning in games and it appears as though the three approaches are fundamentally different. Indeed these models differ in terms of the information that agents use and whether they optimize. Despite these differences, existing research has found some connections between the three approaches. For example, the learning trajectories for RL and replicator dynamics were shown to match for certain PD games (Kaisers & Tuyls, 2010). However, many of the findings that connect the different learning approaches are not readily relevant to dilemma games. For example, Tuyls et al. (2003) showed that by plotting the direction field of replicator dynamics, it is possible to better tune the parameters for reinforcement learning such that learning converges to Nash equilibrium. Hopkins (2002) drew similarities between RL and fictitious play in that both converge to a point that is close to Nash equilibrium. However, in the context of dilemmas, the challenge is to avoid deficient equilibria. The question therefore arises 'how do the the different learning methods inter-relate in the context of dilemmas and how can insights gained from one approach be utilized for addressing issues in others'? In this regard, hybrid models that encompass several individual models as special cases can be useful, especially those in which switching between different methods can be achieved by means of a simple parameter adjustment. Further, software tools such as OpenSpiel (Lanctot et al., 2019) can be greatly useful for doing an empirical comparative analysis.

**P2:** The existing research has modeled repeated interactions with a payoff matrix that remains constant over the iterations. However, in dynamic and uncertain environments, the payoff matrix may change in random ways during the course of an interaction. For example, the context of resource allocation dilemmas, the available resource and consequently the payoff matrix changes as the players take actions toward consuming the resources from a common pool. How can such dynamic and uncertain environments be modeled? A possibility is to use a *stochastic game* in which the payoff matrix is a part of state and there is uncertainty over the state transitions. Research on the use of stochastic games for the study of dilemmas has only just begun (Hilbe et al., 2018).

**P3:** Another avenue for future investigation is to address the issue of incomplete information about the payoffs for dilemma games. In the existing literature dilemmas, it is commonly assumed that the payoffs are exogenously known. However, in real-world dilemmas, payoffs may be unknown. How can the payoffs for a dilemma game be learnt? To this end, inverse reinforcement learning (IRL) seems to be a promising approach to investigate. However, a challenge is that existing IRL methods (Arora & Doshi, 2021; Fu et al., 2021) are based on the assumption that the underlying game has a unique equilibrium, but many dilemmas have multiple equilibria. A related aspect is the correctness of information pertaining to individual players. For example, reputation-based mechanisms proposed in the literature are useful for resolving dilemmas only if the reputations are true. The question that remains open is as follows. How can fake reputations be detected and avoided, and to what extent can noise in reputations be tolerated?

**P4:** Another key topic for future research is human-computer interactions in dilemmas. A majority of the existing literature on dilemmas is devoted to computer only tournaments or else human-human interactions. However, human-computer interactions are crucial for future applications such as autonomous driving; autonomous vehicles must be designed to interact and coordinate with human-driven cars. It remains to be seen how effective the computational learning models are at resolving dilemmas against humans.

**P5:** While existing research has uncovered various different mechanism for enhancing cooperation in dilemmas, little attention has been paid to how cooperation can be enhanced by *continual learning* (CL) (Khetarpal et al., 2022) and *transfer of learning* (ToL) (Da Silva & Costa, 2019). CL and ToL are vital in light of sociological theories (Blau, 1964) of cooperation that describe the development of cooperation as a slow process, starting with less risky interactions and eventually expanding into much riskier situations that require significantly higher levels of trust to generate cooperative behaviour. Trust building works because a counterpart's behaviour in the less risky situation serves as a precedent for their behaviour in the riskier situation. There is experimental evidence of these theories; studies of dilemmas with human participants not only confirmed that previous experience in a less risky PD enhances future cooperation in more risky PDs (Bettenhausen & Murnighan, 1991), but that cooperation also transfers from coordination dilemmas to PDs (Knez & Camerer, 2000). The question 'how can this human aspect be introduced into computational models of learning?' remains to be investigated.

**P6:** Much of the existing literature has focused on *homogeneous* individuals, i.e., all interactants use the same learning method. However, given that humans vary in their cognitive abilities (Boogert et al., 2018), a more reasonable approach is to study the evolution of cooperation for heterogenous learners. This knowledge will be particularly relevant to human-computer interactions.

**P7:** Although the existing literature has studied the relation between cooperation indices and actual cooperation rates in dilemmas in the context of best-reply models, we have not found any such studies for EGT and RL models. Thus there is scope for further enhancing our understanding of the relation between exogenously fixed payoff matrix and the EGT and reinforcement learning dynamics.

**P8:** The field has just started to explore issues such as the role of an individual's emotion, mood, and culture on their learning in social dilemmas. Further investigation is needed to understand how the interplay between *affect* and *rationality* impacts the evolution of cooperation.

**P9:** In the experimental literature on individual learning models such as RL, best-reply, and their hybrids, little attention has been paid to the problem of *heterogeneity bias*. The empirical comparisons typically begin with a pooled estimate, i.e., a single shared vector of learning model parameters is assumed for all subjects in a sample. However if subjects are in fact heterogeneous, pooled estimates tend to bias empirical comparisons in favor of a particular learning model (Wilcox, 2006; Salmon, 2001; Erev & Haruvy, 2005). Aside from the statistical shortcomings, prior research (Collaboration, 2015; Klein et al., 2018; Serra-Garcia & Gneezy, 2021) has raised concerns about the robustness and replicability of the results generated in the experimental economics and psychology literature, triggering attempts to develop agreed-upon best practices (Sánchez, 2018; Dreber & Johannesson, 2019; Muthukrishna & Henrich, 2019; Fréchette et al., 2022). Awareness of the crisis and the recommended best-practices is crucial for future research.

## Acknowledgements

We are very thankful to the reviewers, their detailed and insightful comments helped us to significantly improve the article.

## References

Adami, C., Schossau, J., & Hintze, A. (2016). Evolutionary game theory using agent-based methods. *Physics of life reviews*, *19*, 1–26.

Ahn, T.-K., Lee, M., Ruttan, L., & Walker, J. (2007). Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games. *Public Choice*, *132*(3), 353–366.

Ahn, T.-K., Ostrom, E., Schmidt, D., Shupp, R., & Walker, J. (2001). Cooperation in PD games: Fear, greed, and history of play. *Public Choice*, *106*(1), 137–155.

Airiau, S., Sen, S., & Villatoro, D. (2014). Emergence of conventions through social learning: Heterogeneous learners in complex networks. *Autonomous Agents and Multi-Agent Systems*, *28*(5), 779–804.

Aktipis, C. A. (2004). Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology*, *231*(2), 249–260.

Alexander, R. D. (2017). *The biology of moral systems.* Routledge.

Allen, B., Lippner, G., Chen, Y.-T., Fotouhi, B., Momeni, N., Yau, S.-T., & Nowak, M. A. (2017). Evolutionary dynamics on any population structure. *Nature*, *544*(7649), 227–230.

Allen, B., & Nowak, M. A. (2014). Games on graphs. *EMS surveys in mathematical sciences*, *1*(1), 113–151.

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The economic journal*, *103*(418), 570–585.

Arefin, M., Kabir, K., Jusup, M., Ito, H., Tanimoto, J., et al. (2020). Social efficiency deficit deciphers social dilemmas. *Scientific Reports*, *10*(1), 1–9.

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, *297*, 103500.

Ashlock, D., Smucker, M. D., Stanley, E. A., & Tesfatsion, L. (1996). Preferential partner selection in an evolutionary study of prisoner's dilemma. *BioSystems*, *37*(1-2), 99–125.

Axelrod, R. (1980). More effective choice in the prisoner's dilemma. *Journal of conflict resolution*, *24*(3), 379–403.

Axelrod, R. (1990). *The Evolution of Cooperation.* Penguin.

Axelrod, R. (1997). The complexity of cooperation. In *The Complexity of Cooperation.* Princeton University Press.

Babes, M., Munoz de Cote, E., & Littman, M. L. (2008). Social reward shaping in the prisoner's dilemma. In *Proceedings of AAMAS*, pp. 1389–1392.

Bala, V., & Goyal, S. (2001). Conformism and diversity under social learning. *Economic theory*, *17*(1), 101–120.

Bales, R. F., Flood, M. M., & Householder, A. S. (1952). *Some group interaction models*. Rand Corporation.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39–57.

Bandyopadhyay, A., & Kar, S. (2018). Coevolution of cooperation and network structure in social dilemmas in evolutionary dynamic complex network. *Applied Mathematics and Computation*, *320*, 710–730.

Banerjee, D., & Sen, S. (2007). Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems*, *15*(1), 91–108.

Bara, J., Turrini, P., & Andrighetto, G. (2022). Enabling imitation-based cooperation in dynamic social networks. *Autonomous Agents and Multi-Agent Systems*, *36*.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Basu, K. (1977). Information and strategy in iterated prisoner's dilemma. *Theory and Decision*, *8*(3), 293.

Battalio, R., Samuelson, L., & Van Huyck, J. (2001). Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica*, *69*(3), 749–764.

Bauch, C. T., & Earn, D. J. (2004). Vaccination and the theory of games. *Proceedings of the National Academy of Sciences*, *101*(36), 13391–13394.

Baumann, T., Graepel, T., & Shawe-Taylor, J. (2020). Adaptive mechanism design: Learning to promote cooperation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE.

Bazzan, A. L., Peleteiro, A., & Burguillo, J. C. (2011). Learning to cooperate in the iterated prisoner's dilemma by means of social attachments. *Journal of the Brazilian computer society*, *17*(3), 163–174.

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory*, *122*(1), 1–36.

Bellman, R. E., & Dreyfus, S. E. (1962). Applied dynamic programming. Tech. rep., The RAND Corporation.

Belloc, M., Bilancini, E., Boncinelli, L., & D'Alessandro, S. (2019). Intuition and deliberation in the stag hunt game. *Scientific Reports*, *9*(1), 1–7.

Benaım, M., & Hirsch, M. W. (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, *29*(1-2), 36–72.

Bendor, J., Mookherjee, D., & Ray, D. (1994). Aspirations, adaptive learning and cooperation in repeated games. Tech. rep., Planning Unit, Indian Statistical Institute, New Delhi.

Bendor, J., Mookherjee, D., & Ray, D. (2001a). Aspiration-based reinforcement learning in repeated interaction games: An overview. *International Game Theory Review*, *3*(2–3), 159–174.

Bendor, J., Mookherjee, D., & Ray, D. (2001b). Reinforcement learning in repeated interaction games. *The BE Journal of Theoretical Economics*, *1*(1).

Berger, U. (2007). Brown's original fictitious play. *Journal of Economic Theory*, *135*(1), 572–578.

Berger, U., & Grüne, A. (2016). On the stability of cooperation under indirect reciprocity with first-order information. *Games and Economic Behavior*, *98*, 19–33.

Bernard, M., Fanning, J., & Yuksel, S. (2018). Finding cooperators: Sorting through repeated interaction. *Journal of Economic Behavior & Organization*, *147*, 76–94.

Bettenhausen, K. L., & Murnighan, J. K. (1991). The development of an intragroup norm and the effects of interpersonal and structural challenges. *Administrative Science Quarterly*, *36*(1), 20–35.

Biely, C., Dragosits, K., & Thurner, S. (2007). The prisoner's dilemma on co-evolving networks under perfect rationality. *Physica D: Nonlinear Phenomena*, *228*(1), 40–48.

Bigoni, M., Camera, G., & Casari, M. (2013). Strategies of cooperation and punishment among students and clerical workers. *Journal of Economic Behavior & Organization*, *94*, 172–182.

Bitsch, F., Berger, P., Nagels, A., Falkenberg, I., & Straube, B. (2018). The role of the right temporo–parietal junction in social decision-making. *Human Brain Mapping*, *39*(7), 3072–3085.

Blau, P. M. (1964). *Exchange and power in social life.* NJ: Transaction Publishers.

Bloembergen, D., Tuyls, K., Hennes, D., & Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, *53*, 659–697.

Blonski, M., Ockenfels, P., & Spagnolo, G. (2011). Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics*, *3*(3), 164–192.

Bó, P. D. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American economic review*, *95*(5), 1591–1604.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (2018). Measuring and understanding individual differences in cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170280.

Boone, R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior*, *27*(3), 163–182.

Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, *64*(1), 265–276.

Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, *77*(1), 1–14.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical population biology*, *65*(1), 17–28.

Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, *136*(2), 215–250.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535.

Brafman, R., & Tennenholtz, M. (2004). Efficient learning equilibrium. *Artificial intelligence*, *159*(1-2), 27–47.

Brafman, R. I., & Tennenholtz, M. (2002). R-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*(Oct), 213–231.

Brandt, H., & Sigmund, K. (2004). The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology*, *231*(4), 475–486.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, *13*(1), 374.

Burguillo-Rial, J. C. (2009). A memetic framework for describing and simulating spatial prisoner's dilemma with coalition formation. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 441–448.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning.. *Psychological review*, *58*(5), 313.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.* John Wiley & Sons, Inc.

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

Busoniu, L., Babuska, R., & Schutter, B. D. (2010). Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications*, pp. 183–221. Springer.

Cabral, L., Ozbay, E. Y., & Schotter, A. (2014). Intrinsic and instrumental reciprocity: An experimental study. *Games and Economic Behavior*, *87*, 100–121.

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2019). Algorithmic pricing what implications for competition policy?. *Review of industrial organization*, *55*, 155–171.

Calvano, E., Calzolari, G., Denicolo, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, *110*(10), 3267–3297.

Camera, G., & Casari, M. (2009). Cooperation among strangers under the shadow of the future. *American Economic Review*, *99*(3), 979–1005.

Camerer, C., Ho, T., & Chong, K. (2003). Models of thinking, learning, and teaching in games. *American Economic Review*, *93*(2), 192–195.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, *104*(1), 137–188.

Capraro, V., Venanzi, M., Polukarov, M., & Jennings, N. R. (2013). Cooperative equilibria in iterated social dilemmas. In *International Symposium on Algorithmic Game Theory*, pp. 146–158. Springer.

Cardillo, A., Gómez-Gardeñes, J., Vilone, D., & Sánchez, A. (2010). Co-evolution of strategies and update rules in the prisoner's dilemma game on complex networks. *New Journal of Physics*, *12*(10), 103034.

Casasnovas, J. P. (2012). *Evolutionary Games in Complex Topologies: Interplay Between Structure and Dynamics*. Springer Science & Business Media.

Cassar, A. (2007). Coordination and cooperation in local, random and small world networks: Experimental evidence. *Games and Economic Behavior*, *58*(2), 209–230.

Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*(4), 1009–1040.

Chammah, A. M. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor, U. of Michigan P.

Chen, W., Wang, J., Yu, F., He, J., Xu, W., & Wang, R. (2021). Effects of emotion on the evolution of cooperation in a spatial prisoner's dilemma game. *Applied Mathematics and Computation*, *411*, 126497.

Chen, X., Sasaki, T., Brännström, Å., & Dieckmann, U. (2015). First carrot, then stick: How the adaptive hybridization of incentives promotes cooperation. *Journal of the Royal Society Interface*, *12*(102), 20140935.

Chen, X., & Wang, L. (2008). Promotion of cooperation induced by appropriate payoff aspirations in a small-world networked game. *Physical Review E*, *77*(1), 017103.

Chen, Y., & Tang, F.-F. (1998). Learning and incentive-compatible mechanisms for public goods provision: An experimental study. *Journal of Political Economy*, *106*(3), 633–662.

Cheng, H., Li, H., Dai, Q., Zhu, Y., & Yang, J. (2010). Motion depending on the strategies of players enhances cooperation in a co-evolutionary prisoner's dilemma game. *New Journal of Physics*, *12*(12), 123014.

Cheung, Y.-W., & Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, *19*(1), 46–76.

Cimini, G., & Sánchez, A. (2014). Learning dynamics explains human behaviour in prisoner's dilemma on networks. *Journal of The Royal Society Interface*, *11*(94), 20131186.

Coleman, J. S. (1994). *Foundations of social theory.* Harvard University Press.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Collenette, J., Atkinson, K., Bloembergen, D., & Tuyls, K. (2017a). Environmental effects on simulated emotional and moody agents. *The Knowledge Engineering Review*, *32*.

Collenette, J., Atkinson, K., Bloembergen, D., & Tuyls, K. (2017b). Mood modelling within reinforcement learning. In *ECAL 2017, the Fourteenth European Conference on Artificial Life*, pp. 106–113. MIT Press.

Colman, A. M. (1982). *Game theory and its applications: In the social and biological sciences.* Psychology Press.

Colyvan, M., Justus, J., & Regan, H. M. (2011). The conservation game. *Biological conservation*, *144*(4), 1246–1253.

Cournot, A.-A. (1838). *Researches sur les principes mathématiques del la théorie des richesses.* Paris: Hachette. Translated into English by N. Bacon as *Researches in the Mathematical Principles of the Theory of Wealth.* London: Haffner, 1960.

Crandall, J. W., & Goodrich, M. A. (2005). Learning to compete, compromise, and cooperate in repeated general-sum games. In *Proceedings of the 22nd international conference on Machine learning*, pp. 161–168.

Crandall, J. W., & Goodrich, M. A. (2011). Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Machine Learning*, *82*, 281–314.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature communications*, *9*(1), 233.

Crawford, V. P. (1995). Adaptive dynamics in coordination games. *Econometrica*, *63*(1), 103–143.

Cross, J. G. (1973). A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, *87*(2), 239–266.

Da Silva, F. L., & Costa, A. H. R. (2019). A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, *64*, 645–703.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature*, *593*(7857), 33–36.

Dal Bó, P., & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, *101*(1), 411–29.

Dal Bó, P., & Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, *56*(1), 60–114.

Dal Bó, P., Fréchette, G. R., & Kim, J. (2021). The determinants of efficient behavior in coordination games. *Games and Economic Behavior*, *130*, 352–368.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. Princeton University Press.

Darwin, C. (1909). *The origin of species*. PF Collier & son New York.

Dawes, R. (1975). Formal models of dilemmas in social decision-making. In Kaplan, M., & Schwartz, S. (Eds.), *Human Judgement and Decision Processes*. New York, Academic Press.

Dawes, R. M. (1974). *Formal models of dilemmas in social decision-making*. Oregon Research Institute.

Dawes, R. M. (1980). Social dilemmas. *Annual review of psychology*, *31*(1), 169–193.

De Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of personality and social psychology*, *106*(1), 73.

de Melo, C. M., & Terada, K. (2019). Cooperation with autonomous machines through culture and emotion. *PloS one*, *14*(11), e0224758.

de Melo, C. M., & Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific reports*, *10*(1), 14959.

Débarre, F., Hauert, C., & Doebeli, M. (2014). Social evolution in structured populations. *Nature Communications*, *5*(1), 1–7.

Dekel, E., & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory*, *52*(2), 243–267.

Dercole, F., Della Rossa, F., & Piccardi, C. (2019). Direct reciprocity and model-predictive rationality explain network reciprocity over social ties. *Scientific Reports*, *9*(1), 1–13.

Díaz, J., Goldberg, L. A., Mertzios, G. B., Richerby, D., Serna, M., & Spirakis, P. G. (2014). Approximating fixation probabilities in the generalized moran process. *Algorithmica*, *69*(1), 78–91.

Díaz, J., & Mitsche, D. (2021). A survey of the modified moran process and evolutionary graph theory. *Computer Science Review*, *39*, 100347.

Dreber, A., & Johannesson, M. (2019). Statistical significance and the replication crisis in the social sciences. In *Oxford research encyclopedia of economics and finance*. Oxford University Press.

Droz, M., Szwabiński, J., & Szabó, G. (2009). Motion of influential players can support cooperation in prisoner's dilemma. *The European Physical Journal B*, *71*(4), 579–585.

Du, C., Guo, K., Lu, Y., Jin, H., & Shi, L. (2023). Aspiration driven exit-option resolves social dilemmas in the network. *Applied Mathematics and Computation*, *438*, 127617.

Duffy, J., & Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, *66*(2), 785–812.

Duong, M. H., et al. (2020). On equilibrium properties of the replicator–mutator equation in deterministic and random games. *Dynamic Games and Applications*, *10*(3), 641–663.

Ebel, H., & Bornholdt, S. (2002). Coevolutionary games on networks. *Physical Review E*, *66*(5), 056118.

Eguíluz, V. M., Zimmermann, M. G., Cela-Conde, C. J., & Miguel, M. S. (2005). Cooperation and the emergence of role differentiation in the dynamics of social networks. *American journal of sociology*, *110*(4), 977–1008.

Eimontaite, I., Schindler, I., De Marco, M., Duzzi, D., Venneri, A., & Goel, V. (2019). Left amygdala and putamen activation modulate emotion driven decisions in the iterated prisoner's dilemma game. *Frontiers in Neuroscience*, *13*, 741.

Embrey, M., Fréchette, G. R., & Yuksel, S. (2018). Cooperation in the finitely repeated prisoner's dilemma. *The Quarterly Journal of Economics*, *133*(1), 509–551.

Engle-Warnick, J., & Slonim, R. L. (2006). Learning to trust in indefinitely repeated games. *Games and Economic Behavior*, *54*(1), 95–114.

Epstein, J. M. (2001). Learning to be thoughtless: Social norms and individual computation. *Computational economics*, *18*(1), 9–24.

Erev, I., Bereby-Meyer, Y., & Roth, A. E. (1999). The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior & Organization*, *39*(1), 111–128.

Erev, I., & Haruvy, E. (2005). On the potential uses and current limitations of data driven learning models. *J Math Psychol*, *49*(5), 357–371.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, *88*(4), 848–881.

Erev, I., & Roth, A. E. (2002). Simple reinforcement learning models and reciprocation in the prisoner's dilemma game. In *Bounded rationality: The adaptive toolbox*, pp. 215–231. MIT Press.

Erev, I., & Roth, A. E. (2007). Multi-agent learning and the descriptive value of simple models. *Artificial Intelligence*, *171*(7), 423–428.

Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, *79*(4), 1331–1335.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological review*, *57*(2), 94–107.

Estes, W. K. (1967). Reinforcement in human learning. Tech. rep., Stanford University Stanford United States.

Ezaki, T., Horita, Y., Takezawa, M., & Masuda, N. (2016). Reinforcement learning explains conditional cooperation and its moody cousin. *PLoS Computational Biology*, *12*(7), e1005034.

Ezaki, T., & Masuda, N. (2017). Reinforcement learning account of network reciprocity. *PloS One*, *12*(12), e0189220.

Fan, L., Song, Z., Wang, L., Liu, Y., & Wang, Z. (2022). Incorporating social payoff into reinforcement learning promotes cooperation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(12), 123140.

Feehan, G., & Fatima, S. (2022). Augmenting reinforcement learning to enhance cooperation in the iterated prisoner's dilemma. In *Proceedings of ICAART (3)*, pp. 146–157.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in cognitive sciences*, *8*(4), 185–190.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fernández-Berrocal, P., Extremera, N., Lopes, P. N., & Ruiz-Aranda, D. (2014). When to cooperate and when to compete: Emotional intelligence in interpersonal decision-making. *Journal of Research in Personality*, *49*, 21–24.

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford at the Clarendon Press.

Flache, A., & Macy, M. W. (2002). Stochastic collusion and the power law of learning: A general reinforcement learning model of cooperation. *Journal of Conflict Resolution*, *46*(5), 629–653.

Flood, M. M. (1958). Some experimental games. *Management Science*, *5*(1), 5–26.

Flood, M. M. (1952). Testing organization theories. Tech. rep., RAND CORP SANTA MONICA CA.

Foerster, J., Chen, R., M, A.-S., Whiteson, S., P, A., & I, M. (2018). Learning with opponent-learning awareness. In *Proceedings of AAMAS*, pp. 122–130.

Fouraker, L. E., & Siegel, S. (1963). *Bargaining behavior*. McGraw-Hill.

Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, *107*(12), 5334–5338.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co.

Fréchette, G. R., Sarnoff, K., & Yariv, L. (2022). Experimental economics: Past and future. *Annual Review of Economics*, *14*, 777–794.

Fréchette, G. R., & Yuksel, S. (2017). Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination. *Experimental Economics*, *20*(2), 279–308.

Frijda, N. H., et al. (1986). *The emotions*. Cambridge University Press.

Fu, F., Hauert, C., Nowak, M. A., & Wang, L. (2008). Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*, *78*(2), 026117.

Fu, F., Nowak, M. A., & Hauert, C. (2010). Invasion and expansion of cooperators in lattice populations: Prisoner's dilemma vs. snowdrift games. *Journal of Theoretical Biology*, *266*(3), 358–366.

Fu, F., Rosenbloom, D. I., Wang, L., & Nowak, M. A. (2011). Imitation dynamics of vaccination behaviour on social networks. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1702), 42–49.

Fu, J., Tacchetti, A., Perolat, J., & Bachrach, Y. (2021). Evaluating strategic structures in multi-agent inverse reinforcement learning. *Journal of Artificial Intelligence Research*, *71*, 925–951.

Fudenberg, D., Drew, F., Levine, D. K., & Levine, D. K. (1998). *The theory of learning in games*, Vol. 2. MIT press.

Fudenberg, D., & Harris, C. (1992). Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory*, *57*(2), 420–441.

Fudenberg, D., & Kreps, D. M. (1993). Learning mixed equilibria. *Games and Economic Behavior*, *5*(3), 320–367.

Fudenberg, D., & Levine, D. K. (1995). Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, *19*(5-7), 1065–1089.

Galla, T., & Farmer, J. D. (2013). Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, *110*(4), 1232–1236.

Gokhale, C. S., & Traulsen, A. (2010). Evolutionary games in the multiverse. *Proceedings of the National Academy of Sciences*, *107*(12), 5500–5504.

Gotts, N. M., Polhill, J. G., & Law, A. N. R. (2003). Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review*, *19*, 3–92.

Gracia-Lázaro, C., Cuesta, J. A., Sánchez, A., & Moreno, Y. (2012a). Human behavior in prisoner's dilemma experiments suppresses network reciprocity. *Scientific Reports*, *2*(1), 1–4.

Gracia-Lázaro, C., Ferrer, A., Ruiz, G., Tarancón, A., Cuesta, J. A., Sánchez, A., & Moreno, Y. (2012b). Heterogeneous networks do not promote cooperation when humans play a prisoner's dilemma. *Proceedings of the National Academy of Sciences*, *109*(32), 12922–12926.

Grimm, V., & Mengel, F. (2011). Matching technology and the choice of punishment institutions in a prisoner's dilemma game. *Journal of Economic Behavior & Organization*, *78*(3), 333–348.

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, *55*, 1–49.

Gross, J., & De Dreu, C. K. (2019). The rise and fall of cooperation through reputation and group polarization. *Nature Communications*, *10*(1), 1–10.

Gross, T., Rudolf, L., Levin, S. A., & Dieckmann, U. (2009). Generalized models reveal stabilizing factors in food webs. *Science*, *325*(5941), 747–750.

Grujić, J., Fosco, C., Araujo, L., Cuesta, J. A., & Sánchez, A. (2010). Social experiments in the mesoscale: Humans playing a spatial prisoner's dilemma. *PloS One*, *5*(11), e13749.

Grujić, J., Gracia-Lázaro, C., Milinski, M., Semmann, D., Traulsen, A., Cuesta, J. A., Moreno, Y., & Sánchez, A. (2014). A comparative analysis of spatial prisoner's dilemma experiments: Conditional cooperation and payoff irrelevance. *Scientific Reports*, *4*(1), 1–9.

Gurerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108–111.

Gutiérrez-Roig, M., Gracia-Lázaro, C., Perelló, J., Moreno, Y., & Sánchez, A. (2014). Transition from reciprocal cooperation to persistent behaviour in social dilemmas at the end of adolescence. *Nature Communications*, *5*(1), 1–7.

Hamilton, I. M., & Taborsky, M. (2005). Contingent movement and cooperation evolve under generalized reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1578), 2259–2267.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, *97*(896), 354–356.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology*, *7*(1), 1–16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology*, *7*(1), 17–52.

Han, T., Pereira, L., & Santos, Santos, F. (2011). The role of intention recognition in the evolution of cooperative behavior. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1684–1689.

Han, T. A., Pereira, L. M., & Lenaerts, T. (2017). Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems*, *31*(3), 561–583.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2013). Why is it so hard to say sorry? Evolution of apology with commitments in the iterated Prisoner's Dilemma. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 177–183.

Harris, R. J. (1969). A geometric classification system for 2×2 interval-symmetric games. *Behavioral Science*, *14*(2), 138–146.

Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, *2*(1), 1–23.

Harsanyi, J. C., & Selten, R. (1988). *A general theory of equilibrium selection in games*. The MIT Press.

Hauert, C., De Monte, S., Hofbauer, J., & Sigmund, K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science*, *296*(5570), 1129–1132.

Hauert, C., & Doebeli, M. (2004). Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature*, *428*(6983), 643–646.

Hauert, C., & Szabo, G. (2003). Prisoner's dilemma and public goods games in different geometries: Compulsory versus voluntary interactions. *Complexity*, *8*(4), 31–38.

Hauk, E. (2001). Leaving the prison: Permitting partner choice and refusal in prisoner's dilemma games. *Computational Economics*, *18*(1), 65–87.

Hays, D. G., & Bush, R. R. (1954). A study of group action. *American Sociological Review*, *19*(6), 693–701.

Helbing, D., & Yu, W. (2008). Migration as a mechanism to promote cooperation. *Advances in Complex Systems*, *11*(04), 641–652.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Hernandez-Leal, P., Kaisers, M., Baarslag, T., & De Cote, E. M. (2017). A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, *0*.

Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 750–797.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, *4*(3), 267.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the experimental analysis of behavior*, *13*(2), 243–266.

Hertel, G., Neuhof, J., Theuer, T., & Kerr, N. L. (2000). Mood effects on cooperation in small groups: Does positive mood simply lead to more cooperation?. *Cognition & Emotion*, *14*(4), 441–472.

Hilbe, C., Šimsa, Š., Chatterjee, K., & Nowak, M. A. (2018). Evolution of cooperation in stochastic games. *Nature*, *559*(7713), 246–249.

Ho, T. H., Camerer, C. F., & Chong, J.-K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, *133*(1), 177–198.

Ho, T.-H., & Weigelt, K. (1996). Task complexity, equilibrium selection, and learning: An experimental study. *Management Science*, *42*(5), 659–679.

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, *40*(4), 479–519.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, *70*(6), 2141–2166.

Horita, Y., Takezawa, M., Inukai, K., Kita, T., & Masuda, N. (2017). Reinforcement learning accounts for moody conditional cooperation behavior: Experimental results. *Scientific Reports*, *7*(1), 1–10.

Hsiao, V., & Nau, D. (2022). A mean field game model of spatial evolutionary games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 606–614.

Huang, C., Dai, Q., & Li, H. (2018). Leaders should be more persistent in evolutionary social dilemmas. *EPL (Europhysics Letters)*, *124*(1), 18001.

Huberman, B. A., & Lukose, R. M. (1997). Social dilemmas and internet congestion. *Science*, *277*(5325), 535–537.

Ibsen-Jensen, R., Chatterjee, K., & Nowak, M. A. (2015). Computational complexity of ecological and evolutionary spatial dynamics. *Proceedings of the National Academy of Sciences*, *112*(51), 15636–15641.

Ifti, M., Killingback, T., & Doebeli, M. (2004). Effects of neighbourhood size and connectivity on the spatial continuous prisoner's dilemma. *Journal of Theoretical Biology*, *231*(1), 97–106.

Imhof, L. A., & Nowak, M. A. (2006). Evolutionary game dynamics in a wright-fisher process. *Journal of Mathematical Biology*, *52*(5), 667–681.

Irwin, A. J., & Taylor, P. D. (2001). Evolution of altruism in stepping-stone populations with overlapping generations. *Theoretical Population Biology*, *60*(4), 315–325.

Ivanov, D., Zisman, I., & Chernyshev, K. (2023). Mediated multi-agent reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 49–57.

Iwagami, A., & Masuda, N. (2010). Upstream reciprocity in heterogeneous networks. *Journal of Theoretical Biology*, *265*(3), 297–305.

Iyer, S., & Killingback, T. (2016). Evolution of cooperation in social dilemmas on complex networks. *PLoS Computational Biology*, *12*(2), e1004779.

Iyer, S., & Killingback, T. (2020). Evolution of cooperation in social dilemmas with assortative interactions. *Games*, *11*(4), 41.

Izquierdo, L. R., & Izquierdo, S. S. (2008). Dynamics of the Bush-Mosteller learning algorithm in 2x2 games. In Weber, C., Elshaw, M., & Mayer, N. M. (Eds.), *Reinforcement Learning*, chap. 11. IntechOpen, Rijeka.

Izquierdo, L. R., Izquierdo, S. S., Gotts, N. M., & Polhill, J. G. (2007). Transient and asymptotic dynamics of reinforcement learning in games. *Games and Economic Behavior*, *61*(2), 259–276.

Izquierdo, S. S., Izquierdo, L. R., & Gotts, N. M. (2008). Reinforcement learning dynamics in social dilemmas. *Journal of Artificial Societies and Social Simulation*, *11*(2), 1–22.

Jafari, A., Greenwald, A., Gondek, D., & Ercal, G. (2001). On no-regret learning, fictitious play, and nash equilibrium. In *ICML*, Vol. 1, pp. 226–233.

Jansen, V. A., & Van Baalen, M. (2006). Altruism through beard chromodynamics. *Nature*, *440*(7084), 663–666.

Jiang, L.-L., Wang, W.-X., Lai, Y.-C., & Wang, B.-H. (2010). Role of adaptive migration in promoting cooperation in spatial games. *Physical Review E*, *81*(3), 036108.

Johnson, P., Levine, D. K., Pesendorfer, W., et al. (1998). *Evolution and information in a prisoner's dilemma game*, Vol. 9805. ITAM, Centro de Investigación Económica.

Jusup, M., Holme, P., Kanazawa, K., Takayasu, M., Romić, I., Wang, Z., Geček, S., Lipić, T., Podobnik, B., Wang, L., et al. (2022). Social physics. *Physics Reports*, *948*, 1–148.

Kagel, J. H. (2018). Cooperation through communication: Teams and individuals in finitely repeated prisoners' dilemma games. *Journal of Economic Behavior & Organization*, *146*, 55–64.

Kahn, H. (2017). *On escalation metaphors and scenarios.* Routledge.

Kaisers, M., & Tuyls, K. (2010). Frequency adjusted multi-agent q-learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pp. 309–316.

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, *59*(1), 63–80.

Kandori, M., & Rob, R. (1995). Evolution of equilibria in the long run: A general theory and applications. *Journal of Economic Theory*, *65*(2), 383–414.

Kaniovski, Y. M., & Young, H. P. (1995). Learning dynamics in games with stochastic perturbations. *Games and Economic Behavior*, *11*(2), 330–363.

Karandikar, R., Mookherjee, D., Ray, D., & Vega-Redondo, F. (1998). Evolving aspirations and cooperation. *Journal of Economic Theory*, *80*(2), 292–331.

Keller, L. (1999). *Levels of selection in evolution*, Vol. 66. Princeton University Press.

Kendall, G., Yao, X., & Chong, S. Y. (2007). *The iterated prisoners' dilemma: 20 years on*, Vol. 4. World Scientific.

Khetarpal, K., Riemer, M., Rish, I., & Precup, D. (2022). Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, *75*, 1401–1476.

Killingback, T., Doebeli, M., & Hauert, C. (2010). Cooperation and defection in the tragedy of the commons. *Biological Theory*, *5*, 3–6.

Killingback, T., & Doebeli, M. (1996). Spatial evolutionary game theory: Hawks and doves revisited. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *263*(1374), 1135–1144.

Killingback, T., & Doebeli, M. (2002). The continuous prisoner's dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *The American Naturalist*, *160*(4), 421–438.

Kim, B. J., Trusina, A., Holme, P., Minnhagen, P., Chung, J. S., & Choi, M. (2002). Dynamic instabilities induced by asymmetric influence: prisoners' dilemma game in small-world networks. *Physical Review E*, *66*(2), 021907.

Kim, Y. (1999). Satisficing and optimality in 2×2 common interest games. *Economic Theory*, *13*(2), 365–375.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Knez, M., & Camerer, C. (2000). Increasing cooperation in prisoner's dilemmas by establishing a precedent of efficiency in coordination games. *Organizational Behavior and Human Decision Processes*, *82*(2), 194–216.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*, 183–214.

Konno, T. (2011). A condition for cooperation in a game on complex networks. *Journal of Theoretical Biology*, *269*(1), 224–233.

Kosfeld, M., & Riedl, A. (2004). The design of (de) centralized punishment institutions for sustaining cooperation. Tinbergen Institute Discussion Paper no. TI 04-025/1.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory*, *27*(2), 245–252.

Lahkar, R. (2017). Equilibrium selection in the stag hunt game under generalized reinforcement learning. *Journal of Economic Behavior & Organization*, *138*, 63–68.

Lanctot, M., Lockhart, E., Lespiau, J., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., & Ryan-Davis, J. (2019). Openspiel: A framework for reinforcement learning in games.. arXiv.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, 2017*, pp. 464–473.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, *66*, 799–823.

Li, A., Zhou, L., Su, Q., Cornelius, S. P., Liu, Y.-Y., Wang, L., & Levin, S. A. (2020). Evolution of cooperation on temporal networks. *Nature Communications*, *11*(1), 1–9.

LiCalzi, M., & Mühlenbernd, R. (2022). Feature-weighted categorized play across symmetric games. *Experimental Economics*, *25*(3), 1052–1078.

Lieberman, E., Hauert, C., & Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, *433*(7023), 312–316.

Liebrand, W. B. (1983). A classification of social dilemma games. *Simulation & Games*, *14*(2), 123–138.

Lindskold, S., & Finch, M. L. (1981). Styles of announcing conciliation. *Journal of Conflict Resolution*, *25*(1), 145–155.

Littman, M. L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature*, *521*(7553), 445–451.

Liu, C., Shi, J., Li, T., & Liu, J. (2019). Aspiration driven coevolution resolves social dilemmas in networks. *Applied Mathematics and Computation*, *342*, 247–254.

Liu, X., He, M., Kang, Y., & Pan, Q. (2016). Aspiration promotes cooperation in the prisoner's dilemma game with the imitation rule. *Physical Review E*, *94*(1), 012124.

Locodi, A., & O'Riordan, C. (2021). Introducing a graph topology for robust cooperation. *Royal Society Open Science*, *8*(5), 201958.

Lopez, C. R., Cihon, T. M., de Borba Vasconcelos Neto, A., & Becker, A. (2022). An exploration of cooperation during an asymmetric iterated prisoner's dilemma game. *Behavior and Social Issues*, *31*, 106–132.

Luce, R. D., & Raiffa, H. (1957). *Games and Decisions*. New York: John Wiley and Sons.

Luce, R. D., & Raiffa, H. (1989). *Games and Decisions: Introduction and Critical Survey*. Courier Corporation.

Lütz, A. F., Amaral, M. A., & Wardil, L. (2021). Moderate immigration may promote a peak of cooperation among natives. *Physical Review E*, *104*(1), 014304.

MacKay, D. J. (1992). The evidence framework applied to classification networks. *Neural Computation*, *4*(5), 720–736.

Macy, M. W. (1991). Learning to cooperate: Stochastic and tacit collusion in social exchange. *American Journal of Sociology*, *97*(3), 808–843.

Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, *99*(suppl_3), 7229–7236.

Mantas, V., Pehlivanidis, A., Kotoula, V., Papanikolaou, K., Vassiliou, G., Papaiakovou, A., & Papageorgiou, C. (2022). Factors of influence in prisoner's dilemma task: A review of medical literature. *PeerJ*, *10*, e12829.

Marsella, S., Gratch, J., Petta, P., et al. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, *11*(1), 21–46.

Marshall, J. A. (2011). Group selection and kin selection: formally equivalent approaches. *Trends in Ecology & Evolution*, *26*(7), 325–332.

Masuda, N. (2012). Evolution of cooperation driven by zealots. *Scientific Reports*, *2*(1), 1–5.

Masuda, N., & Ohtsuki, H. (2009). A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bulletin of Mathematical Biology*, *71*(8), 1818–1850.

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267–275.

Matsuzawa, R., Tanimoto, J., & Fukuda, E. (2016). Spatial prisoner's dilemma games with zealous cooperators. *Physical Review E*, *94*(2), 022114.

May, R. M. (2006). Network structure and the biology of populations. *Trends in Ecology & Evolution*, *21*(7), 394–399.

McAllister, P. H. (1991). Adaptive approaches to stochastic programming. *Annals of Operations Research*, *30*(1), 45–62.

Meloni, S., Buscarino, A., Fortuna, L., Frasca, M., Gómez-Gardeñes, J., Latora, V., & Moreno, Y. (2009). Effects of mobility in a population of prisoner's dilemma players. *Physical Review E*, *79*(6), 067101.

Mengel, F. (2012). Learning across games. *Games and Economic Behavior*, *74*(2), 601–619.

Mengel, F. (2018). Risk and temptation: A meta-study on prisoner's dilemma games. *The Economic Journal*, *128*(616), 3182–3209.

Mengel, F., Orlandi, L., & Weidenholzer, S. (2022). Match length realization and cooperation in indefinitely repeated games. *Journal of Economic Theory*, *200*, 105416.

Michod, R. E. (2000). *Darwinian dynamics: evolutionary transitions in fitness and individuality*. Princeton University Press.

Moisan, F., ten Brincke, R., Murphy, R. O., & Gonzalez, C. (2018). Not all prisoner's dilemma games are equal: Incentives, social preferences, and cooperation.. *Decision*, *5*(4), 306.

Mookherjee, D., & Sopher, B. (1997). Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, *19*(1), 97–132.

Moran, P. A. P. (1962). *The statistical process of evolutionary theory*. Oxford: Clarendon Press.

Mosteller, F. (1956). Stochastic learning models. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 151–167. University of California Press Berkeley, CA, USA.

Murnighan, J. K., & Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games: A test of several models. *Journal of conflict resolution*, *27*(2), 279–300.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Myerson, R. B. (1997). *Game theory: Analysis of conflict*. Harvard university press.

Nakamaru, M., & Iwasa, Y. (2005). The evolution of altruism by costly punishment in lattice-structured populations: Score-dependent viability versus score-dependent fertility. *Evolutionary ecology research*, *7*(6), 853–870.

Nakamaru, M., Matsuda, H., & Iwasa, Y. (1997). The evolution of cooperation in a lattice-structured population. *Journal of theoretical Biology*, *184*(1), 65–81.

Nakamaru, M., Nogami, H., & Iwasa, Y. (1998). Score-dependent fertility model for the evolution of cooperation in a lattice. *Journal of Theoretical Biology*, *194*(1), 101–124.

Narendra, K. S., & Thathachar, M. A. (1974). Learning automata: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-4*(4), 323–334.

Narendra, K. S., & Thathachar, M. A. (2012). *Learning Automata: An Introduction*. Dover Publications.

Newman, M. E., & Watts, D. J. (1999). Scaling and percolation in the small-world network model. *Physical review E*, *60*(6), 7332.

Newton, J. (2018). Evolutionary game theory: A renaissance. *Games*, *9*(2), 31.

Nguyen, T. T., Nguyen, N. D., & Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, *50*(9), 3826–3839.

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11.

Noordman, C., & Vreeswijk, G. A. (2019). Evolving novelty strategies for the iterated prisoner's dilemma in deceptive tournaments. *Theoretical Computer Science*, *785*, 1–16.

Norman, M. F. (1968). Some convergence theorems for stochastic learning models with distance diminishing operators. *Journal of Mathematical Psychology*, *5*(1), 61–101.

Norman, M. F. (1972). *Markov processes and learning models*, Vol. 84. Academic Press New York.

Normann, H.-T., & Wallace, B. (2012). The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *International Journal of Game Theory*, *41*(3), 707–718.

Nowak, M. A. (2006a). *Evolutionary dynamics: Exploring the equations of life*. Harvard University Press.

Nowak, M. A. (2006b). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*(6398), 826–829.

Nowak, M. A., & Roch, S. (2007). Upstream reciprocity and the evolution of gratitude. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 605–610.

Nowak, M. A., & Sarah, C. (2013). *Evolution, Games, and God*. Harvard University Press.

Nowak, M. A., Sasaki, A., Taylor, C., & Fudenberg, D. (2004). Emergence of cooperation and evolutionary stability in finite populations. *Nature*, *428*(6983), 646–650.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298.

Nowak, M. A., Tarnita, C. E., & Antal, T. (2010). Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 19–30.

Ohtsuki, H. (2010). Evolutionary games in wright's island model: kin selection meets evolutionary game theory. *Evolution: International Journal of Organic Evolution*, *64*(12), 3344–3353.

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*(7092), 502–505.

Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness? Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology*, *231*(1), 107–120.

Ohtsuki, H., & Nowak, M. A. (2006a). Evolutionary games on cycles. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1598), 2249–2256.

Ohtsuki, H., & Nowak, M. A. (2006b). The replicator equation on graphs. *Journal of Theoretical Biology*, *243*(1), 86–97.

Ohtsuki, H., & Nowak, M. A. (2007). Direct reciprocity on graphs. *Journal of Theoretical Biology*, *247*(3), 462–470.

Ozaita, J., Baronchelli, A., & Sánchez, A. (2020). Ethnic markers and the emergence of group-specific norms. *Scientific Reports*, *10*(1), 1–13.

Pacheco, J. M., Traulsen, A., & Nowak, M. A. (2006). Active linking in evolutionary games. *Journal of Theoretical Biology*, *243*(3), 437–443.

Paiva, A., Santos, F., & Santos, F. (2018). Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Palomino, F., & Vega-Redondo, F. (1999). Convergence of aspirations and (partial) cooperation in the prisoner's dilemma. *International Journal of Game Theory*, *28*(4), 465–488.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502.

Peck, J. R., & Feldman, M. W. (1986). The evolution of helping behavior in large, randomly mixed populations. *The American Naturalist*, *127*(2), 209–221.

Perc, M., Gómez-Gardenes, J., Szolnoki, A., Floría, L. M., & Moreno, Y. (2013). Evolutionary dynamics of group interactions on structured populations: A review. *Journal of the Royal Society Interface*, *10*(80), 20120997.

Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., & Szolnoki, A. (2017). Statistical physics of human cooperation. *Physics Reports*, *687*, 1–51.

Perc, M., Ozer, M., & Hojnik, J. (2019). Social and juristic challenges of artificial intelligence. *Palgrave Communications*, *5*(1).

Perc, M., & Szolnoki, A. (2008). Social diversity and promotion of cooperation in the spatial prisoner's dilemma game. *Physical Review E*, *77*(1), 011904.

Perc, M., & Szolnoki, A. (2010). Coevolutionary games: A mini review. *BioSystems*, *99*(2), 109–125.

Peysakhovich, A., & Lerer, A. (2017). Prosocial learning agents solve generalized stag hunts better than selfish ones.. arXiv preprint.

Peysakhovich, A., & Lerer, A. (2018a). Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceeding of AAMAS 2018*, pp. 2043–2044.

Peysakhovich, A., & Lerer, A. (2018b). Towards ai that can solve social dilemmas. In *2018 AAAI Spring Symposium Series*.

Phelps, S. (2013). Emergence of social networks via direct and indirect reciprocity. *Autonomous Agents and Multi-Agent Systems*, *27*(3), 355–374.

Phelps, S., & Wooldridge, M. (2013). Game theory and evolution. *IEEE Intelligent Systems*, *28*(04), 76–81.

Pincus, J., & Bixenstine, V. E. (1977). Cooperation in the decomposed prisoner's dilemma game: A question of revealing or concealing information. *Journal of Conflict Resolution*, *21*(3), 519–530.

Pincus, J., & Bixenstine, V. E. (1979). Cognitive factors and cooperation in the prisoner's dilemma game. *The Psychological Record*, *29*, 463–471.

Pinheiro, F. L., Santos, F. C., & Pacheco, J. M. (2016). Linking individual and collective behavior in adaptive social networks. *Physical Review Letters,*, *116*(12), 128702.

Poppe, M. (1980). *Social comparison in two-person experimental games*. Ph.D. thesis, Tilburg University.

Postman, L. (1947). The history and present status of the law of effect. *Psychological Bulletin*, *44*(6), 489–563.

Rainey, P. B., & Rainey, K. (2003). Evolution of cooperation and conflict in experimental bacterial populations. *Nature*, *425*(6953), 72–74.

Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, *108*(48), 19193–19198.

Rand, D. G., Nowak, M. A., Fowler, J. H., & Christakis, N. A. (2014). Static network structure can stabilize human cooperation. *Proceedings of the National Academy of Sciences*, *111*(48), 17093–17098.

Ranjbar-Sahraei, B., Bou Ammar, H., Bloembergen, D., Tuyls, K., & Weiss, G. (2014). Evolution of cooperation in arbitrary complex networks. In *Proceedings of the 2014 international conference on Autonomous Agents and Multi-Agent Systems*, pp. 677–684.

Rankin, D. J., & Taborsky, M. (2009). Assortment and the evolution of generalized reciprocity. *Evolution: International Journal of Organic Evolution*, *63*(7), 1913–1922.

Rapoport, A., & Mowshowitz, A. (1966). Experimental studies of stochastic models for the prisoner's dilemma. *Behavioral Science*, *11*(6), 444–458.

Rapoport, A. (1967). A note on the 'index of cooperation' for prisoner's dilemma. *Journal of Conflict Resolution*, *11*(1), 100–103.

Rapoport, A., & Chammah, A. M. (1966). The game of chicken. *American Behavioral Scientist*, *10*(3), 10–28.

Rapoport, A., Chammah, A. M., & Orwant, C. J. (1965). *Prisoner's dilemma: A study in conflict and cooperation*, Vol. 165. University of Michigan Press.

Raub, W. (1988). Problematic social situations and the "large-number dilemma" a game-theoretical analysis. *Journal of Mathematical Sociology*, *13*(4), 311–357.

Realpe-Gómez, J., Andrighetto, G., Nardin, L. G., & Montoya, J. A. (2018a). Balancing selfishness and norm conformity can explain human behavior in large-scale prisoner's dilemma games and can poise human groups near criticality. *Physical Review E*, *97*(4), 042321.

Realpe-Gómez, J., Vilone, D., Andrighetto, G., Nardin, L. G., & Montoya, J. A. (2018b). Learning dynamics and norm psychology supports human cooperation in a large-scale prisoner's dilemma on networks. *Games*, *9*(4), 90.

Rezaei, G., & Kirley, M. (2012). Dynamic social networks facilitate cooperation in the n-player prisoner's dilemma. *Physica A: Statistical Mechanics and its Applications*, *391*(23), 6199–6211.

Rezek, I., Leslie, D. S., Reece, S., Roberts, S. J., Rogers, A., Dash, R. K., & Jennings, N. R. (2008). On similarities between inference in game theory and machine learning. *Journal of Artificial Intelligence Research*, *33*, 259–283.

Richter, H. (2019). Properties of network structures, structure coefficients, and benefit-to-cost ratios. *BioSystems*, *180*, 88–100.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, *54*(2), 296–301.

Roca, C. P., & Helbing, D. (2011). Emergence of social cohesion in a model society of greedy, mobile individuals. *Proceedings of the National Academy of Sciences*, *108*(28), 11370–11374.

Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2020). A structural solution to sequential moral dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1152–1160.

Rogers, A. R. (1990). Group selection by selective emigration: The effects of migration and kin structure. *The American Naturalist*, *135*(3), 398–413.

Rogers, A., Dash, R. K., Ramchurn, S. D., Vytelingum, P., & Jennings, N. R. (2007). Coordinating team players within a noisy iterated prisoner's dilemma tournament. *Theoretical Computer Science*, *377*(1-3), 243–259.

Rong, Z., Wu, Z.-X., Li, X., Holme, P., & Chen, G. (2019). Heterogeneous cooperative leadership structure emerging from random regular graphs. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *29*(10), 103103.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*(1), 164–212.

Roth, A. E., & Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, *17*(2), 189–198.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory*, *39*(1), 83–96.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*, Vol. 37. Citeseer.

Rustichini, A. (1999). Optimal properties of stimulus—response learning models. *Games and Economic Behavior*, *29*(1-2), 244–273.

Sabater-Grande, G., & Georgantzis, N. (2002). Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior & Organization*, *48*(1), 37–50.

Salazar, N., Rodriguez-Aguilar, J. A., Arcos, J. L., Peleteiro, A., & Burguillo-Rial, J. C. (2011). Emerging cooperation on complex networks. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 669–676.

Salmon, T. C. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, *69*(6), 1597–1628.

Samuelson, L. (1997). *Evolutionary games and equilibrium selection*, Vol. 1. MIT press.

Sánchez, A. (2018). Physics of human cooperation: experimental evidence and theoretical models. *Journal of Statistical Mechanics: Theory and Experiment*, *2018*(2), 024001.

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, *37*(1-2), 147–166.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics.* MIT press.

Santos, F., Pacheco, J., & Santos, F. (2018). Social norms of cooperation with costly reputation building. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Santos, F. P., Mascarenhas, S., Santos, F. C., Correia, F., Gomes, S., & Paiva, A. (2020). Picky losers and carefree winners prevail in collective risk dilemmas with partner selection. *Autonomous Agents and Multi-Agent Systems*, *34*(2), 1–29.

Santos, F. P., Mascarenhas, S. F., Santos, F. C., Correia, F., Gomes, S., & Paiva, A. (2019). Outcome-based partner selection in collective risk dilemmas.. In *AAMAS*, pp. 1556–1564.

Santos, F. P., Santos, F. C., Pacheco, J. M., & Levin, S. A. (2021). Social network interventions to prevent reciprocity-driven polarization. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1643–1645.

Santos, F. C., & Pacheco, J. M. (2005). Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, *95*(9), 098104.

Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006a). Cooperation prevails when individuals adjust their social ties. *PLoS Computational Biology*, *2*(10), e140.

Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2006b). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, *103*(9), 3490–3494.

Santos, F. C., Rodrigues, J., & Pacheco, J. M. (2006c). Graph topology plays a determinant role in the evolution of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), 51–55.

Santos, F. C., Rodrigues, J. F., & Pacheco, J. M. (2005). Epidemic spreading and cooperation dynamics on homogeneous small-world networks. *Physical Review E*, *72*(5), 056128.

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual review of psychology*, *70*(1), 719–745.

Schoenherr, J. R., & Thomson, R. (2020). Beyond the prisoner's dilemma: The social dilemmas of cybersecurity. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–7. IEEE.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, *14*(4), 433–440.

Sela, A., & Herreiner, D. (1999). Fictitious play in coordination games. *International Journal of Game Theory*, *28*(2), 189–197.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach. *Journal of Economic Behavior & Organization*, *7*(1), 47–70.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science advances*, *7*(21), eabd1705.

Shakarian, P., Roos, P., & Johnson, A. (2012). A review of evolutionary graph theory with applications to game theory. *Biosystems*, *107*(2), 66–80.

Shamma, J. S., & Arslan, G. (2005). Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria. *IEEE Transactions on Automatic Control*, *50*(3), 312–327.

Shapley, L. (1964). Some topics in two-person games. *Annals of Mathematics Studies*, *5*, 1–28.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095–1100.

Sherstyuk, K., Tarui, N., & Saijo, T. (2013). Payment schemes in infinite-horizon experimental games. *Experimental Economics*, *16*(1), 125–153.

Shoham, Y., Powers, R., & Grenager, T. (2003). Multi-agent reinforcement learning: A critical survey. Tech. rep., Technical report, Stanford University.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question?. *Artificial intelligence*, *171*(7), 365–377.

Sigmund, K. (2010). The calculus of selfishness. In *The Calculus of Selfishness*. Princeton University Press.

Sigmund, K., Hauert, C., & Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences*, *98*(19), 10757–10762.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1983). *Reason in Human Affairs Stanford University Press*. Stanford.

Skinner, B. (1938). *The behavior or organisms*. New York: Appleton-Century.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. OUP Oxford.

Skyrms, B., & Pemantle, R. (2009). A dynamic model of social network formation. In *Adaptive networks*, pp. 231–251. Springer.

Smale, S. (1980). The prisoner's dilemma and dynamical systems associated to non-cooperative games. *Econometrica: Journal of the Econometric Society*, *48*(7), 1617–1634.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Smith, M. J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18.

Stanley, E., Ashlock, D., & Smucker, M. D. (1995). Iterated prisoner's dilemma with choice and refusal of partners: Evolutionary results. In *European Conference on Artificial Life*, pp. 490–502. Springer.

Stanley, E. A., Ashlock, D., Tesfatsion, L., et al. (1994). Iterated prisoner's dilemma with choice and refusal of partners. In *Santa Fe Institute Studies in the Sciences of Complexity-Proceedings*, Vol. 17, pp. 131–175. Addison-Wesley Publishing Co.

Steele, M. W., & Tedeschi, J. T. (1967). Matrix indices and strategy choices in mixed-motive games. *Journal of Conflict Resolution*, *11*(2), 198–205.

Steinfatt, T. M. (1973). The prisoner's dilemma and a creative alternative game: The effects of communications under conditions of real reward. *Simulation & Games*, *4*(4), 389–409.

Stephens, C. (1996). Modelling reciprocal altruism. *The British Journal for the Philosophy of Science*, *47*(4), 533–551.

Stimpson, J. L., & Goodrich, M. A. (2003). Learning to cooperate in a social dilemma: A satisficing approach to bargaining. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 728–735.

Su, Q., Allen, B., & Plotkin, J. B. (2022). Evolution of cooperation with asymmetric social interactions. *Proceedings of the National Academy of Sciences*, *119*(1).

Sun, J., Fan, R., Luo, M., Zhang, Y., & Dong, L. (2018). The evolution of cooperation in spatial prisoner's dilemma game with dynamic relationship-based preferential learning. *Physica A: Statistical Mechanics and Its Applications*, *512*, 598–611.

Suppes, P., & Atkinson, R. C. (1960). *Markov learning models for multiperson interactions.* Stanford University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Szabó, G., & Fath, G. (2007). Evolutionary games on graphs. *Physics reports*, *446*(4-6), 97–216.

Szabó, G., & Hauert, C. (2002). Phase transitions and volunteering in spatial public goods games. *Physical Review Letters,*, *89*(11), 118101.

Szolnoki, A., & Perc, M. (2021). The self-organizing impact of averaged payoffs on the evolution of cooperation. *New Journal of Physics*, *23*(6), 063068.

Szolnoki, A., & Szabó, G. (2007). Cooperation enhanced by inhomogeneous activity of teaching for evolutionary prisoner's dilemma games. *EPL (Europhysics Letters)*, *77*(3), 30004.

Szolnoki, A., Xie, N.-G., Wang, C., & Perc, M. (2011). Imitating emotions instead of strategies in spatial games elevates social welfare. *EPL (Europhysics Letters)*, *96*(3), 38002.

Szolnoki, A., Xie, N.-G., Ye, Y., & Perc, M. (2013). Evolution of emotions on networks leads to the evolution of cooperation in social dilemmas. *Physical Review E*, *87*(4), 042805.

Tanimoto, J. (2009). A simple scaling of the effectiveness of supporting mutual cooperation in donor-recipient games by various reciprocity mechanisms. *BioSystems*, *96*(1), 29–34.

Tanimoto, J. (2015). *Fundamentals of evolutionary game theory and its applications*. Springer.

Tanimoto, J. (2017). Coevolution of discrete, mixed, and continuous strategy systems boosts in the spatial prisoner's dilemma and chicken games. *Applied Mathematics and Computation*, *304*, 20–27.

Tanimoto, J., & Sagara, H. (2007). Relationship between dilemma occurrence and the existence of a weakly dominant strategy in a two-player symmetric game. *BioSystems*, *90*(1), 105–114.

Tarnita, C. E., Antal, T., Ohtsuki, H., & Nowak, M. A. (2009). Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences*, *106*(21), 8601–8604.

Taylor, C., Fudenberg, D., Sasaki, A., & Nowak, M. A. (2004). Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology*, *66*(6), 1621–1644.

Taylor, C., & Nowak, M. A. (2006). Evolutionary game dynamics with non-uniform interaction rates. *Theoretical population biology*, *69*(3), 243–252.

Taylor, P. D., & Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, *40*(1-2), 145–156.

Taylor, P. D., & Wilson, D. S. (1988). A mathematical model for altruism in haystacks. *Evolution*, *42*(1), 193–196.

Tembine, H., Tempone, R., & Vilanova, P. (2012). Mean-field learning: A survey.. arXiv:1210.4657.

Thathachar, M. A., & Sastry, P. S. (2003). *Networks of learning automata: Techniques for online stochastic optimization*. Springer.

Thibaut, J., & Kelly, H. (1959). *The Social Psychology of Groups (13)*. New York, NY.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals.. *The Psychological Review: Monograph Supplements*, *2*(4), 1–109.

Thorndike, E. L. (1931). *Human learning*. New York: Appleton-Century-Crofts,.

Thorndike, E. L. (2013). *The elements of psychology*. Routledge.

Traulsen, A., & Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, *103*(29), 10952–10955.

Traulsen, A., Pacheco, J. M., & Nowak, M. A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, *246*(3), 522–529.

Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2010). Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences*, *107*(7), 2962–2966.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Tuyls, K., Heytens, D., Nowe, A., & Manderick, B. (2003). Extended replicator dynamics as a key to reinforcement learning in multi-agent systems. In *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 14*, pp. 421–431. Springer.

Tuyls, K., Hoen, P. J., & Vanschoenwinkel, B. (2006). An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, *12*(1), 115–153.

Tuyls, K., & Parsons, S. (2007). What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence*, *171*(7), 406–416.

Tuyls, K., & Stone, P. (2017). Multiagent learning paradigms. In *Multi-Agent Systems and Agreement Technologies*, pp. 3–21. Springer.

Tuyls, K., Verbeeck, K., & Lenaerts, T. (2003). A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 693–700.

Tuyls, K., & Weiss, G. (2012). Multiagent learning: Basics, challenges, and prospects. *AI Magazine*, *33*(3), 41–41.

Vainstein, M. H., Silva, A. T., & Arenzon, J. J. (2007). Does mobility decrease cooperation?. *Journal of Theoretical Biology*, *244*(4), 722–728.

Van Baalen, M., & Rand, D. A. (1998). The unit of selection in viscous populations and the evolution of altruism. *Journal of Theoretical Biology*, *193*(4), 631–648.

Van Doorn, G. S., & Taborsky, M. (2011). The evolution of generalized reciprocity on social interaction networks. *Evolution: International Journal of Organic Evolution*, *66*(3), 651–664.

Van Kleef, G. A., De Dreu, C. K., & Manstead, A. S. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. In *Advances in experimental social psychology*, Vol. 42, pp. 45–96. Elsevier.

Van Lange, P. A., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125–141.

Vassiliades, V., Cleanthous, A., & Christodoulou, C. (2011). Multiagent reinforcement learning: spiking and nonspiking agents in the iterated prisoner's dilemma. *IEEE Transactions on Neural Networks*, *22*(4), 639–653.

Vazifedan, A., & Izadi, M. (2023). A dynamic graph model of strategy learning for predicting human behavior in repeated games. *The BE Journal of Theoretical Economics*, *23*(1), 371–403.

Vincent, T. L., & Brown, J. S. (2005). *Evolutionary game theory, natural selection, and Darwinian dynamics.* Cambridge University Press.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior, 2nd rev.* Princeton University Press.

Vrancx, P., Tuyls, K., & Westra, R. (2008). Switching dynamics of multi-agent learning. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 307–313.

Walliser, B. (1998). A spectrum of equilibration processes in game theory. *Journal of Evolutionary Economics*, *8*(1), 67–87.

Wang, S.-Y., Liu, Y.-P., Zhang, F., & Wang, R.-W. (2021). Super-rational aspiration induced strategy updating promotes cooperation in the asymmetric prisoner's dilemma game. *Applied Mathematics and Computation*, *403*, 126180.

Wang, Z., Kokubo, S., Jusup, M., & Tanimoto, J. (2015). Universal scaling for the dilemma strength in evolutionary games. *Physics of Life Reviews*, *14*, 1–30.

Wang, Z., Kokubo, S., Tanimoto, J., Fukuda, E., & Shigaki, K. (2013). Insight into the so-called spatial reciprocity. *Physical Review E*, *88*(4), 042145.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279–292.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Weibull, J. W. (1997). *Evolutionary game theory.* MIT press.

Wilcox, N. T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, *74*(5), 1271–1292.

Willer, R., Kuwabara, K., & Macy, M. W. (2009). The false enforcement of unpopular norms. *American Journal of Sociology*, *115*(2), 451–490.

Willi, T., Letcher, A. H., Treutlein, J., & Foerster, J. (2022). Cola: consistent learning with opponent-learning awareness. In *International Conference on Machine Learning*, pp. 23804–23831. PMLR.

Williams, G. C., & Williams, D. C. (1957). Natural selection of individually harmful social adaptations among sibs with special reference to social insects. *Evolution*, *11*(1), 32–39.

Wilson, E. O., & Hölldobler, B. (2005). Eusociality: origin and consequences. *Proceedings of the National Academy of Sciences*, *102*(38), 13367–13371.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, *56*(645), 330–338.

Wright, S. (1929). The evolution of dominance. *The American Naturalist*, *63*(689), 556–561.

Wright, S. (1945). Tempo and mode in evolution: A critical review. *Ecology*, *26*(4), 415–419.

Wyer, R. S. (1969). Prediction of behavior in two-person games.. *Journal of Personality and Social Psychology*, *13*(3), 222.

Xu, J., Garcia, J., & Handfield, T. (2019). Cooperation with bottom-up reputation dynamics. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 269–276.

Yang, H.-X., Wu, Z.-X., & Wang, B.-H. (2010). Role of aspiration-induced migration in cooperation. *Physical Review E*, *81*(6), 065101.

You, T., Shi, L., Wang, X., Mengibaev, M., Zhang, Y., & Zhang, P. (2021). The effects of aspiration under multiple strategy updating rules on cooperation in prisoner's dilemma game. *Applied Mathematics and Computation*, *394*, 125770.

Young, H. P. (1993). The evolution of conventions. *Econometrica: Journal of the Econometric Society*, *61*(1), 57–84.

Yu, C., Zhang, M., Ren, F., & Tan, G. (2015). Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(12), 3083–3096.

Zeng, W., Li, M., & Feng, N. (2017). The effects of heterogeneous interaction and risk attitude adaptation on the evolution of cooperation. *Journal of Evolutionary Economics*, *27*(3), 435–459.

Zhang, C., Liu, S., Wang, Z., Weissing, F. J., & Zhang, J. (2022). The "self-bad, partner-worse" strategy inhibits cooperation in networked populations. *Information Sciences*, *585*, 58–69.

Zhang, L., Huang, C., Li, H., & Dai, Q. (2019). Aspiration-dependent strategy persistence promotes cooperation in spatial prisoner's dilemma game. *EPL (Europhysics Letters)*, *126*(1), 18001.

Zhang, L., Li, H., Dai, Q., & Yang, J. (2022). Adaptive persistence based on environment comparison enhances cooperation in evolutionary games. *Applied Mathematics and Computation*, *421*, 126912.

Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences*, *109*(5), 1419–1424.

Zimmermann, M. G., Eguiluz, V. M., & Miguel, M. S. (2001). Cooperation, adaptation and the emergence of leadership. In *Economics with heterogeneous interacting agents*, pp. 73–86. Springer.