University of Nottingham
UK | CHINA | MALAYSIA

How does experience influence game reasoning? A survey for repeated prisoner's dilemma

Name: Tong Fang

Department: School of Economics

Supervisor: Dr. Alex Possajennikov

Dr. Fabio Tufano

Date: August 2022

Acknowledgement

I am deeply grateful for the support and encouragement I have received in the completion of my dissertation. I would like to express my gratitude to my supervisors, Dr. Alex Possajennikov and Dr. Fabio Tufano, for their guidance and suggestions in the writing of this dissertation, especially their reassuring and encouraging words when I struggled. And to my family, this is not possible without their unconditional support, love, and patience throughout my study and life.

Abstract

This dissertation reviews past literature on belief updating/belief-based learning and sophistication (reasoning), links the two models with recent experimental findings in repeated prisoner's dilemma, supergame strategies, attention and incentives, and tries to answer how these two models interact with each other with experience. These two models have good performance in explaining part of the repeated games separately but fail to a smooth and united explanation. By re-examining the literature from the perspective of the information source to generate different types of belief, elements of two models can be reconciled into one. Recent new findings in attention and incentives also contribute to the evolution of sophistication, which was not captured by the sophistication models. This helps to examine sophistication models in a dynamic way and makes both models comparable from the dynamic point of view.

Keywords: Prisoner's dilemma, Repeated games, Belief, Sophistication, Supergame strategy, Incentive

Contents

1 Introduction

Strategic interactions between agents play a foundational role in many economic models. Games, like prisoner's dilemma, coordination game, ultimatum game, etc., have been applied to a number of sub-fields in economics. Prisoner's dilemma is the oldest game discussed in economics. It has been widely studied due to its unique structure on the tension between individual interest and collective benefit. Being a non-cooperative and non-zero-sum game, it contains richer situations in strategic interaction that can be studied and worth studying. Recent empirical evidence and meta-studies on prisoner's dilemma provide many new findings on supergame strategies and payoff structure, which can be linked with much literature on sophistication reasoning and cognitive effort.

Traditional game theory is built upon rational agents with unlimited computational power. However, empirical evidence on these games does not support its prediction. Recent studies on these games are streamed into two channels. One is adaptive learning, which allows agents to make reactions based on past experience, and gradually converge to equilibrium. The belief formation/updating model is one class of this family and the most widely used in economics, where agents learn their opponents' move, form beliefs about their move from experience/learning and respond to that belief. The other channel is behavioural strategic reasoning, an extension of the traditional game theory. This kind of model (best summarized in Crawford et al. (2013)) can make predictions of non-equilibrium plays. The most powerful and popular model is sophistication reasoning including level-k thinking and cognitive hierarchy. Sophistication reasoning roots in the idea that "the natural way of looking at game situations ... is not based on circular concepts, but rather on a step-by-step reasoning procedure" (Selten 1998). The models provide a well-specified set of belief and behavioural rules at each level of sophistication for players with limited rationality and computational power.

A longstanding problem is whether and how sufficient exercise/training can improve people's decision-making. Empirical studies on various experimental settings including payoff structure, matching protocols, number of repetitions, game complexity, rewarding theme[1], etc, give a positive answer to the first question. As for the second question, the mechanism is still not perfectly clear. Sufficient training is defined differently in different models. In the belief formation models Brown (1951), agents learn in a setting, where the same stage game is repeated again and again with a partner/opponent, and the reward is calculated by a (discounted) sum of payoff in each stage, while in sophistication reasoning models, agents are trained in a setting where the same game is replayed with a different opponent, and the reward is calculated by a randomly picked round.

Different model settings generate different information for agents to make predictions and reasoning mechanisms. It is widely acknowledged that sophistication reasoning shows a good performance for the initial play of each repeated game (Crawford et al. 2013, Camerer et al. 2004*b*), while belief updating tracks the belief change stage by stage accurately (Crawford & Broseta 1998, Nyarko & Schotter 2002, Aoyagi et al. 2021). Since many findings, either in theories or experiments, are based on different settings, a systematic review is lacking on how different elements, such as beliefs, information, and incentive, influence people's decision. In this dissertation, I review the literature on belief-based models and sophistication reasoning in order to provide a comprehensive view of how people make strategic decisions. In this dissertation, I

[1]Two typical themes are that subjects are rewarded by accumulated payoff or the payoff of a randomly picked stage

am going to show when beliefs (over stage actions, supergame strategies or sophistication levels) and the information used to form beliefs are classified based on either individual or group level, many different results under different settings can be reconciled.

It is important to review models and to go deep into elements leading to equilibrium play. Empirical evidence suggests that both belief updating and sophistication reasoning works in repeatedly playing a game, but the specific mechanism is vague. The main question is under what conditions people use one or the other, and change from one to the other. Is there any possibility that both are part of a unified reasoning mechanism that people learn to reason? Moreover, by digging into the elements behind each model, can they provide a smoother explanation? Both models use beliefs, but they are generated over different sets (stage actions, supergame strategies, level of sophistication), do they have anything in common? Traditional definitions of belief ignore the information source used to form beliefs, a specific opponent or the opponent population. For example, the stage action belief assumes agents have no history/information about the opponent at the initial stage, but is there truly no information for the agents to make predictions? Can the agent use the population information as a substitute? Besides, do recent new findings in attention and cognitive effort provide new insights into the problem? By making these questions clear, we can have a better understanding of how these scattered studies link to each other and the real gap we are facing with.

The structure of the remaining dissertation is arranged in the following way: In section 2, I will introduce the setup of the prisoner's dilemma and people's practical behaviour in different experimental settings. In section 3 and 4, the two main models, belief updating and sophistication reasoning are introduced successively. I am going to show how the model proprieties capture some specific behaviour in playing prisoner's dilemma in different settings. Finally, I will reexamine beliefs based on their information source and discuss the possibility of people learning to reason by combining the elements of both models together.

2 Behaviour pattern in prisoner's dilemma

2.1 Prisoner's dilemma

The repeated prisoner's dilemma game starts with a **stage game** $G(I, A, \Pi)$, which is a standard two-person $i \in I = \{1, 2\}$ prisoner's dilemma game (PD game) with two actions, cooperation (C) and defection (D). Let $A_i = \{C, D\}$ be the stage action set for both players ($a_i \in A_i$ refers to player i's an action), and let $A = A_1 \times A_2$ be the stage action space for the stage. The stage payoff $\pi_i(a_1, a_2) \in \Pi$ to each player i is jointly determined by both players' simultaneous actions (or stage action profile (a_1, a_2)), which can be shown in the form of a matrix (table 1). For example, when player 1 (2) choose C (D), the payoff to player 1 (2) is $\pi_1(C, D) = b$ ($\pi_2(C, D) = c$), which is shown in table 1 in bold.

A repeated PD game (often called a **supergame** $G(I, A, \Pi, H, T, \delta)$) is to play the stage game repeatedly with a horizon T (the number of repetitions). If horizon $T = 1$, then it is a standard PD game, often called a **one-shot PD game**. If horizon $T \geq 2$, it can be a **finite supergame** ($T < \infty$), where the terminal stage is well and commonly identified before a supergame, or an **infinite supergame** ($T = \infty$), where there is a high probability (**continuation probability**) to extend an extra stage. At any stage $t = 1, 2, ...$, players make the decision $a_i^t(h^{t-1})$ based on the history of past $t-1$ periods, h^{t-1}. For $t = 1$, there is no history $h^0 = \emptyset$. Correspondingly, we have a history space at the stage t denoting the history actions in the past $t-1$

Table 1: Payoff matrix of PD games

	C	D
C	a, a	**b, c**
D	c, b	d, d

Note: Parameters follow $c > a > d > b$ (Mengel 2018), the first (second) letter in each cell refers to row (column) player's payoff.

stages, $H^{t-1} = \cup h^{t-1} = A^{t-1}$. The **(continuation) payoffs** in supergames are calculated as the discounted (arithmetical) sum of payoffs in each stage game in the infinite (finite) case[2].

$$u_i(G) = \begin{cases} \sum_{t=1}^{\infty} \delta^{t-1} \pi_i^t(a_1^t(h^{t-1}), a_2^t(h^{t-1})) & \text{Infinite case} \\ \sum_{t=1}^{T} \pi_i^t(a_1^t(h^{t-1}), a_2^t(h^{t-1})) & \text{Finite case} \end{cases} \tag{1}$$

where $\delta \in [0, 1]$ is the discount factor[3]. For the scope of this dissertation, I only consider the full information case.

Players are also assumed to use some patterns/paradigms of thinking/reasoning tactics, called **(supergame) strategies** $\sigma_i : (a_t|h^{t-1}) \to [0, 1]$, mapping player i's action set A_i to a probability space $[0, 1]$ based on history[4]. Normally, by playing a supergame with a horizon of T periods, there are 2^T pure strategies (Aoyagi et al. 2021). However, people tend to play a typical subset of strategies in practice, such as always defection (AD), always cooperation (AC), Tit-For-Tat (TFT), Grim, and threshold-x strategies (T-x) (See Table 2).

Table 2: Description of supergame strategies

Strategy	Description
Always Defect (AD)	Always play D.
Always Cooperate (AC)	Always play C.
Tit-For-Tat (TFT)	Play C unless partner played D last stage.
Grim	Play C until either player plays D, then play D forever.
Threshold-x (T-**x**)	Play Grim until stage **x** then switch to AD.

Note:
1. Source: Aoyagi et al. (2021). Only the most widely used strategies across literatures are listed here.
2. By the definition of threshold strategies, AD can also be interpreted as a threshold-0 strategy.

2.2 Cooperation in supergames

The Folk theorem provides a baseline prediction of people's behaviour that choosing the actions (sequence) that lead to a continuation payoff above the minmax one is optimal. In the infinite supergame, it is to figure out a discount factor $\delta \in [0, 1]$ sustaining a higher payoff (see equation (1)). In repeated PD games, we can reach an AD strategy by applying backward induction from the final stage. This action sequence

[2] In many literatures, a normalized factor $1 - \delta$ is multiplied to the infinite continuation payoffs.
[3] Sometime, this can also be interpreted as a continuation probability in infinite supergame (Fudenberg & Tirole 1991)
[4] In some literature, strategies function is interpreted as a mapping from history to action.

leads to a minmax payoff, which is the lower bound of all possible available payoffs. Any action sequence leading to a payoff lower than minmax should not be chosen. This opens the possibility of choosing any action sequence and mutual cooperation at the early stage of a repeated PD game becomes reasonable. It is often related to the reputation effect. It is also the case in infinitely repeated PD games. Aoyagi et al. (2021) reported high average cooperation (80%) along the first several eight stages with a high continuation probability ($\delta = 7/8$) and a payoff structure helpful to cooperation. Dal Bó & Fréchette (2011) also reported relatively high average cooperation (60-80%) with a continuation probability of $3/4$ and around 35% with a continuation probability of $1/2$, when C-C is supported by both the equilibrium play and risk dominance.

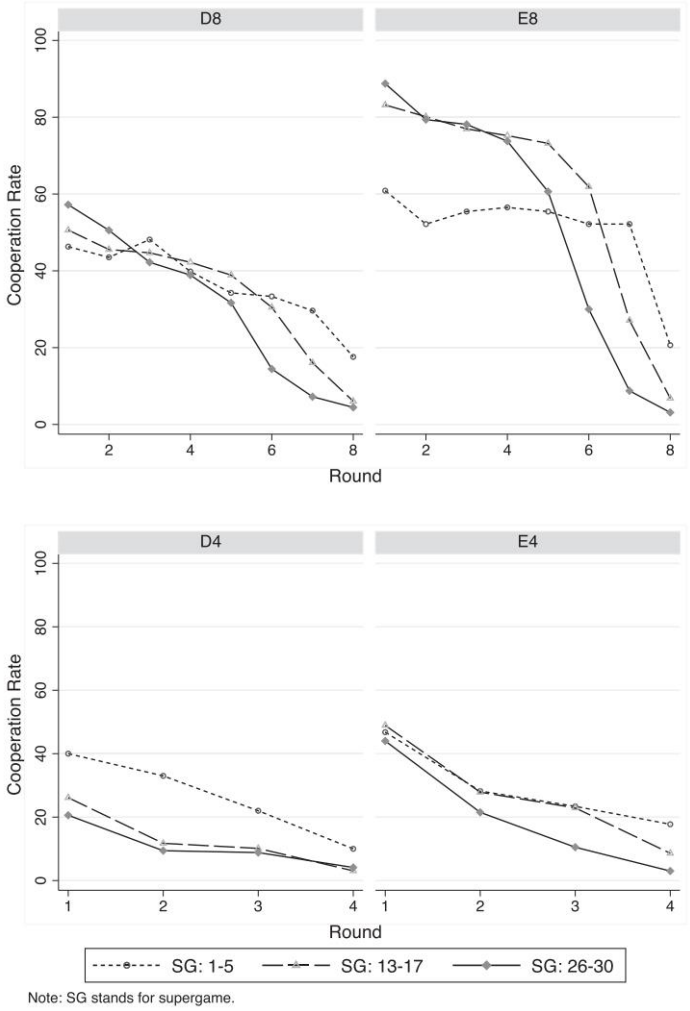
Sensitive to game parameters Dal Bó & Fréchette (2011) first claimed that cooperation might not increase with experience, even though C-C is supported as the equilibrium and risk-dominant play. In their case, the cooperation rate decreases with experience (from 30% to 5%), when C-C is not supported by the equilibrium and sustains at a low level (around 20-30%) with experience when C-C is supported by the equilibrium play but not the risk-dominance. It seems to be contradictory whether experience enhances cooperation. Dal Bó & Fréchette (2018) summarized determinants of cooperation in infinite supergames. They found that the more likely one is going to interact with his/her opponents, the more cooperative he/she is, and this cooperativeness increases with experience. This implies the assumption for the holding of reputation effect (or the shadow of future) that the continuation probability needs to be high (close to an infinite case rather than a finite case), The second important finding is that the cooperation rate is high when the parameters form a game that can hedge the strategic uncertainty from opponents' decisions.

Mengel (2018) provided a more detailed look into this issue. Three key factors are found crucial in (finitely repeated) PD games: temptation (percentage gain when unilaterally defecting against a cooperator), risk (percentage loss when unilaterally cooperating against a defector) and efficiency (the number of gains by mutual cooperation as opposed to mutual defection) (Mengel 2018). She argued that the play of subjects is highly sensitive to the payoff parameters. They carried out experiments, covering a wide range of payoff parameter values, and found that risk (temptation) significantly influences average cooperation rates in one-shot (finitely repeated) PD games, but the influence magnitude of temptation in the repeated setting is smaller than that of risk in a one-shot setting. There is more cooperation in repeated settings if and only if the risk is high and temptation is low. Also, in finitely repeated games, the risk is more important in the early stages of a game and when players are inexperienced, while in later stages of the game and when players are more experienced temptation becomes more important. Besides, Embrey et al. (2018) ran an experiment to compare the cooperation rate between settings with higher risk (risk $>$ temptation) and higher temptation (risk $<$ temptation) and found that games with higher risk generates more cooperation. Also, initial cooperation increases with experience in the high-risk setting but shows no difference with experience in the high-temptation setting.

Endgame effect People prefer to play defection at stages close to the terminal stage in the finite case (Aoyagi et al. 2021, Embrey et al. 2018). Selten & Stoecker (1986) named such a phenomenon **endgame effect**. A Probit regression analysis also shows this negative association between the terminal stage and the cooperation rate (Aoyagi et al. 2021). Besides, the stage to first defection changes with experience. Selten & Stoecker (1986) explored cooperation behaviour in 25 supergames, with 10 stages each and reported a

gradual decrease in the stage to the first defection. Embrey et al. (2018) replicated the result but added a restriction when the horizon of the supergame is large enough (e.g. eight stages). When the horizon is small (e.g. four stages), the pattern is different (see Figure 1). This is consistent with the Folk theorem that when cooperation cannot bring a high continuation payoff, defection is the optimal choice. Nonetheless, there is also evidence against it, which suggests an opposite change. Andreoni & Miller (1993) compared cooperation between 200 one-shot PD games with random matching and 20 finite ten-stage supergame with human subjects. The mean round to the first defection increases with subjects gaining experience, starting below two in the first supergame and ending above 5 in the last.

Figure 1: Cooperation rate by rounds



Source: Embrey et al. (2018). The top two panels (D8 & E8) are treatments with eight stages per supergame, while the bottom two (D4 & E4) are treatments with four stages. With more supergame played (from SG 1-5 to SG 13-17 to SG 26-30), the cooperation rates at the early stages go up under the long horizon (the top two panels), but this is not the case in the short horizon (the bottom two panels). There is no single stage where the cooperation rate increases with experience.

Aoyagi et al. (2021) compared initial cooperation rates[5] in both finite and infinite cases in their experiment and found that the initial rates in both cases are similar. Though mutual cooperation can be a long-standing stable equilibrium in the infinite situation, it dropped dramatically when it is close to the terminal stage in the finite case. In Andreoni & Miller (1993)'s experiment, they also found in the repeated setting,

[5]They used a payoff structure leading to high initial cooperation

the cooperation rate begins at a high level (above 60%), lasts at 50% for 6 stages and dropped dramatically (10%)in the last stage, while less than 30% of subjects in the one-shot setting choose to cooperate in the last ten rounds (Equivalently, the last one supergame in the repeated setting). The comparison implies that extra information allows subjects to make more inferences about their opponents, and to be more responsive to the belief. Besides, past experience has a systematic impact on decisions in the subsequent stages. Cooper et al. (1996) compared cooperation between 20 one-shot PD games and two finite supergame with ten stages, and found that cooperation rates start above 50% in the finitely repeated game and end below, but are always lower for the one-shot game. However, the data in this study is not sufficient to attribute this to any learning or evolution. Moreover, some literatures (Huck & Weizsäcker 2002, Aoyagi et al. 2021, Embrey et al. 2018) also claimed that people's behaviour is forward-looking even though a move cannot be supported by past action frequency.

2.3 Summary

Empirical evidence supports the Folk theorem that mutual cooperation is practically possible or rational when the payoff of mutual cooperation is higher than that of mutual defection. Mutual cooperation as an SPE is a necessary condition for a high cooperation rate. Moreover, indicators, risk and temptation, explain the reason why SPE is a necessary condition in a behavioural aspect. When the temptation is high, extra information from experience is risky to be used. Besides, the Folk theorem also supports the endgame effect if agents examine the continuation payoff in every stage, then they should defect only at the final stage in the finite supergame. However, it fails to capture the gradual decrease with experience in the stage to the first defection.

3 Belief updating

The Folk theorem provides theoretical support to cooperation as an equilibrium play in the repeated PD games, but it does not tell which equilibrium play would be sustained along the horizon. A typical way is to assume people follow a belief and respond to that belief. The **belief** is a probability distribution over opponents' choice. It can be an **belief over actions** if subjects treat each PD game as one stage of a supergame and predict opponents' next stage actions, or **belief over supergame strategies** if subjects treat the supergame as a whole and predict opponents' supergame strategies. This section is going to introduce models and experiments on the belief over stage actions and supergame strategies.

3.1 Belief over stage actions

Early studies focus on belief learning, which is unobservable and is recovered from action data. To the best of my knowledge, Nyarko & Schotter (2002) are the first to study stated belief. Though the accuracy of eliciting belief in experiments and its effect on subjects making better decisions are still disputable[6], it

[6]Gächter & Renner (2010) discussed the behaviour of incentivised stated belief in the repeated public good games and found that incentives to belief would additionally raise the proportion of contribution and the belief. However, in most belief elicitation experiments in repeated PD games, there is no report on the effect of elicitation, mostly because they avoided belief elicitation throughout the whole experiment(Aoyagi et al. 2021, Gill & Rosokha 2020). A comprehensive review of belief elicitation can

comes to the fact that action data are best response to subjects' stated beliefs, no matter whether it is a finite or infinite supergame (Aoyagi et al. 2021).

Fictitious play was first introduced by Brown (1951) as a cognitive algorithm to compute the equilibrium by simulation and was later interpreted as a model learning from actual behaviour history. This model, in most studies, was applied to two-player simultaneous move games, so there is no compatible problem with applying the model to our case (Dhami 2016). Agents are assumed to use opponents' historical action frequencies to construct the distribution and make the best response conditional on that distribution. The most critical assumption in this model is the stationary assumption of the distribution from which opponents draw their actions. Then agents are assumed to maximise their current-period payoff.

In our case $G(I, A, \Pi, H)$, at any stage t, agents own a weight $w_i^t(C_j, D_j)$ (i refers to agent self, and j refers to the opponent) and applied the weight to form belief over actions (x_i^t) at stage t (Equation 2).

$$x_i^t(a_j) = \frac{w_i^t(a_j)}{w_i^t(C_j) + w_i^t(D_j)} \geq 0 \tag{2}$$

where $a_j \in A_j = \{C, D\}$. Then, agents make the best response a_i^* conditional on this belief (Equation 3).

$$a_i^* \in BR(x_i^t) = \{a_i \in A_i : \pi_i(a_i, x_i^t) \geq \pi_i(\hat{a}_i, x_i^t), \forall \hat{a}_i \in A_i\} \tag{3}$$

Belief updating follows the frequency rule. If an action is played at stage t, then its weight for next stage is augmented by 1 unit and the unplayed one remains the same (Equation 4). The initial belief is also formed by the exogenous initial weight.

$$w_i^{t+1} = \begin{cases} w_i^t(a_j) + 1 & \text{if } a_j \text{ is played at stage } t \\ w_i^t(a_j) & \text{otherwise} \end{cases} \tag{4}$$

One problem of this original model is the non-differentiable best response function (Dhami 2016). Hence, a smooth version of fictitious play was introduced by adding an additive shock to the payoff (Equation 5).

$$\begin{cases} \pi_i^{Smooth} = \pi_i(a_i, a_j) + \eta_i \\ lnPr(\eta_i < y) = -e^{-\lambda_i y} \\ x_i^t(a_i|a_j) = \frac{e^{-\lambda_i \pi_i(a_i, a_j)}}{e^{-\lambda_i[\pi_i(C_i, a_j) + \pi_i(D_i, a_j)]}} \end{cases} \tag{5}$$

where η is the random shock and λ is an individual-specific parameter. In this smooth fictitious play, action belief is not explicitly disclosed. Instead, it is displayed as the probability of a player i's action given his/her opponents' plays a_j (or sometimes mixed strategies). Another more general variant is weighted fictitious play, introduced by Cheung & Friedman (1997), which takes into account the distance of historical events to the current period. Earlier experience is heavily discounted by the rate ϕ. This somehow models the recency effect or forgetting effect of the faraway past, solving the unpractical problem of unlimited memory in the original model.

be checked on Charness et al. (2021).

$$x_i^t(a_j) = \frac{1(a_j, a_j^t) + \sum_{k=1}^{t-1} \phi_i^k 1(a_j, a_j^{t-k})}{1 + \sum_{k=1}^{t-1} \phi_i^k} \tag{6}$$

where $1(\cdot, \cdot)$ is the indicator function and equal to one if the two input is the same and zero otherwise. When ϕ comes to 1, it becomes the original model, while when ϕ goes to 0, it becomes Cournot play.

Though fictitious play is a behaviour learning model, it is still possible to reach the same result (SPE) as the traditional game theory suggests. This means the perfect rational result can be nested in the fictitious play model. Fudenberg & Levine (1995) showed that if the following two conditions are met both predictions are consistent: 1) The initial belief over stage actions are given by the Dirichlet distribution. 2) Every player believes that opponents' play sequence is i.i.d. multinational random at each stage. This covers the link between practical behaviour and extreme behaviour (full rationality).

Fictitious play is a popular model of belief-based learning models. Some other models include general belief learning (Crawford & Broseta 1998), and Bayesian learning (Jordan 1991) . There are fewer studies on these models theoretically and empirically. The general belief learning model was proposed by Crawford & Broseta (1998) to explain the convergence of scattered initial beliefs and adaptive dynamics in order-statistic coordination games (Van Huyck et al. 1991). In an order-statistic coordination game, subjects guess a number $(s_i(t))$ between 1 to 7 and are incentivised positively by their guess and negatively by the difference $(y(t))$ between their guess and the order statistic[7]. In their model, the initial choice is given by a time-specific group element (α_0) and an individual-time-varying element (ξ_{i0}), and the subsequent choice is updated by a weighted sum of the guess $(s_i(t))$ and the difference $(y(t))$, apart from the two variables above. (See equation 7)

$$\begin{cases} s_i(0) = \alpha_0 + \xi_{i0} \\ s_i(t) = \alpha_t + \xi_{it} + (1 - \phi_t) y(t-1) + \phi_t s_i(t-1) \end{cases} \tag{7}$$

where the weight (ϕ_t) is time varied, representing an agent's degree of improvement from the previous order statistic. Crawford & Broseta (1998) pointed out that though it is sensitive to the selection of initial parameters, the model predicts that people will eventually reach the equilibrium by keeping modifying their choice from feedback and by shrinking the group belief and individual shocks.

The Bayesian learning (Jordan 1991) models behaviours in an environment where agents are not sure about the exact payoff matrix of their opponents, but they have common knowledge about the potential alternatives of opponents' payoff matrix. So, agents can learn which payoff matrix is applied to their opponent after continuous playing. However, this model cares more about which payoff matrix is used rather than what action their opponents will do. Agents try to use action beliefs to learn opponents' payoffs. When the potential payoff matrix becomes known to every subject, the model prediction that subjects should immediately reach the equilibrium was not observed so often in Cox et al. (2001)'s experiment. Camerer et al. (2004*a*) criticized the fragility of the model that its correct prediction stands on the right plays of the first several rounds. However, some theorists love this model because it captures the unrealistic perfect anticipation of others' moves with belief updating, which is not a typical feature in belief-based learning models.

Fictitious play (also other belief-based learning models) only allows agents to passively adapt to his/her

[7]In Van Huyck et al. (1991), the order statistic is the minimum

opponent by simulating the data from the previous stages, but cannot model active strategic moves. However, empirical evidence does not fully support this. Nyarko & Schotter (2002) designed an experiment to play a finitely repeated 2×2 game with a unique mixed Nash equilibrium in fixed and random matching with/without belief (over actions) elicitation. They compared the incentivized stated belief over actions with the belief estimated by the weighted fictitious play model using past opponents' actions. The weight ϕ is fitted close to 1. They found three interesting results: 1) subjects' behaviour best responds to their stated belief; 2) stated belief varies from stage to stage and does not converge; 3) estimated belief (from weighted fictitious play) better predicts opponents' decisions. So, the belief subjects held is not an accurate one or a pure summary belief that can lead them to a higher payoff. This shows that there must have more elements in forming people's beliefs.

A recent study comprehensively compares people's cooperation in finite and infinite PD games. Aoyagi et al. (2021) asked subjects to play the same stage PD game repeatedly in finite (eight stages) and infinite (essential eight stages with a continuation probability of 7/8) cases. The payoff parameters are carefully designed to have high cooperation rates in the initial response in both cases (of their interest). Subjects were asked to elicit their action belief stage by stage only after four supergames in order to avoid the incentivised belief effect. They confirmed and extended Nyarko & Schotter (2002)'s result that people best respond to their belief in both cases. Furthermore, by having a detailed look into stated action belief, they found two patterns of systematic deviation, an early pessimism in the infinite supergames and a late optimism in the finite case. This provides a potential explanation for the gap between stated action belief and estimated belief in Nyarko & Schotter (2002). Besides, they also found the same action frequencies do not reach the same belief in the finite and infinite cases and vice versa. This is not covered in the fictitious play. They also noticed that stated action beliefs predict the change of their opponents' behaviour (forward-looking) within a game, which is not a feature of belief-based models as well. The main difference in stated action belief in both settings is the drop of cooperation in the terminal stage and people seem to anticipate such breakdown. All these evidence shows that the way people play repeated PD games is similar at the beginning in both cases with beliefs responding to their action beliefs. Nonetheless, the difference in belief and play in both settings is led by forward-looking the potential breakdown of cooperation in the final stage of the finite supergame. Since there is no terminal stage in the infinite supergame, there is no need to forward look anything, with no dramatic drop in the cooperation rate in the last one or two stages.

3.2 Belief over supergame strategies

Some literatures then put their emphasis to the belief over supergame strategies due to the failure of action belief to capture the endgame effect or forward-looking and the fact that elicited action belief is not purely the summary of past action frequency. It becomes a complementary method to model people's behaviour in addition to simple belief-based learning. Dal Bó & Fréchette (2011) tried to propose one to explain the change in strategy.

Dal Bó & Fréchette (2011) report that AD, Grim, and TFT (slightly) account for the strategies played in the infinite supergames[8]. By focusing on AD and Grim, they used a belief-based model (for stage action

[8]AD and Grim account for 88% of the data. The proportion to each specific strategy is summed up more than one because Dal Bó & Fréchette (2011) only ensure the behaviour is consistent with the behavioural rule of a specific strategy. (It is hard to

belief) to model the evolution of supergame belief. Given subjects' weights on AD β_{it}^{AD} and on Grim β_{it}^{G}, they are updated in a way of discounted fictitious play.

$$\beta_{i(t+1)}^{k} = \theta_i \beta_{it}^{k} + 1(a_j^k) \tag{8}$$

where θ_i is an individual-specific discount rate and k is specified to a strategy (AD or G). Their beliefs are calculated in the following way.

$$p_{it}^{a} = \frac{e^{\frac{1}{\lambda_i t} U_{it}^{a}}}{e^{\frac{1}{\lambda_i t} U_{it}^{AD}} + e^{\frac{1}{\lambda_i t} U_{it}^{G}}} \tag{9}$$

where a refers to a specific action (C or D) and U_{it}^{a} is a random utility function.

$$U_{it}^{a} = \frac{\beta_{it}^{AD}}{\beta_{it}^{AD} + \beta_{it}^{G}} u^{a}(a_j^{AD}) + \frac{\beta_{it}^{G}}{\beta_{it}^{AD} + \beta_{it}^{G}} u^{a}(a_j^{G}) + \lambda_{it} \varepsilon_{it}^{a} \tag{10}$$

where $u^{a}(a_j^{k})$ is the average utility function when agent i plays action a with agent j playing action a_j. λ_{it} is a parameter representing how an agent best responds to his/her belief. ε_{it} is an idiosyncratic error term. Dal Bó & Fréchette (2011) reported a good model performance with the empirical data. This model follows the manner of fictitious play, but it is only limited to two strategies and is burdensome to extend it to more strategies. Hence, most studies now are focused on experiments.

Gill & Rosokha (2020) asked participants to play 25 rounds of infinite supergame with a continuation probability of $3/4$, by choosing within ten supergame strategies[9] at the beginning of each supergame. Participants were incentivized to elicit their beliefs on supergame strategies at the first and final supergame. They manipulated the value of mutual cooperation payoff as the way in Dal Bó & Fréchette (2011). Basically, supergame belief is generally consistent with the chosen strategies at the aggregate level in the first supergame even though there is no history of this experiment and its accuracy increases in the payoff value of mutual cooperation. However, when goes to the individual level, only 25% of subjects perfectly best respond to their supergame belief. Besides, they replicated that the cooperation rate increases in the payoff value of mutual cooperation (Dal Bó & Fréchette 2011), and found that there is a positive relationship between optimism and cooperation. Optimism is defined as the expected cooperation rate and also increases in the payoff value of mutual cooperation. The accuracy of supergame belief is decreased by optimism when the payoff of mutual cooperation is low but is slightly higher due to optimism when the payoff is high. In addition, they found that optimism obstacles people from best responding to their beliefs when the payoff is low under an approximation definition of best response[10]

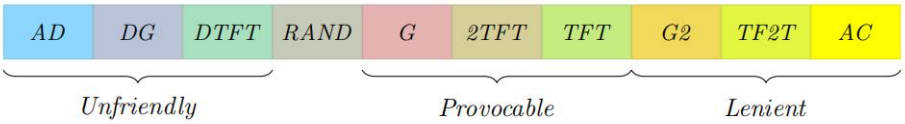
Gill & Rosokha (2020) also used the regression model to study the transition between strategies types (see Figure2). They classified the ten strategies into three categories based on the time of the first defection: unfriendly (defection in the first round), provocable (defection as long as the opponents defect), and (lenient) delayed defection after being unravelling. One of their regression results shows that subjects using provocable strategies and experiencing long horizons in previous strategies tend to keep applying provoca-

tell the difference (between AC, TFT and G) if a player always plays D). Dal Bó & Fréchette (2018) replicated the result and constrained that punishment in TFT and G is credible to be useful only if cooperation is supported by the equilibrium play.

[9]Participants familiarized the strategy set by preplay it with themselves.

[10]A strategy is the best response if its payoff is in the vicinity of 3.15% of the best response

Figure 2: Strategy category



Source: Gill & Rosokha (2020)

ble strategies and lower the probability of using unfriendly strategies. These findings show both beliefs on supergame strategies and their transitions are only affected by opponents' actions. If opponents prefer cooperation more, subjects are less likely to change beliefs greatly in one or two supergames. This is also verified in their regression analysis. When others' strategy is cooperative in the previous supergame, subjects tend to remain at the strategy used in the previous round, but the magnitude of the probability of remaining in the same strategy category is a quarter lower when subjects played unfriendly strategies compared with the other two categories.

Embrey et al. (2018) recruited participants to play finite supergames of 20 or 30 rounds, providing ample experience. Controlling the payoff and horizon of the supergame, they found that the fraction of subjects who learn to make threshold strategies implied by the backward induction increases with the number of supergame. Even though various strategies are carried out at the beginning, it converges into threshold strategies when a great amount of experience is gained. This is also supported by the experiment result that the skewness of the probability distribution of breakdown of cooperation (over stages in a supergame) shifts to the left when more rounds of supergames are played. Types' shifts from normal supergame strategies (AD, TFT, Grim, etc.) to threshold strategies were also found in Aoyagi et al. (2021). This shows the possibility of the existence of strategic reasoning over supergame strategies because the best response to threshold strategy m is threshold strategy $m-1$. By anticipating opponents' supergame strategy distribution, one can take out his/her best response, similar to the logic of level-k thinking. Players who anticipate their opponents to play a threshold-x strategy tend to carry out a threshold-(x-1) strategy. In this manner, applying supergame strategies becomes a number guessing game, choosing the stage of the first defection.

Aoyagi et al. (2021) used elicited action belief $\mu_i^t(a_j|h^{t-1})$ in the last one to three supergames to recover subjects' supergame belief $\tilde{p}_i^t(\sigma_j|h^{t-1})$ via a Bayesian rule and assume agents use supergame belief at the first stage $\tilde{p}_i^1(\sigma_j|h^{t-1})$ to compute one's own supergame strategies applied to that whole supergame.

$$\begin{cases} \mu_i^t(C|h^{t-1}) = \sum_{\sigma_j \in Z_j} \tilde{p}_i^t(\sigma_j|h^{t-1})\sigma_j(C|h^{t-1}) \\ \tilde{p}_i^t(\sigma_j|h^{t-1}) = \frac{\tilde{p}_i^{t-1}(\sigma_j|h^{t-1})\sigma_j^{t-1}(C|h^{t-1})}{\sum_{\sigma_j \in Z_j} \tilde{p}_i^{t-1}(\sigma_j|h^{t-1})\sigma_j^{t-1}(C|h^{t-1})} \\ \sigma_i \in \arg\max_{\tilde{\sigma}_i \in Z_i} \tilde{p}_i^1(\sigma_j) u_i(\tilde{\sigma}_i, \sigma_j) \end{cases} \tag{11}$$

where $Z_i \subseteq \Sigma_i$ is a finite subset of the full strategy set. Aoyagi et al. (2021) built up the subset with 16 strategies based on literature evidence. They used SFEM[11] (Dal Bó & Fréchette 2011) to estimate a group-level strategies distribution. The results confirmed findings in Embrey et al. (2018) that subjects like to use

[11] They used the 0.06 for implementation error $(1-\beta)$

threshold (T-x) strategies in the finitely repeated PD games (T7 for30%, T8 for 22%, AD for 12%, TFT for 9%, and T6 for 8%) and in Dal Bó & Fréchette (2018) that people prefer to apply conditional cooperative strategies, such as TFT (36%), Grim(18%), Grim2 (11%), AC (11%)[12]. Subjects were classified into different types based on the most frequent strategies used in the experiment, and Aoyagi et al. (2021) built the supergame belief by various types. They found subjects managed to predict popular strategies in each game setting and subjects using different supergame strategies held heterogeneous supergame beliefs. To be specific, subjects tend to overestimate others' use of the same strategies as themselves and to underestimate others' use of less cooperative strategies than themselves. Aoyagi et al. (2021) ranked strategies based on expected payoffs when there is no implementation error ($\beta \to 0$). The cooperativeness ranked from least to most as follows: AD, STFT, T6, T7, T8, Grim, TFT, Grim2, TF2T, and AC.

It is innovative to include belief over supergame strategies into belief over actions. However, there are still some questions left unanswered. The first one is how people narrow down the large supergame strategies to a feasible subset of strategies. As the potential pure strategies space in a supergame is 2^T (T is the horizon of a supergame), the criteria that agents apply to choose the subset of supergame strategies lack study. Second, the time people begin to consider supergame strategies is unknown. As mentioned above, the consideration of choice (stage actions or supergame strategies) to supergame strategies explicitly shows ways agents treat the supergame, either stage by stage or as a whole. It is unclear what the condition that triggers the transfer is. Lastly, though I classify belief over supergame strategies as one of belief formation and updating, it contains the property of sophistication reasoning (forward-looking), which will be discussed in the next section[13].

3.3 Summary

Theoretical action belief is basically a summary of past opponents' actions, usually in a manner of the weighted average of frequency. It fails to predict anything unrealized in history. However, elicited action belief shows a different pattern and the difference between finite and infinite supergame cannot be captured by belief-based learning. Nonetheless, elicited beliefs are not only accurate to predict opponents' moves but also are best responded to by the players themselves. The empirical evidence ensures the validity of the idea of using beliefs to predict players' actions and the assumption that people are subjective rational (Boudon 1989), which defines people's behaviour to be rational as long as it is reasonable, and is not necessary to carry out decisions resulting in the objective maximal payoff. The difference between theoretical action beliefs and the stated one is studied from the view of supergames strategies. Supergame strategies play a role as a paradigm guiding agents to make choices. This explains why some actions with tiny probability based on history are played. Several studies found that even though players apply various strategies at the beginning, their behaviour steadily converges to some typical strategies with experience, Grim and TFT in the infinite supergame and threshold-x strategies in the finite case.

[12] AD is only for 9% in the infinite case.

[13] A better interpretation of belief over supergame strategies is a way of strategic reasoning in the form of belief updating.

4 Sophistication reasoning

The other mainstream of considering people's reasoning is the Sophistication model. It is a class of strategic reasoning in a more behavioural view. It is good at predicting players' initial play in repeated games, but cannot track the subsequent plays in the same repeated game. Interestingly, when agents have no experience in playing the game and knowledge of their opponents, the models can capture players' behaviour. The word 'sophisticated' is usually used in describing and comparing the depth/number of iterations of reasoning. The term 'sophistication' refers to one's cognitive ability to carry out reasoning based on game structure/payoff without learning (about the opponent) in most cases. The two most popular models are level-k thinking and cognitive hierarchy. It provides good predictions for the initial play in contrast to belief updating, which often assumes an exogenous initial weight/belief, and in some cases, the prediction may be sensitive to initial parameters.

4.1 Models and empirical evidences

Level-k thinking Level-k thinking, proposed by Stahl & Wilson (1995) and Nagel (1995), is a model predicts non-equilibrium strategic reasoning model. The model assumes the population of players are heterogeneous in sophistication (a cognitive resource used to make reasoning). There is no restriction or assumption on the component of the population. The model fixes the behaviour of each level of sophistication thinking to be the same. Level-0 ($L\emptyset$) is assumed to follow a naive and non-strategic behaviour with a uniform distribution over all feasible actions. $L1$ players suppose all his/her opponents to be $L\emptyset$ players and make the best response to the $L\emptyset$ pattern. $L2$ players make the best responses to the $L1$ pattern. In summary, Lk players only realize all players are one level lower than themselves and make the best responses to the $L(k-1)$ pattern. A large number of studies show that there are rare L$\emptyset$ players (Crawford & Broseta 1998, Dhami 2016, Camerer et al. 2004*a*). Most players are $L1$ type, a minority is $L2$ and few are $L3$ or above. Hence, the $L\emptyset$ behaviour type is more likely a stereotype model/ anchor type in $L1$ reasoning. This model is a deterministic one, avoiding the effect of the noisiness of others' responses.

Cognitive hierarchy Cognitive hierarchy (CH), proposed by Camerer et al. (2004*b*), is a generalization of level-k thinking. The main motivation for this generalization is when k goes up, the proportion of higher level does not shrink, instead, it goes up as well because level-k thinking model assumes all players hold sophistication one level lower that the players, which is contradictory to empirical studies that higher $k(\geq 3)$ is rare. For example, the actual $L3$ players are rare in practice, but for a $L4$ player, all other players are $L3$ type. CH-k players realize the existence of players with sophistication from $CH-0$ to $CH-(K-1)$, including all levels lower than theirs. Due to cognitive limits, equal or higher level than k is unrecognized by a $CH-k$ player. The composition of players in a population follows a distribution, which is common knowledge to all players but is restricted to the sophistication level $CH-k$ one holds. The Poisson distribution is the most popular one.

$$f(k) = \frac{e^{-\tau}\tau^k}{k!}; k = 0, 1, 2, ... \tag{12}$$

where τ (the mean and variance of a Poisson distribution) can be interpreted as the expected level of sophistication of the population. $f(k)$ tells the proportion of $CH-k$ players with different level k in population. Its property that $\frac{f(k)}{f(k-1)}$ decreases at a rate proportional to k is consistent with empirical evidence. In each $CH-k$ player's point of view, the relative frequency $(g_k(h))$ of players of each levels lower than his/her $CH-h(h<k)$ is calculated as

$$g_k(h) = \frac{f(h)}{\sum_{l=0}^{k-1} f(l)} \tag{13}$$

Another big difference between level-k thinking and cognitive hierarchy is the error structure. In level-k thinking, players are allowed to have implementation errors to create fluctuations because the only random type/level is L$\emptyset$, which rarely exists in practice. However, in CH, the L$\emptyset$ disturbance though rarely happens yet was repeated again as a $CH-k$ player considers also levels of players below him/her. This adds more randomness than level-k thinking does. Hence, CH does not allow errors for agents. One drawback of this generalized model is the sensitivity to the population distribution. Different distribution assumptions affect the model predictions. Camerer et al. (2004*b*) report the τ on 24 p-beauty contests with various p, and found that the parameter τ ranges from 0.1 to 4.9, with a median close to the golden ratio (1.618)[14]. They also estimated the parameter in other games (with mixed strategy equilibrium) in the early rounds without feedback, and the value ranged from 0 to 15.9. One explanation they provided is the value of the payoff. They argued that a higher payoff pushes participants to make more steps of reasoning, which would be discussed later.

In dominance-solvable games (such as prisoner's dilemma), $CH-k$ agents always assume their opponents with a lower level of sophistication $CH-h$ $(k>h)$. The $CH-0$ is assume to have uniform randomization on their actions, so the action distribution $(CH-0)$ of is $p(a_j) = \frac{1}{\text{\# of actions in the action set}}$. In the prisoner's dilemma case, it would be $p(C_j) = p(D_j) = 1/2$. Then, we can iterate to get the action distribution for each level h below k By knowing the relative frequency of players with lower sophistication $g_k(h)$, we can have the opponents' action distribution by summing the weighted probability up.

$$p_k(a_j^h) = \sum_{h=0}^{k-1} g_k(h) p(a_j^h) \tag{14}$$

Dhami (2016) pointed out that when $\tau \to \infty$, agents' behaviour is close to the behaviour in Nash equilibrium and level-k thinking. Crawford et al. (2013) summarized the model performance in various games that cognitive hierarchy performs at least better than Nash equilibrium.

Alaoui & Penta (2016) and Alaoui & Penta (2022) proposed a model of strategic thinking of initial responses to extend the argument in Camerer et al. (2004*b*) that payoff stake might change people's depth of reasoning (expectation of sophistication)[15]. In their study, the choice is a result of depth of reasoning, determined by the value and the cost of reasoning. Each step of reasoning leads to a better understanding of their opponent in the game. The value of reasoning (benefit) increases in games' payoff, while the cost is negatively associated with one's exogenous reasoning ability. A higher level of cognitive reasoning ability

[14]The background of subjects are various from students to professionals.

[15]Players' behaviour is fixed for each level in level-k thinking model. Hence, agents are insensitive to incentives/payoffs. Another method considering the influence of incentives on deviation is quantal response equilibrium (QRE). However, it is not of my focus here because this model is highly sensitive to the noise distribution Haile et al. (2008).

(upper limit of one's sophistication) exerts less effort/cost to the reasoning for the same depth of reasoning. In a game, the cost of reasoning is fixed. Higher the payoff, the more steps of reasoning. The incentive for a further step of reasoning is the payoff difference between the one a player could get if he chose the best response (one more step) and the one he would receive given his current action. This model disentangles the actual sophistication in the reasoning process from intrinsic cognitive sophistication ability/state. This implies one's reasoning sophistication can be inconsistent with the actual sophistication reflected by his/her action. For example, a level-3 player only needs to play a level-2 strategy when he/she thinks his/her opponent is level-1. If the opponent is thought to be less sophisticated than the player, the best response of that player is not to play at his/her full strategic sophistication, but some level lower, which is the actual sophistication. Based on one's belief about their opponent's sophistication/reasoning cost, one's depth of reasoning changes.

Their findings are supported by an experiment. The task they used in the experiment is the 11-20 game. The game is played in the following way: Two players report a number between 11-20. They can get the number of tokens they report. Besides, if the number that one reports is just one below his/her opponent's, then one will get x tokens as a bonus. Also, if tied, then both subjects gain 10 tokens as a bonus. There are two kinds of treatments. One is to manipulate incentives/value of reasoning, by setting x to be either 20 or 80. The other is to manipulate subjects' beliefs about their opponents' reasoning sophistication/cost of reasoning. This piece of information is given to the subjects in the form of either their opponents' programme types ("science and maths" or "humanities") or test levels (high or low). Participants were asked to play 18 games in a row, covering all the combinations of treatments. Partners were randomly assigned for each game. The result shows that the depth of reasoning (weakly) decreases in the payoff. The steps of reasoning are also influenced by the sophistication belief of opponents and higher-order belief of the sophistication belief (i.e. a second-order belief is that one believes his/her opponent believes the sophistication one is)

However, Esteban-Casanelles & Gonçalves (2020) criticized Alaoui & Penta (2016)'s implication that only the difference between payoff (payoff distortion) matters and provided some empirical evidence for how full incentive scale up also changes people's behaviour and beliefs via choice implementation mistakes (for noisy best response) and belief formation (in sophistication reasoning). They used diagonal games (Gonçalves 2020), an experimental tool that can vary the steps of reasoning by keeping other things (e.g. payoff structure and the number of actions) unchanged, to run a $2 \times 2 \times 2$ experiment, with two levels of players' incentive, two levels of opponents' incentives and two levels of the number of iteration to the dominant solution. The level of payoff stake reaches a ratio of 40, with a large absolute value, which magnifies the incentive effect. Each subject played the one-shot game once with a completion time of 22 minutes. Choices and action beliefs about their opponents were elicited. They used a very special experiment setting, adding the position of an observer (Huck & Weizsäcker 2002). A player (A') is assumed to play with an opponent (B) who is also playing the same game with a player (A), who has the same incentive treatment as A'. The payoff of players A' is jointly determined by both A' and B's choice, but not vice versa. The authors focused on the behaviour and belief of player A's and asked player A's to elicit the expected belief of player B's action. Basically, they found high own incentives raise people's belief and action tendency toward non-dominated strategies (it is only the case of the change of opponents' incentives if subjects' own incentive is high), while complexity (the number of iterations to reach the dominance solution) decrease

subjects' belief and action tendency to dominant solutions. In terms of the implementation mistakes, they showed that when own incentive is high, people are more likely to best respond to one's belief with fewer implementation errors and there is a distributional shift down in losses. As for belief formation, they found the belief of subjects with high own incentives are less uniformly random and are more accurate (only when their opponents have high incentives). They also explained that subjects with high incentives indirectly influence the accuracy of beliefs by increasing response time (effort).

Dynamic level-k Both two sophistication models do not allow agents to improve their prediction with experience, which violates the reality. Agents in level-k thinking always prefer all other players to be one level below theirs, while agents in CH estimate a distribution of the population sophistication level and holds it along the supergame. Ho & Su (2013) proposed a dynamic level-k thinking model to explain violation in backward induction, the limited induction and repetition unravelling. Limited induction refers to the phenomenon that the number of stages deviated from backward induction increases with the horizon of a sequential game. Repetition unravelling is a violation that game behaviour converges to the prediction of backward induction when the game is repeated more enough. They constructed a rule hierarchy (a set of potential moves for the whole sequential game or centipede game in their case), with each corresponding to a specific level of sophistication. Unlike level-k thinking or CH, they allow participants to build the composition of the population by trials. Agents update their sophistication belief (the belief over opponents' sophistication level in the population) stage by stage by following the rule of fictitious play, given the initial belief. Hence, the subjects are assumed to be homogeneous (in contrast to level-k thinking and CH) but heterogeneous in their sophistication beliefs. The model explains the limited induction by the sophistication belief. A longer horizon allows the sophistication belief to be updated more times, leading to more deviation. Repetition unravelling is explained by a best response to sophistication belief and shifting to higher level of sophistication with few people choosing low-level rule. This model shows how people improve their decision through learning.

4.2 Attention

Attention is widely measured as an indicator of sophistication. Eye-tracking experiments have verified its validity as a measure of information collection and a reflection of the use of sophistication (Polonio et al. 2015). A recent study tracked the use of attention to show the evolution in sophistication with experience, which again proves that sophistication is not a fixed characteristic, but is something that can be learned. Marchiori et al. (2021) explored whether and how previous experience affects the level of sophistication reasoning. The sophistication was measured via visual fixation and lookup patterns on various payoff areas, which separates sophistication from belief updating.

The authors ran an experiment with two-person 3×3 normal-form dominance-solvable games, evaluating subjects' level of sophistication in three games requiring various levels of analytic effort: the one with dominant strategy only for players (DS), the one with dominant strategy only for the opponent(DO) and the one with no dominant strategy for both (ND). Players played with an algorithm (always played as the opponent) which always plays the equilibrium move[16]. They compared players' decisions and eye fixations

[16]Participants were informed of the rationality of this algorithm before the experiment

before and after the learning stage to see whether the learning experience makes a difference in subjects' choice, eye fixation and visual moving patterns. The learning stage contains 2×2 treatments with a binary treatment of with/without feedback and two levels of analytic effort required for the task.

One of their main results shows that experiencing cognitive-difficult games with feedback greatly increases subjects' frequencies of equilibrium choice and sophistication level. Subjects in the difficult with feedback treatment had a significant transition from self-payoff to opponent-payoff, while subjects in easy treatment did not. Participants increase their attention (fixation time) to other-sum and other-dominant payoff regions, decrease in own-sum fixation, but do not decrease in own-dominant. By contrast, participants in the easy treatment with feedback group did not learn to consider others' incentives.

The study showed the possibility of the learning effect in understanding the game structure. To be more specific, having experience requiring more analytic effort with feedback shifts subjects' attention assignment to others' dominant strategy. This study shows how sophistication evolves in non-zero-sum games. The main problem is here payoff scheme. Subjects were randomly rewarded by one trial, which means subjects had to take each round separately, eliminating the reputation effect. Each round is a one-shot game. This differs from the repeated games, as they only focus on learning the game rather than learning the opponent.

4.3 Summary

This section introduced the two main sophistication models, in which action beliefs and behavioural rules are pre-set and agents do not learn or improve their choice with experience. This corresponds to the opinion that sophistication is an intrinsic mental cognitive ability, which should not change in a short time. However, as technology allows sophistication to be measured in experiments, evidence supports that sophistication is also possible to evolve. This separates the concept of sophistication as a cognitive ability and reasoning level reflected in choice. Moreover, the incentive plays an important role in sophistication reasoning, but the mechanism between incentives and presented sophistication is not well set. Either full scale-up or partial distortion seems to influence people's reasoning.

5 Discussion

In this dissertation, three beliefs are mentioned, belief over stage action, belief over supergame strategies and belief over the level of sophistication (either in CH or in dynamic level-k thinking). Beliefs[17] are widely considered as a distribution over a set (stage actions, supergame strategies or sophistication level) to the opponent whom one is playing with, but it fails to identify the belief type by the information source, a group belief or an individual belief. A **group belief** is a belief using population information (estimated by the samples), while an **individual belief** refers to a belief using the information of a specific opponent.

When using supergame belief or sophistication belief, subjects are asked to make a decision without specific information about the randomly matched new opponent. In the supergame literature, the value or source of the initial supergame belief is not mentioned, either recovered from stated action belief or elicited, and the supergame strategy implemented cannot be changed in a supergame, while the sophistication belief

[17] In this section, beliefs refer to all three beliefs if not specified.

is decided by the expected sophistication level τ. It is a question of how people have these beliefs or expectations even if they do not have the information about their specific opponents. In other words, players should have consistent beliefs and actions when there is no individual-specific information. However, empirical evidence supports that people's initial performance improves (e.g. an increasing initial cooperation rate) even if their history in a specific supergame is still empty. Hence, there must be some information in this improvement and group information is the only option left. People tend to use population information as a reference to form their beliefs. This implies that a change in these two beliefs reflects subjects' knowledge and estimation of the group. Hence, both supergame belief and sophistication belief are group beliefs. The convergence of supergame strategies chosen at the beginning of a supergame in both finitely (threshold strategies) and infinitely (conditional cooperative) repeated games can be evidence that subjects gradually reach an agreement on the group action distribution, not on a specific individual (In the prisoner's dilemma case, it is the cooperation rate). This is also why the estimation of supergame belief can be transformed into sophistication belief (Gill & Rosokha 2020) because some supergame strategies are a mutual representation of some sophistication levels.

Such an inference can also be applied to stage action belief of the initial stage in a supergame because the initial belief is not formed by any specific information about their opponent. A player A should always form the same belief in a supergame no matter whether he/she plays with opponent B or C because he/she does not know which move the opponent will play and use the population belief as an approximation to the individual one. In the subsequent play, agents collect information specific to that opponent. They can make specific action sequences in the manner of fictitious play. However, many literatures on PD games showed that people's elicited belief, though best responded, is different from fictitious play. From the scope of belief type, the question is whether the elicited belief is an individual belief of stage actions which is proceeded by fictitious play. The endgame effect is a counter-example to this. The decrease in the round to the first defection is only possible when group information is used. Otherwise, a player should not know when his/her opponent defects.

Hence, I propose a potential mechanism for people to generate beliefs. The elicited belief $b^e_{it}(a_j)$ is a joint distribution of both summarized stage action beliefs $b^s_t(a_j)$ to a specific opponent by summarizing history frequency in a supergame and a group belief $b^g_t(a_j)$ over action sequence across supergames.

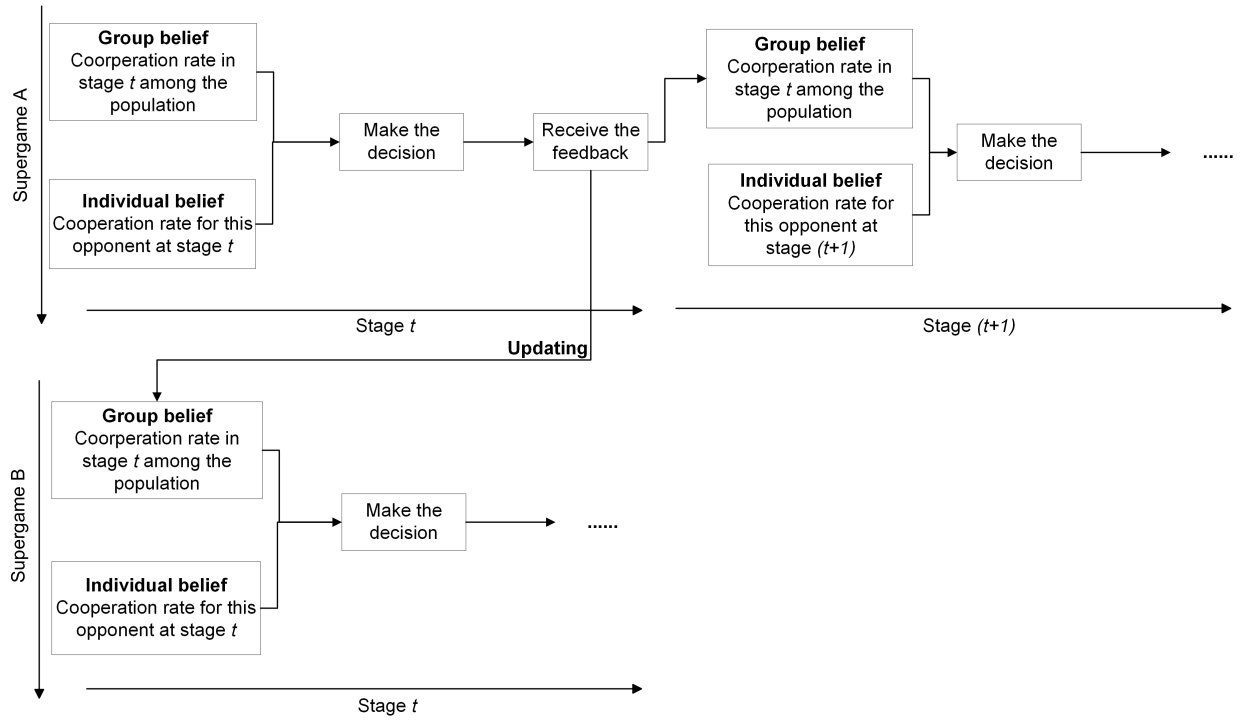
$$b^e_{it}(a_j) = \rho_{it} b^s_t(a_j) + (1 - \rho_{it}) b^g_t(a_j) \tag{15}$$

where ρ_{it} is a stage-varying preference for individual-specific or group belief, related to the learning paths/action sequence (e.g. initial cooperation/defection) and the game structure[18]. An interpretation of this parameter is the level of trust in one's opponent. If one trusts his/her opponent, one should weigh individual belief more than group belief because individual belief provides more information to the opponent whom one is playing with. After receiving the feedback, both group and individual beliefs are updated marginally and jointly. (see Figure 3)

While the marginal update is an adaptive behaviour to their previous marginal belief and the joint update is an adaption to the weight between individual belief and group belief. On one side, it is decided by the history of a specific opponent. If the cooperation between the two does not unravel before, the inertia keeps

[18]The influence of game structure will be discussed in next paragraph.

Figure 3: Marginal updating process



the player to infer the opponent to be trustful. On the other hand, the incentive is an important component of the trust parameter ρ_t. Trust should not be built purely on experience but also better be sustained by some practical assurance. Incentives, constructing the game's structure, provide such an assurance. If mutual cooperation is not an SPE or the temptation is sufficiently high, one can hardly believe his/her opponent will not betray even if it has not happened before.

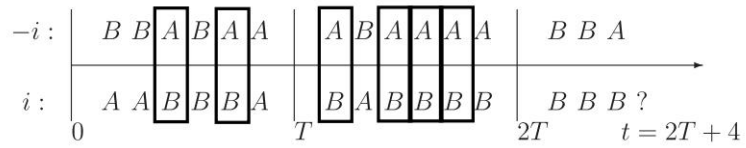
The incentive is not covered in the previous two classes of models. Either model fails to sustain its belief with some practical elements/facts. Given the same initial belief, incentives do not generate different beliefs if the history is the same, and the incentive does not change the behavioural rules and stage action beliefs in each level of sophistication. However, Dal Bó & Fréchette (2011), Dal Bó & Fréchette (2018) argued that mutual cooperation is possible when it is an SPE with a high continuation probability in the infinite case. Besides, (Mengel 2018) has shown that regardless of a repeated (finite (Embrey et al. 2018) or one-shot game, incentives are important to players in a prisoner's dilemma game. The temptation and risk critically influence people's cooperation choices. Besides, incentive structure influences people's cognitive effort on forming beliefs, shifting subjects' attention on some specific actions (Marchiori et al. 2021), changing depth of reasoning (Alaoui & Penta 2016, 2022), and requiring various response time (Esteban-Casanelles & Gonçalves 2020). The link between incentives and belief is worth a deep study. It definitely influences people's initial decision, which is repeated again and again in each stage of the game.

If I call the belief updating to a specific opponent learning, and to a group of opponent reasoning, then this mechanism can be called learning to reason. In simple words, an elicited stage action belief is not only conditional on the history of a supergame but also conditional on the history of that specific stage across supergames played. Hence, the endgame effect or forward-looking can be explained by beliefs across supergames at stage t. It can be a potential direction for a simulation and an experiment in the future. This mechanism might solve another question of how supergame strategies and behaviour related to a specific

sophistication level are consciously recognized by an agent. Theoretical models do not explain how these are formulated by people from inexperience to experience. By putting group beliefs as adaptive behaviour evolving during play, these pre-set behaviour and belief patterns can be a result of repeated playing.

Recently, there is one study close to this idea. Mengel (2014) proposed a model combining belief updating and forward-looking. Agents in this model are classified as a group role. For example, in our PD games case, one player is randomly chosen from group i, and the other player is also randomly picked from a group j. Hence, a group stage action belief is updated in each supergame. She assume agents have limited memory $M_i^t(H|m)$, which can recall opponents' action in the last m-instance starting from the h-stage history $H^t(h) = \{a^{t-h}, a^{t-h+1}, ..., a^{t-1}\}$. (see Figure 4 for example) After knowing the history, agents randomly

Figure 4: Example (Mengel 2014)



Assume that h = 1 and m = 5. The T means a supergame. At $t = 2T + 1$, the history is $H^t = \{(B, A)\}$, and the memory is $M_i^t((B, A)|5)$. I emphasize the history (B, A) on the timeline. The last five actions of opponent $-i$ are $\{A, B, A, A, A\}$. The last one in the second supergame does not count because there is no reaction for that history at that supergame.

sample $\rho < m$ from the memory and form opponents' stage action belief $\mu_i^t(a_{-i}|H)$. Besides, agents have an action plan by looking at k-stages forward. The stage $t + k$ will be a temporary terminal stage for the action plan and a belief of terminal stage $\mu_i^t((a_i^\tau, a_{-i}^\tau)_{t=\tau}^{t+k}|(a_i^\tau)_{\tau=t}^{t+k})$ is constructed based on their action plans. Action plans are made by best response to belief along these forward-looked k stages. Hence, the belief in Mengel (2014)'s model is a group belief specific to a history ($H_i^t(h)$). An agent summarizes history to estimate the stage action belief of the opponent population, and make future decision plans by constructing opponents' future action based on the group belief and best responding to each group action belief at each future stage. As an implication, it allows individuals to make an actual decision different to their plans. However, in this way, individual-specific information is wasted, and only group information is retained to estimate group belief.

$$max_{(a_i^\tau)_\tau = t,..,t+k} V(\mu_i^t(H), (a_i^\tau)) = \sum_{(a_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k}} \mu_i^t((a_i^\tau, a_{-i}^\tau)_{t=\tau}^{t+k}|(a_i^\tau)_{\tau=t}^{t+k}) \sum_{\tau=t}^{t+k} \pi_i(a_i^\tau, a_{-i}^\tau) \tag{16}$$

Other popular streams also explain people's cooperation in repeated PD games, including team reasoning (Sugden 2003, 2011, Gold et al. 2012), conditional cooperation (Gächter & Renner 2010, Neugebauer et al. 2009, Chaudhuri et al. 2017), social or other-regarding preference (Yamagishi et al. 2013, Pei et al. 2021), and inequality aversion (Backhaus & Breitmoser 2021). However, Dal Bó & Fréchette (2018) argued that there is no robust evidence on whether those behavioural traits systematically influence the play in the infinite supergames where C-C is sustained as an equilibrium play, and strategic motivation behind the cooperation in infinite supergame prevails in empirical evidence(Dal Bó & Fréchette 2018). Besides, it is still unknown whether these characteristics are the factors for cooperation or a result of cooperative interaction. This needs to be checked in future studies.

References

Alaoui, L. & Penta, A. (2016), 'Endogenous depth of reasoning', *The Review of Economic Studies* **83**(4), 1297–1333.

Alaoui, L. & Penta, A. (2022), 'Cost-benefit analysis in reasoning', *Journal of Political Economy* **130**(4), 881–925.

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

Aoyagi, M., Fréchette, G. R. & Yuksel, S. (2021), 'Beliefs in repeated games', *ISER DP* (1119R).

Backhaus, T. & Breitmoser, Y. (2021), Inequity aversion and limited foresight in the repeated prisoner's dilemma, Technical report, Center for Mathematical Economics Working Papers.

Boudon, R. (1989), 'Subjective rationality and the explanation of social behavior', *Rationality and society* **1**(2), 173–196.

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

Camerer, C. F., Ho, T.-H. & Chong, J. K. (2004*a*), Behavioural game theory: thinking, learning and teaching, *in* 'Advances in understanding strategic behaviour', Springer, pp. 120–180.

Camerer, C. F., Ho, T.-H. & Chong, J.-K. (2004*b*), 'A cognitive hierarchy model of games', *The Quarterly Journal of Economics* **119**(3), 861–898.

Camerer, C., Ho, T.-H. et al. (1997), Experience-weighted attraction learning in games: A unifying approach, Technical report.

Camerer, C. & Hua Ho, T. (1999), 'Experience-weighted attraction learning in normal form games', *Econometrica* **67**(4), 827–874.

Charness, G., Gneezy, U. & Rasocha, V. (2021), 'Experimental methods: Eliciting beliefs', *Journal of Economic Behavior & Organization* **189**, 234–256.

Chaudhuri, A., Paichayontvijit, T. & Smith, A. (2017), 'Belief heterogeneity and contributions decay among conditional cooperators in public goods games', *Journal of Economic Psychology* **58**, 15–30.

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

Cooper, R., DeJong, D. V., Forsythe, R. & Ross, T. W. (1996), 'Cooperation without reputation: Experimental evidence from prisoner's dilemma games', *Games and Economic Behavior* **12**(2), 187–218.

Cox, J. C., Shachat, J. & Walker, M. (2001), 'An experiment to evaluate bayesian learning of nash equilibrium play', *Games and Economic Behavior* **34**(1), 11–33.

Crawford, V. & Broseta, B. (1998), 'What price coordination? the efficiency-enhancing effect of auctioning the right to play', *American Economic Review* pp. 198–225.

Crawford, V. P., Costa-Gomes, M. A. & Iriberri, N. (2013), 'Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications', *Journal of Economic Literature* **51**(1), 5–62.

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.

Dhami, S. (2016), *The foundations of behavioral economic analysis*, Oxford University Press.

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.

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

Esteban-Casanelles, T. & Gonçalves, D. (2020), The effect of incentives on choices and beliefs in games: An experiment, Technical report, Mimeo.

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

Fudenberg, D. & Tirole, J. (1991), *Game theory*, MIT press.

Gächter, S. & Renner, E. (2010), 'The effects of (incentivized) belief elicitation in public goods experiments', *Experimental Economics* **13**(3), 364–377.

Gill, D. & Rosokha, Y. (2020), 'Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma', *Available at SSRN 3652318* .

Gold, N. et al. (2012), 'Team reasoning, framing and cooperation', *Evolution and rationality: Decisions, co-operation and strategic behaviour* pp. 185–212.

Gonçalves, D. (2020), Diagonal games: A tool for experiments and theory, Technical report, Working Paper.

Haile, P. A., Hortaçsu, A. & Kosenok, G. (2008), 'On the empirical content of quantal response equilibrium', *American Economic Review* **98**(1), 180–200.

Ho, T.-H. & Su, X. (2013), 'A dynamic level-k model in sequential games', *Management Science* **59**(2), 452–469.

Huck, S. & Weizsäcker, G. (2002), 'Do players correctly estimate what others do?: Evidence of conservatism in beliefs', *Journal of Economic Behavior & Organization* **47**(1), 71–85.

Jordan, J. S. (1991), 'Bayesian learning in normal form games', *Games and Economic Behavior* **3**(1), 60–81.

Marchiori, D., Di Guida, S. & Polonio, L. (2021), 'Plasticity of strategic sophistication in interactive decision-making', *Journal of Economic Theory* **196**, 105291.

Mengel, F. (2014), 'Learning by (limited) forward looking players', *Journal of Economic Behavior & Organization* **108**, 59–77.

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

Nagel, R. (1995), 'Unraveling in guessing games: An experimental study', *The American economic review* **85**(5), 1313–1326.

Neugebauer, T., Perote, J., Schmidt, U. & Loos, M. (2009), 'Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments', *Journal of Economic Psychology* **30**(1), 52–60.

Nyarko, Y. & Schotter, A. (2002), 'An experimental study of belief learning using elicited beliefs', *Econometrica* **70**(3), 971–1005.

Pei, H., Yan, G. & Wang, H. (2021), 'Reciprocal rewards promote the evolution of cooperation in spatial prisoner's dilemma game', *Physics Letters A* **390**, 127108.

Polonio, L., Di Guida, S. & Coricelli, G. (2015), 'Strategic sophistication and attention in games: An eye-tracking study', *Games and Economic Behavior* **94**, 80–96.

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.

Selten, R. (1998), 'Features of experimentally observed bounded rationality', *European Economic Review* **42**(3-5), 413–436.

Selten, R., Abbink, K. & Cox, R. (2005), 'Learning direction theory and the winner's curse', *Experimental Economics* **8**(1), 5–20.

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.

Stahl, D. O. & Wilson, P. W. (1995), 'On players′ models of other players: Theory and experimental evidence', *Games and Economic Behavior* **10**(1), 218–254.

Sugden, R. (2003), 'The logic of team reasoning', *Philosophical explorations* **6**(3), 165–181.

Sugden, R. (2011), 'Mutual advantage, conventions and team reasoning', *International Review of Economics* **58**(1), 9–20.

Van Huyck, J. B., Battalio, R. C. & Beil, R. O. (1991), 'Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games', *The Quarterly Journal of Economics* **106**(3), 885–910.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., Inukai, K., Tanida, S., Kiyonari, T. et al. (2013), 'Is behavioral pro-sociality game-specific? pro-social preference and expectations of pro-sociality', *Organizational Behavior and Human Decision Processes* **120**(2), 260–271.

A Brief literature review for other learning models

Belief-based models are one family of adaptive learning models. Other adaptive learning models, such as the reinforcement learning (Roth & Erev 1995, Erev & Roth 1998), experience-weighted attraction (EWA) learning (Camerer et al. 1997, Camerer & Hua Ho 1999), and learning direction theory might also reach similar prediction by focusing on different aspects. Since it is not our focus, I will have a simple summary and comparison here.

A closely related model to belief-based learning is the learning direction model. Rigorously speaking, this model does not show belief explicitly as well. There is no clear specification on this model. It only provides a qualitative prediction of the direction of the next play relative to the current play. The well-known analogous 'archer shooting' that an archer exactly aims at the target by moving close to it little by little best describes this theory. The concept of impulse balance theory (Selten et al. 2005), summarized the theory that the long-term goal is to find a point (target), which has the same probability to move in either direction. Suppose an agent chooses an action s_i^t and his/her opponent(s) chooses actions s_{-i}^t, an agent i's ex-post best response b_i^t is

$$b_i^t \in \arg\max_{s_i \in S_i} \pi_i(s_i, s_{-i}^t) \tag{17}$$

Then for the next stage $t+1$, the agents' action choice s_i^{t+1} lies between b_i^t and s_i^t. Selten & Stoecker (1986) used this theory to explain the endgame effects in finitely repeated PD games. Subjects in their experiment played 25 supergame with 10 stages each. In their model, each agent is assumed to naturally play defection at stage k, given interaction begins with cooperation. So, agents lower down the first stage to defection k by one unit with probability α if he/she was defected earlier by his/her opponents or with probability β if he/she defects at the same stage with his/her opponent[19], or increase k by one unit with probability γ if he/she defects earlier than his/her opponent. They reported that the behaviour of 65% of subjects is consistent with the model. The median estimated values of α, β and γ are 0.5, 0.135, 0.225 respectively. They suggested that people are predicted to play defection immediately at the beginning of a supergame under these parameters.

A parallel model to belief-based learning is value-based learning, usually referred to as reinforcement learning. Agents using reinforcement learning are unaware of the structure of the game they played or the action one's opponents are going to play. Basically, reinforcement learning updates the propensity of action actually played by reinforcing prediction error with the learning rate. Agents always choose the action with the highest action value but do not understand the incentive behind their opponents' play and the next potential action. Dhami (2016) commented that using reinforcement learning only makes better decisions but not the optimal ones. A large number of empirical studies compared the fitness between belief-based learning and reinforcement learning but do not reach an agreement as either model performs better in some games not all. The EWA model contains core elements in both reinforcement learning and weighted fictitious play. It used the information of one's own play history[20] to form propensity to each own actions and allows the propensity of unplayed moves to be discounted [21]. From this point of view, EWA models can also be viewed as an overarching model of belief-based models and reinforcement learning

[19] $\alpha > \beta$ because agents were betrayed earlier are more motivated to change their behaviour

[20] A drawback of fictitious play is not

[21] A drawback of reinforcement learning is the unchanged propensity for actions unplayed

models. With some specific sets of parameters, EWA can be transformed into either case. A related model is the learning direction model. Rigorously speaking, this model does not show belief explicitly as well. There is no clear specification on this model. It only provides a qualitative prediction of the direction of the next play relative to the current play. The well-known analogous 'archer shooting' that an archer exactly aims at the target by moving close to it little by little best describes this theory. The concept of impulse balance theory (Selten et al. 2005), summarized the theory that the long-term goal is to find a point (target), which has the same probability to move in either direction. Suppose an agent chooses an action s_i^t and his/her opponent(s) chooses actions s_{-i}^t, an agent i's ex-post best response b_i^t is

$$b_i^t \in \arg\max_{s_i \in S_i} \pi_i(s_i, s_{-i}^t) \tag{18}$$

Then for the next stage $t+1$, the agents' action choice s_i^{t+1} lies between b_i^t and s_i^t. Selten & Stoecker (1986) used this theory to explain the endgame effects in finitely repeated PD games. Subjects in their experiment played 25 supergame with 10 stages each. In their model, each agent is assumed to naturally play defection at stage k, given interaction begins with cooperation. So, agents lower down the first stage to defection k by one unit with probability α if he/she was defected earlier by his/her opponents or with probability β if he/she defects at the same stage with his/her opponent[22], or increase k by one unit with probability γ if he/she defects earlier than his/her opponent. They reported that the behaviour of 65% of subjects is consistent with the model. The median estimated values of α, β and γ are 0.5, 0.135, 0.225 respectively. They suggested that people are predicted to play defection immediately at the beginning of a supergame under these parameters. A summary of how information is used in each model can be seen in Table 3

Table 3: Information used in each Learning model

	Learning theories				
Information	Reinforcement	Beliefs	Learning direction model	EWA	Sophistication
i's choice a_i^t	X		X	X	
$-i$'s choice a_{-i}^t		X		X	X
i's received payoff $\pi_i(s_i^t, s_{-i}^t)$	X	X		X	
i's forgone payoff $\pi_i(\hat{s_i^t}, s_{-i}^t)$		X		X	X
i's best response $b_i(s_{-i}^t)$			X		
$-i$'s received payoff $\pi_{-i}(s_i^t, s_{-i}^t)$					X
$-i$'s forgone payoff $\pi_i(s_i^t, \hat{s_{-i}^t})$					X

Note: From table 6.3 in Camerer et al. (2004*a*). I only picked the models I discussed in the dissertation, which are also the main models. The learning direction model is listed beyond the beliefs due to its unique information on the best response.

[22] $\alpha > \beta$ because agents were betrayed earlier are more motivated to change their behaviour