# Essays in Behavioral Economics and Game Theory

Thesis by

Meng-Jhang Fong

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended September 27, 2023

Meng-Jhang Fong
ORCID: 0009-0008-8832-3985

# ACKNOWLEDGEMENTS

I would like to first express my gratitude to my advisor, Marina Agranov, for her invaluable guidance, patience, and advice throughout my PhD study. I am also deeply indebted to Thomas R. Palfrey for his continuous mentorship and encouragement as my thesis committee chair and as a senior collaborator. I am grateful to Charles D. Sprenger and Kirby Nielsen for their insightful feedback on my research projects at different stages. I would also like to thank Luciano Pomatto for his support as a member of my committee.

I sincerely appreciate my close friend and collaborator, Po-Hsuan Lin, for his unwavering support in various areas, from research to life. I have been lucky to have Po as a friend since high school, which makes our bond even more special. In addition, my co-author on the first chapter, Wei James Chen, deserves special mention; without his backing, we could not have managed to run the experiment during the COVID-19 pandemic. My appreciation also goes to my master advisor, Joseph Tao-yi Wang, who has never hesitated to offer his assistance since introducing me to Caltech.

I am truly grateful to have Shunto J. Kobayashi and Danny Ebanks as companions on my PhD journey. Their treasured support, both academically and emotionally, goes beyond what I can express in a few sentences. I will never forget the picnics and conversations with Shunto during the pandemic, just as I will remember Danny's encouraging words when I faced uncertainties about the job market.

I am thankful to my roommates, Chen-Hsuan Lu, Kuang-Ming Shang, and Wen Chao, for their companionship, especially during the challenging times of the pandemic. I will always cherish those Chinese board-game nights.

Special thanks to I-Hui Huang, who helped me get through the last year of my PhD and motivated me to pursue a career outside of academia. Last but not least, I would like to convey my thanks to my parents, Chin-Tzu Fong and Yueh-Lien Tseng, as this journey would not have been possible without their care and love. The supportive and liberal environment of my family has played a pivotal role in shaping my identity.

# ABSTRACT

This thesis consists of three papers. Chapter 1 conducts experimental research on individual bounded rationality in games, Chapter 2 introduces a novel equilibrium solution concept in behavioral game theory, and Chapter 3 investigates confirmation bias within the framework of game theory.

In Chapter 1 (joint with Wei James Chen and Po-Hsuan Lin), we investigate individual strategic reasoning depths by matching human subjects with fully rational computer players in a lab, allowing for the isolation of limited reasoning ability from beliefs about opponent players and social preferences. Our findings reveal that when matched with robots, subjects demonstrate higher stability in their strategic thinking depths across games, in contrast to when matched with humans.

In Chapter 2 (joint with Po-Hsuan Lin and Thomas R. Palfrey), we investigate how players' misunderstanding about the relationship between opponents' private information and strategies influence their equilibrium behavior in dynamic environments. This theoretical study introduces a framework that extends the analysis of cursed equilibrium from the strategic form to multi-stage games and applies it to various applications in economics and political science.

In Chapter 3, I employ a game-theoretic framework to model how decision makers strategically interpret signals, particularly when they face a utility loss from holding beliefs that differ from their partners. The study reveals that the emergence of confirmation bias is positively associated with the strength of prior beliefs about a state, while the impact of signal accuracy remains ambiguous.

# PUBLISHED CONTENT AND CONTRIBUTIONS

Chen, W. J., Fong, M.-J., & Lin, P.-H. (2023). Measuring higher-order rationality with belief control. *arXiv preprint arXiv:2309.07427*. https://arxiv.org/abs/2309.07427
Meng-Jhang Fong contributed to the design of experiment, data analysis, and the preparation of the manuscript.

Fong, M.-J., Lin, P.-H., & Palfrey, T. R. (2023). Cursed sequential equilibrium. *arXiv preprint arXiv:2301.11971*. https://arxiv.org/abs/2301.11971
Meng-Jhang Fong contributed to the derivation of propositions, remark, and the preparation of the manuscript.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# SUMMARY INTRODUCTION

Motivated by studying the underlying mechanisms that drive human behavior to deviate from the predictions of standard economic theory, a substantial theme of my thesis centers around investigating how people form their beliefs and the profound impact belief formation has on individual decision-making processes.

Using choice data to infer an individual's strategic reasoning ability is challenging, as sophisticated players may hold non-equilibrium beliefs about others, leading to non-equilibrium behavior. In Chapter 1, *'Measuring Higher-Order Rationality with Belief Control'* (joint with Wei James Chen and Po-Hsuan Lin), we conduct an experiment to identify individual rationality bounds by matching human subjects with computer players known to be fully rational. The introduction of robot players enables the disentanglement of limited reasoning ability from belief formation and social preferences. Our findings reveal that, compared to interactions with humans, subjects exhibit a higher level of rationality and notable stability in rationality levels across games when matched with robots. This implies that strategic reasoning ability may be a persistent personality trait, and the revealed rationality bound could potentially serve as a proxy for gauging an individual's strategic ability when beliefs about others are properly controlled. In other words, the robot approach has the potential to become a standard tool for measuring a player's actual strategic thinking capacity.

Chapter 2, *'Cursed Sequential Equilibrium'*, (joint with Po-Hsuan Lin and Thomas R. Palfrey) develops a framework to extend the strategic form analysis of cursed equilibrium (CE) developed by Eyster and Rabin (2005) to multi-stage games. The approach uses behavioral strategies rather than normal form mixed strategies, and imposes sequential rationality. We define cursed sequential equilibrium (CSE) and compare it to sequential equilibrium and CE. We provide a general characterization of CSE and apply it to five applications in economics and political science. These applications illustrate a wide range of differences between CSE and Bayesian Nash equilibrium or CE: in signaling games; games with preplay communication; reputation building; sequential voting; and the dirty faces game where higher order beliefs play a key role. Several of these applications illustrate how and why CSE implies systematically cursed behavior in dynamic games of incomplete information with private values, while CE coincides with Bayesian Nash equilibrium for such games. A key mechanism underlying these distinct predictions is that players' misunderstandings of the association between others' types and actions affect the evolution of their beliefs about others' types over time.[1]

To examine the backfire effect of new information, Chapter 3, *'Conformity and Confirmation Bias'*, utilizes a game-theoretic framework to model how decision-makers strategically interpret a signal in the scenario where they experience a utility loss due to differing (posterior) beliefs among themselves. Specifically, we consider a two-player environment with two states, two signals, and two policy choices. The players have a common prior that is in favor of one state, and each player receives a signal before making her policy choice. However, a player may misinterpret the signal and form her posterior belief (and policy choice) accordingly. We characterize the conditions that support the following two types of equilibria: (i) Bayesian Updating Equilibrium (BUE), in which players always correctly interpret their signals; (ii) Confirmatory Bias Equilibrium (CBE), in which players always interpret the signal as supporting their prior beliefs. We show the existence of equilibria and examine how equilibrium conditions change in the strength of the prior belief and the accuracy of a signal. We find that the emergence of confirmation bias is positively associated with the strength of prior, whereas the impact of a signal's accuracy is ambiguous. Interestingly, when the final policy choice is relatively unimportant, higher signal accuracy can paradoxically increase the tendency to misinterpret conflicting evidence, as it incurs a higher cost of having misaligned posterior beliefs with a partner. This finding aligns with empirical evidence suggesting that exposure to conflicting information can sometimes backfire, reinforcing individuals' existing opinions.

[1] In a related work titled *'A Note on Cursed Sequential Equilibrium and Sequential Cursed Equilibrium'* (joint with Po-Hsuan Lin and Thomas R. Palfrey), we further compare our CSE framework with the sequential cursed equilibrium proposed by Cohen and Li (2023), which also extends the notion of cursedness to dynamic games but in different ways.

*Chapter 1*

# MEASURING HIGHER-ORDER RATIONALITY WITH BELIEF CONTROL

## 1.1 Introduction

Studying if people would make optimal choices in strategic interactions is a central research question in economics. Different from an individual decision-making problem, a game involves multiple players whose payoffs depend on each other's choice. Thus, to achieve a mutual best response, or an equilibrium, a player needs to be not only (first-order) rational—be able to take a payoff-maximizing strategy based on her belief about her opponents' actions—but also higher-order rational—be able to anticipate that her opponents are rational and that her opponents expect other opponents are also rational (and so on), and form her own belief and optimal strategies accordingly. Inspired by the well-documented evidence of systematic deviations from Nash-equilibrium play (see, for example, Camerer, 2003), behavioral research has thus proposed that the inability to engage in iterative reasoning and to exhibit infinite orders of rationality could be a critical factor that hinders individuals from reaching an equilibrium.

Beyond investigating how many steps of iterative reasoning one can perform, a recent study on bounded strategic sophistication by Georganas et al. (2015) asked another fundamental question: Can we observe stability in an individual's strategic reasoning depth across games? Measuring strategic reasoning abilities of interacting individuals can facilitate our understanding and prediction of individuals' behavioral patterns. It also helps us evaluate whether the observed non-equilibrium actions are driven by bounded rationality or by other factors. Nevertheless, if we observe no regularity when measuring one's strategic reasoning depth in different environments, there may not even exist such a persistent trait called "strategic thinking ability."

The main challenge behind inferring individual strategic reasoning ability from choice data is that the strategic sophistication revealed by one's choices does not directly imply how many steps of iterative reasoning one is able to perform. A player's observed depth of reasoning is determined not only by her reasoning capability but also by her beliefs about the opponents' (revealed) sophistication, as suggested by Agranov et al. (2012) and Alaoui and Penta (2016). An individual that can carry out more than $k$ steps of reasoning would act as a $k$th-order rational player when she believes that her opponent exhibits $(k-1)$th-order rationality. In other words, measuring an individual's revealed strategic sophistication only yields a lower-bound estimate of her actual sophistication. In addition, psychological factors other than bounded rationality, such as lying aversion and fairness concern may also motivate a player to deviate from an equilibrium. Without controlling for a player's beliefs and social preferences, the estimation of her strategic reasoning ability could be unstable and lack external validity. In fact, Georganas et al. (2015) had their subjects interact in two different families of games but found little persistence of individual strategic sophistication across the two games.

In this study, we demonstrate a way to recover the stability of individual strategic sophistication and to possibly pin down the upper bound of an individual's depth of iterative strategic reasoning in the lab: having human subjects interact with equilibrium-type *computer* players induced by infinite order of rationality. By informing human players that they are facing fully rational computer players, we are able to unify players' expectations about their opponents. In addition, introducing computer players precludes the possible effect of social preferences (Houser and Kurzban, 2002; Johnson et al., 2002; Van den Bos et al., 2008). Thus, human players with an infinite order of rationality are expected to select an equilibrium strategy. In this setting, out-of-equilibrium actions would provide us a solid ground to identify an individual's order of rationality for inferring her strategic reasoning ability since those actions are likely driven by bounded rationality.

To investigate the stability of individual strategic sophistication across games, we conduct an experiment with two classes of dominance-solvable games—ring games and guessing games. Proposed by Kneeland (2015) for identifying higher-order rationality, an $n$-player ring game can be characterized by $n$ payoff matrices and has the following ring structure: the $k$th player's payoff is determined by the $k$th player's and $(k+1)$th player's actions, and the payoff of the last ($n$th) player, who has a strictly dominant strategy, is determined by the last and the first player's actions. The guessing games we adopt are a symmetric variant of the two-person guessing games studied by Costa-Gomes and Crawford (2006), in which a player's payoff is single-peaked and maximized if the player's guess equals its opponent's guess times a predetermined number.

Among the games that have been used to study strategic reasoning, we choose to implement ring games and guessing games in our experiment for two reasons. First, our instruction of a fully rational computer player's behavior is tailored to align with the payoff structure of dominance-solvable games, in which the computer players' actions can be unambiguously determined (see Section 1.5 for details). Furthermore, these dominance-solvable games enable a structure-free identification approach, leveraging the notion of rationalizable strategy sets (Bernheim, 1984; Pearce, 1984). The core idea behind this identification approach is that, within a dominance solvable game, we can gauge an individual's depth of reasoning by assessing how many rounds of iterated deletion of dominated strategies the individual's chosen action would survive. Importantly, this approach does not impose structural assumptions on (the beliefs about) non-rational players' behavior. Therefore, these classes of games provide a plausible, structure-free method to empirically categorize individuals into distinct levels of rationality.

Second, we intend to implement two types of games that are sufficiently different so that, if we observe any stability in individual strategic reasoning levels across games, the stability does not result from the similarity between games. We believe that ring games and guessing games are dissimilar to each other. On one hand, a ring game is a four-player discrete game presented in matrix forms. On the other hand, a guessing game is a two-player game with a large strategy space, which is more like a continuous game. In fact, Cerigioni et al. (2019) report that the correlation of their experimental subjects' reasoning levels between ring games and beauty contest games (which also falls into the family of guessing games) is only 0.10.[1]

Our experiment comprises two treatments within each game family: the Robot Treatment and the History Treatment. In the Robot Treatment, subjects encounter computer players employing equilibrium strategies. In the History Treatment, subjects confront choice data from human players in the Robot Treatment. The History treatment simulates an environment where human subjects interact without displaying social preferences and serves two main objectives. First, by examining if a subject's observed order of rationality in the Robot Treatment exceeds that in the History Treatment, we can evaluate whether the subject responds to equilibrium-type computer players by employing a strategy that reaches her full capacity for strategic reasoning. Second, by comparing the individual orders of rationality inferred from data in both the Robot and History Treatments, we can investigate whether the introduction of robot players contributes to stabilizing observed strategic thinking levels across various games.

[1]We implement a guessing game with a single-peaked payoff structure instead of a standard beauty contest because the latter is not strictly dominant solvable.

Overall, our findings indicate that strategic reasoning ability may be a persistent personality trait deducible from choice data when subjects interact with robot players in strategic scenarios. Relative to interactions involving human opponents, we observe a larger proportion of participants adopting equilibrium strategies and demonstrating higher levels of rationality. This observation is supported by both our between- and within-subject statistical analyses, underscoring the effectiveness of our Robot Treatment and implying that the rationality depths exhibited in this treatment potentially approach subjects' strategic thinking capacity.[2]

Furthermore, our investigation reveals that subjects' rationality levels remain remarkably stable across distinct game classes when interacting with robot players. In terms of absolute levels, a substantial number of first-order and fourth-order rational players retain their respective types while transitioning from ring games to guessing games. In the Robot Treatment, approximately 38% of subjects exhibit constant rationality depths across games. A further statistical test involving 10,000 simulated samples demonstrates that this stability in rationality levels cannot be attributed to two independent type distributions, with the actual proportion of constant-level players exceeding the mean simulated proportion by 6 percentage points (19 percent). Additionally, applying the same statistical analysis to the History Treatment reveals no significant disparities in the proportions of constant-level players between actual and simulated datasets. This indicates that the stability in individual rationality depths is not solely due to game selection but is influenced by our manipulation of subjects' beliefs about opponents' strategic reasoning depths.

In terms of relative levels, the rankings of individual rationality levels also remain consistent in our Robot Treatment. When we randomly select two subjects from our experiment, the probability of their level order remaining unchanged across games in the Robot Treatment exceeds the probability of order switching by approximately 30 percentage points, or more than threefold. A similar pattern emerges in the History Treatment, where a constant ranking of two subjects' levels across games is observed more frequently than a switching ranking, although not as frequently as in the presence of robot players. Additionally, we find evidence suggesting that individual rationality levels in a game can serve as an indicator of the degree of game complexity when subjects interact with robot players. However, this evidence

[2]One might doubt if a subject has the motivation to act rationally upon the presence of an opponent with a (much) higher rationality level than the subject has. In Section 1.6, we argue that a subject does have the incentive to exhibit the highest order of rationality she can achieve when she knows her opponent is at least as rational as herself.

is comparatively less substantial and robust in the context of interactions with human opponents. In summary, the above results demonstrate that, when we use computer players to control for beliefs, the observed rationality levels of a subject may effectively capture her overall strategic thinking ability across various types of games.

A subject's performance in other cognitive tests could potentially hold predictive power regarding her strategic reasoning performance in games. As such, we incorporate tasks measuring cognitive reflection, short-term memory, and backward induction abilities (see Section 1.5 for details) into our experiment. We observe that a subject's cognitive reflection and backward induction abilities are positively correlated with her rationality depths, whereas no significant correlation is found with her short-term memory capacity.

The rest of the paper proceeds as follows. The next subsection reviews related literature. Section 1.2 summarizes the theoretical framework upon which our identification approach and hypotheses to be tested are based. Section 1.3 describes the ring games and guessing games implemented in our experiment. Section 1.4 discusses how we identify a subject's order of rationality given choice data. Section 1.5 presents our experimental design, including the experiment procedure and instruction for the robot strategy. Section 1.6 lists the hypotheses to be tested and discuss their implication. Section 1.7 reports the experiment results, and Section 1.8 concludes.[3]

**Related Literature**

Our work is closely related to Georganas et al. (2015) in terms of the model, experiment protocol, and data analysis procedure. In particular, we follow their model to formalize our hypotheses to be tested.[4] Georganas et al. (2015) examine experimentally whether a subject's sophistication type is persistent across games. Our work, however, differs from their work in several ways. First, we substitute the ring games for the undercutting games in Georganas et al. (2015) and use a simplified, symmetric version of the guessing games. Second, we employ an identification strategy distinct from the standard level-$k$ model to determine a subject's strategic sophistication. We use dominance solvable games in order to identify higher-order rationality without imposing strong and ad hoc assumptions on players' first-order

[3] An English translation of the experiment instructions can be found at https://mjfong.github.io.

[4] For a brief summary of the model in Georganas et al. (2015), see Section 1.2; also, see Section 1.6 for the hypotheses.

beliefs, which can in turn reduce the noise in the estimation of individual reasoning depth using a level-$k$ model.[5] More importantly, we control for human subjects' beliefs about opponents' sophistication (and social preferences) using computer players. As a result, we observe a higher correlation in subjects' types across games compared to Georganas et al. (2015), in which subjects are matched with each other.

Ring games, first utilized for identifying higher-order rationality by Kneeland (2015), are subsequently studied by Lim and Xiong (2016) and Cerigioni et al. (2019), who investigate two variants of the ring games. In this study, we follow the *revealed rationality approach* adopted by Lim and Xiong (2016) and Cerigioni et al. (2019) as our identification approach (discussed in Section 1.4). It is worth noting that Cerigioni et al. (2019) also find little correlation in subjects' estimated types across various games, including ring games, e-ring games, $p$-beauty contests, and a $4 \times 4$ matrix game. Again, our results suggest that the lack of persistence in the identified order of rationality at the individual level is driven by subjects' heterogeneous beliefs about the rationality of their opponents.

Indeed, several empirical studies have shown that beliefs about others' rationality levels can alter a player's strategy formation. Friedenberg et al. (2018) indicate that some non-equilibrium players observed in the ring games (Kneeland, 2015) may actually possess high cognitive abilities but follow an irrational behavioral model to reason about others. Alternatively, Agranov et al. (2012) and Alaoui and Penta (2016) find that, in their experiments, a subject's strategic behavior is responsive to the information she receives about her opponents' strategic abilities.[6] The designs of experiments allow them to manipulate subjects' beliefs, whereas we aim to elicit and identify individual strategic capability by unifying subjects' beliefs about opponents.

Some recent studies have tried to distinguish between non-equilibrium players who are limited by their reasoning abilities and players who are driven by beliefs. Identifying the existence of ability-bounded players is important since, if non-equilibrium behavior was purely driven by beliefs, it would be unnecessary to measure an individual's reasoning depth. Jin (2021) utilizes a sequential version of ring games, finding that around half of the second-order and third-order rational players are bounded by ability. Alaoui et al. (2020) also report the presence of ability-bounded sub-

[5]Burchardi and Penczynski (2014) conduct an experiment in a standard beauty contest with belief elicitation, finding heterogeneity in both level-0 beliefs and level-0 actions within a game.

[6]In Agranov et al. (2012), the subjects play against each other, graduate students from NYU Economics Department, or players taking uniformly random actions. In Alaoui and Penta (2016), the subjects play against opponents majoring in humanities, majoring in math and sciences, getting a relatively high score, or getting a low score in a comprehension test.

jects by showing that an elaboration on the equilibrium strategy shifts the subjects' level-$k$ types toward higher levels. Overall, the existence of both ability-bounded and belief-driven players in the real world indicates the need for an approach that can measure individual reasoning ability without the impact of beliefs. Whereas Jin (2021) and Alaoui et al. (2020) do not pin down the belief-driven players' actual ability limit, we aim to directly measure each subject's strategic reasoning ability.

Bosch-Rosa and Meissner (2020) propose an approach to test a subject's reasoning level in a given game: letting a subject play against herself (i.e., an "one-person" game). Specifically, their subject acts as both players in a modified two-person $p$-beauty contest (Grosskopf & Nagel, 2008), in which a player's payoff decreases in the distance between her guess and the average guess multiplied by $p$, and receives the sum of the two players' payoffs.[7] The one-person game approach eliminates the impact of beliefs that arises from interacting with human players. However, a limitation of this approach is that it can only be applied to the game in which the equilibrium is Pareto optimal. For instance, it would be rational for a payoff-maximizing subject to deviate from the equilibrium and choose (Cooperate, Cooperate) in the prisoner's dilemma since (Cooperate, Cooperate) maximizes the total payoff of both players even though those are not equilibrium strategies.[8] In this study, we employ an alternative approach that overcomes this limitation to measure rationality levels: letting a subject play against equilibrium-type computer players (i.e., the Robot Treatment).

Similar to the motivation of our Robot Treatment, Devetag and Warglien (2003), Grehl and Tutić (2015), and Bayer and Renou (2016) also employ rational computer players to mitigate the impact of beliefs and social preferences on individual decisions in their experiments. While Devetag and Warglien (2003) find a positive correlation between a subject's short-term memory performance and conformity to standard theoretical predictions in strategic behavior, Grehl and Tutić (2015) and Bayer and Renou (2016) explore a player's ability to reason logically about others' types in the incomplete information game known as the dirty faces game. In contrast, our study departs from theirs by focusing on investigating whether playing against computers can provide a robust measure of strategic reasoning ability across

[7] Bosch-Rosa and Meissner (2020) report that 69% of the subjects do not select the equilibrium action (0, 0) when playing the one-person game, which echoes the findings of the presence of ability-bounded players in Jin (2021) and Alaoui et al. (2020).

[8] Also note that in the ring game G1, both the equilibrium strategy profile (P1 chooses b; P2 chooses c; P3 chooses c; P4 chooses b) and another non-equilibrium strategy profile (P1 chooses a; P2 chooses b; P3 chooses a; P4 chooses a) lead to a total payoff of 66 (see Figure 1.1).

different families of games with complete information. Additionally, we also include a memory task to investigate whether the lack of significant predictive power of short-term memory on reasoning levels observed in Georganas et al. (2015) was influenced by uncontrolled beliefs and to offer a robustness check for the findings of Devetag and Warglien (2003) in different game settings.

In previous studies on strategic reasoning, equilibrium-type computer players have been introduced into laboratory experiments to induce human players' equilibrium behavior (e.g., Costa-Gomes and Crawford, 2006; Meijering et al., 2012) and to eliminate strategic uncertainty (e.g., Hanaki et al., 2016).[9] In contrast, our aim is to utilize computer players to uncover individual strategic reasoning ability. Notably, our experimental design avoids fully informing the subjects about either the notion of an equilibrium (Costa-Gomes & Crawford, 2006) or the computer player's exact strategy (Meijering et al., 2012; Hanaki et al., 2016), as such knowledge could potentially bias our estimation of individual strategic reasoning ability. Instead, following the approach of Johnson et al. (2002), in our Robot Treatment we inform subjects that the computer player is third-order rational (i.e., the computer is rational, knows its opponent is rational, knows its opponent knows it is rational) without disclosing further details (see Section 1.5). Our study contributes to the literature by demonstrating that introducing robot players can induce human subjects to exhibit stable reasoning levels across games, thus providing a solid foundation for measuring individual strategic thinking ability.

## 1.2 Theoretical Framework

### The Model in Georganas et al. (2015)

To formalize the idea of the depth of rationality and the hypotheses we are going to test, we introduce the model and notations used in Georganas et al. (2015). In their model, an $n$-person normal form game $\gamma \in \Gamma$ is represented by $(N, S, \{u_i\}_{i \in N})$, where $N = \{1, ..., n\}$ denotes the set of players, $S = S_1 \times \cdots \times S_n = \Pi_{i=1}^n S_i$ denotes the strategy sets, and $u_i : S \rightarrow R$ for $i \in N$ denotes the payoff functions. Following the notation in Georganas et al. (2015), we use $u_i(\sigma)$ to refer to $E_\sigma[u_i(\sigma)]$, where $\sigma = (\sigma_1, ..., \sigma_n)$, when $\sigma$ is a profile of mixed strategies (i.e., $\sigma_i \in \Delta(S_i)$).

Player $i$'s strategic ability is modeled by two functions $(c_i, k_i)$. Let $T$ be the set of *environmental parameters*, which captures the information a player observes about their opponents' cognitive abilities. The function $c_i : \Gamma \rightarrow \mathbb{N}_0$ represents $i$'s *capacity*

[9] For a survey of economics experiments with computer players, see March (2021).

for game $\gamma$, and the function $k_i : \Gamma \times T \to \mathbb{N}_0$ represents $i$'s (realized) *level* for game $\gamma$. A player's level for a game is bounded by her capacity, so $k_i(\gamma, \tau_i) \leq c_i(\gamma)$ for all $\gamma$, $\tau_i \in T$, and $i \in N$. The goal of our experiment is to measure $c_i(\gamma)$ and to test if $c_i(\gamma)$ (or $k_i(\gamma, \tau_i)$, after controlling for $\tau_i$) exhibits any stability across different games (see Section 1.6 for further discussion).

**Level-$k$ Model and Higher-Order Rationality**

In Georganas et al. (2015), a player's behavior is characterized by a standard level-$k$ model (Arad & Rubinstein, 2012; Crawford et al., 2013; Nagel, 1995; Stahl & Wilson, 1994, 1995). Specifically, let $\nu : \mathbb{N}_0 \to \Delta(\mathbb{N}_0)$ be a player's belief about her opponents' levels. In a standard level-$k$ model, $\nu(m) = \mathbb{1}\{m-1\}$ for all $m \geq 1$, and a level-0 player $i$'s strategy is exogenously given as $\sigma_i^0 \in \Delta(S_i)$. A level-$k$ ($k \geq 1$) player $i$'s strategy ($\sigma_i^k$) is defined inductively as a best response to $\nu(k)$. Formally, for all $s_i' \in S_i$, $\sigma_i^k$ satisfies $u_i(\sigma_i^k, \sigma_{-i}^{\nu(k)}) \geq u_i(s_i', \sigma_{-i}^{\nu(k)})$ where $\sigma_{-i}^{\nu(k)} = (\sigma_1^{k-1}, ..., \sigma_{i-1}^{k-1}, \sigma_{i+1}^{k-1}, ..., \sigma_n^{k-1})$. Notice that in order to pin down a level-$k$ player's strategy, we need to impose an assumption on the level-0 strategy. However, some studies have reported variations in level-0 actions and level-0 beliefs across individuals (Burchardi & Penczynski, 2014; Chen et al., 2018). Thus, an individual's identified level of reasoning can be sensitive to the structural assumptions under a level-$k$ model.

To avoid the ad hoc assumptions on level-0 players , we can instead define $k$th-order rationality (Bernheim, 1984; Lim & Xiong, 2016; Pearce, 1984) in the following way. Let $R_i^k(\gamma)$ be the set of strategies that survive $k$ rounds of iterated elimination of strictly dominated strategies (IEDS) for player $i$. In other words, a strategy $s_i$ is in $R_i^1(\gamma)$ if $s_i$ is a best response to some arbitrary $s_{-i}$, and $s_i$ is in $R_i^{k'}(\gamma)$ if $s_i$ is a best response to some $s_{-i} \in R_{-i}^{k'-1}(\gamma)$ for $k' > 1$. We say that a player $i$ exhibits *$k$th-order rationality* in $\gamma$ if and only if $i$ always plays a strategy in $R_i^k(\gamma)$. Equivalently, an individual exhibits $k$th-order rationality if and only if there is a $\sigma_{-i}^0$ such that the individual can be classified as a level-$k$ player in a standard level-$k$ model. Note that given any game $\gamma \in \Gamma$, $R_i^{k+1}(\gamma) \subset R_i^k(\gamma)$ for all $k \in \mathbb{N}_0$. In other words, a player exhibiting $k$th-order rationality also exhibits $j$th-order rationality for all $j \leq k$.

## 1.3 The Games

We study two classes of games: the four-player ring games used in Kneeland (2015) for identifying individuals' higher-order rationality and a variant of the two-person guessing games first studied by Costa-Gomes and Crawford (2006) and used in

Georganas et al. (2015) for identifying players' level-$k$ types.

### Ring Games

A four-player ring game is a simultaneous game characterized by four $3 \times 3$ payoff matrices. Figure 1.1 summarizes the structures of the two ring games, G1 and G2, used in our experiment. As shown in Figure 1.1, each player $i \in \{1, 2, 3, 4\}$ (simultaneously) chooses an action $a_i \in \{a, b, c\}$. Player 4 and Player 1's choices determine Player 4's payoff, and Player $k$ and Player $(k + 1)$'s choices determine Player $k$'s payoff for $k \in \{1, 2, 3\}$. Note that Player 4 has a strictly dominant strategy in each ring game ($b$ in G1 and $c$ in G2), and the two ring games are identical in the payoff matrices of Player 1, 2, and 3.

**G1**

Player 1

| Player 1's actions \ Player 2's actions | a | b | c |
|---|---|---|---|
| a | 8 | 20 | 12 |
| b | 0 | 8 | 16 |
| c | 18 | 12 | 6 |

Player 2

| Player 2's actions \ Player 3's actions | a | b | c |
|---|---|---|---|
| a | 14 | 18 | 4 |
| b | 20 | 8 | 14 |
| c | 0 | 16 | 18 |

Player 3

| Player 3's actions \ Player 4's actions | a | b | c |
|---|---|---|---|
| a | 20 | 14 | 8 |
| b | 16 | 2 | 18 |
| c | 0 | 16 | 16 |

Player 4

| Player 4's actions \ Player 1's actions | a | b | c |
|---|---|---|---|
| a | 6 | 10 | 8 |
| b | 12 | 16 | 14 |
| c | 8 | 12 | 10 |

**G2**

Player 1

| Player 1's actions \ Player 2's actions | a | b | c |
|---|---|---|---|
| a | 8 | 20 | 12 |
| b | 0 | 8 | 16 |
| c | 18 | 12 | 6 |

Player 2

| Player 2's actions \ Player 3's actions | a | b | c |
|---|---|---|---|
| a | 14 | 18 | 4 |
| b | 20 | 8 | 14 |
| c | 0 | 16 | 18 |

Player 3

| Player 3's actions \ Player 4's actions | a | b | c |
|---|---|---|---|
| a | 20 | 14 | 8 |
| b | 16 | 2 | 18 |
| c | 0 | 16 | 16 |

Player 4

| Player 4's actions \ Player 1's actions | a | b | c |
|---|---|---|---|
| a | 8 | 12 | 10 |
| b | 6 | 10 | 8 |
| c | 12 | 16 | 14 |

Figure 1.1: The Ring Games

Note: The Nash Equilibria are highlighted with colored borders.

Given the payoff structure, an individual that exhibits first-order rationality (i.e., maximizes her payoff based on some arbitrary belief) will always choose $b$ in G1 and $c$ in G2 when acting as Player 4. By eliminating dominated strategies, an individual exhibiting second-order rationality will always choose $c$ in G1 and $b$ in G2 when acting as Player 3. Then, by eliminating dominated strategies iteratively, an individual exhibiting third-order rationality will always choose $c$ in G1 and $a$ in G2 when acting as Player 2, and an individual exhibiting fourth-order rationality will always choose $b$ in G1 and $c$ in G2 when acting as Player 1. The unique Nash equilibrium of G1 is thus Player 1, 2, 3, and 4 choosing $b$, $c$, $c$, and $b$, respectively, and the unique Nash equilibrium of G2 is Player 1, 2, 3, and 4 chossing $c$, $a$, $b$, and $c$, respectively, as highlighted in Figure 1.1.

**Guessing Games**

In our experiment, the guessing game is a simultaneous two-player game parameterized by a constant $p \in (0, 1)$. We use $p = \frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$ in our experiment. Each player $i$ simultaneously chooses a positive integer $s_i$ between 1 and 100. Player $i$'s payoff strictly decreases in the difference between the number chosen by $i$, $s_i$, and the number chosen by $i$'s opponent multiplied by a constant $p$, $ps_{-i}$. Specifically, player $i$'s payoff is equal to $\frac{1}{5}(100 - |s_i - ps_{-i}|)$. Thus, a payoff-maximizing player's objective is to make a guess that matches her opponent's guess times $p$. Note that, given $p < 1$, any action (integer) greater than or equal to $100p + 1$ is strictly dominated by $[100p]$ (i.e., $argmax_{x \in \mathbb{N}} |100p - x|$) since $|[100p] - ps_{-i}| < |s'_i - ps_{-i}|$ for all $s_{-i} \in \{1, ..., 100\}$ and $s'_i \in \{[100p + 1], ..., 100\}$.[10]

Given the payoff function, a rational individual will always choose an integer between 1 and $[100p]$. An individual exhibiting second-order rationality will always choose a positive integer less than $[[100p] \cdot p]$, and so on. The unique equilibrium of the two-person guessing game is thus both players choosing 1.

## 1.4 Identification

Our model does not allow us to directly identify one's higher-order rationality from choice data. For example, an equilibrium player will choose 1 in the guessing game with $p = \frac{1}{2}$, while a player choosing 1 may have only performed one step of reasoning if her first-order belief is that her opponent guesses 2. Thus, observing a player $i$ choosing a strategy in $R_i^k(\cdot)$ for $k > 1$ (in a finite number of rounds) does not imply that $i$ exhibits $k$th-order rationality, which renders an individual's higher-order rationality unidentifiable. In fact, following the definition of $R_i^k(\cdot)$, we have $R_i^{k+1}(\cdot) \subset R_i^k(\cdot)$ for all $k \in \mathbb{N}_0$. Namely, every strategy (except for the dominated actions) can be rationalized by some first-order belief.

Following the rationale of higher-order rationality, we use the *revealed rationality approach* (Lim and Xiong, 2016; Brandenburger et al., 2019; Cerigioni et al., 2019) as our identification strategy. As explained below, this approach allows us to identify individual higher-order rationality in a dominance-solvable game. Under

[10] For instance, in a guessing game with $p = \frac{1}{3}$, every integer between 34 and 100 is dominated by 33; when $p = \frac{1}{2}$, every integer between 51 and 100 is dominated by 50; when $p = \frac{2}{3}$, every integer between 68 and 100 is dominated by 67.

the revealed rationality approach, we say that a player $i$ exhibits $k$th-order revealed rationality if (and only if) we observe the player actually playing a strategy that can survive $k$ rounds of IEDS, i.e., $s_i \in R_i^k(\cdot)$. A subject is then identified as a $k$th-order (revealed-)rational player when she exhibits $m$th-order revealed rationality for $m = k$ but not for $m = k + 1$. That is, a player is classified into the upper bound of her (revealed) rationality level.[11]

The idea behind the revealed rationality approach is the "as-if" argument: a subject $i$ selecting $s_i \in R_i^k(\cdot) \setminus R_i^{k+1}(\cdot)$ in finite observations behaves like a $k$th-order rational player, who always selects a strategy in $R_i^k(\cdot)$ but probably not in $R_i^{k+1}(\cdot)$, and thus is identified as a $k$th-order revealed rational player. Under this identification criterion, we can identify an individual's order of (revealed) rationality without requiring her to play in multiple games with different payoff structures. In our data analysis, we will classify subjects into five different types: first-order revealed rational (R1), second-order revealed rational (R2), third-order revealed rational (R3), fourth-order (or fully) revealed rational (R4), and non-rational (R0).[12] Tables 1.1 and 1.2 summarize the predicted actions under the revealed rationality approach for each type of players in our ring games and guessing games, respectively.

| | Ring Games | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | P1 | | P2 | | P3 | | P4 | |
| Type | G1 | G2 | G1 | G2 | G1 | G2 | G1 | G2 |
| R0 | N/A | | N/A | | N/A | | not (b, c) | |
| R1 | N/A | | N/A | | not (c, b) | | (b, c) | |
| R2 | N/A | | not (c, a) | | (c, b) | | (b, c) | |
| R3 | not (b, c) | | (c, a) | | (c, b) | | (b, c) | |
| R4 | (b, c) | | (c, a) | | (c, b) | | (b, c) | |

Table 1.1: Predicted Actions in the Ring Games Under the Revealed Rationality Approach

[11] Kneeland (2015) uses the *exclusion restriction* (ER) as its identification strategy, assuming that a player with low order rationality does not respond to changes in payoff matrices positioned away from herself. However, Lim and Xiong (2016) show that more than three-quarters of their experimental subjects change their actions in two identical ring games, which suggests the failure of the ER assumption since a rational player is predicted to take the same action in two identical games under the exclusion restriction. Also, the ER assumption does not facilitate the identification of higher-order rationality in the guessing games since we cannot separate out first-order payoffs from higher-order ones.

[12] In a four-player ring game, the highest identifiable (revealed) order of rationality is level 4.

| Type | Guessing Games | | |
|---|---|---|---|
| | $p$ = 1/3 | $p$ = 1/2 | $p$ = 2/3 |
| R0 | 34–100 | 51–100 | 68–100 |
| R1 | 12–33 | 26–50 | 46–67 |
| R2 | 5–11 | 14–25 | 31–45 |
| R3 | 2–4 | 8–13 | 21–30 |
| R4 (or above) | 1 | 1–7 | 1–20 |

Table 1.2: Predicted Actions in the Guessing Games Under the Revealed Rationality Approach

## 1.5 Design of the Experiment

### Protocol

We design a laboratory environment to measure the subjects' higher-order rationality. The experiment protocol is summarized in Figure 1.2. The experimental subjects did not receive any feedback about the outcomes of their choices until the end of the experiment.

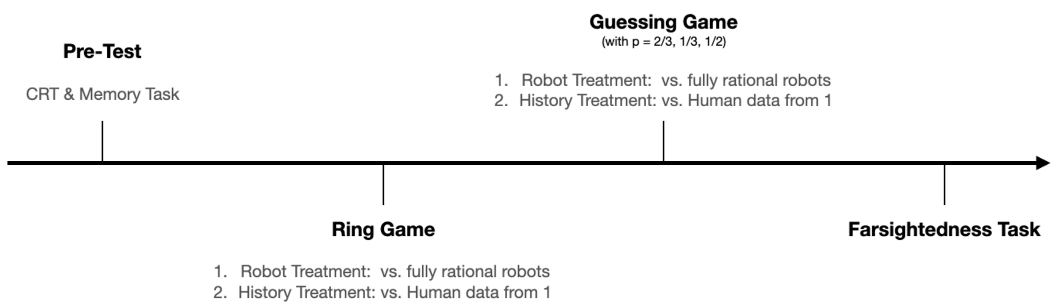


Figure 1.2: Experiment Protocol

At the beginning of the experiment, we let each subject complete two tasks that measure the cognitive abilities that have been found to be correlated with strategic reasoning abilities. The first task is the Cognitive Reflection Test (CRT) (Frederick, 2005), which is designed to evaluate the ability to reflect on intuitive answers. This test contains three questions that often trigger intuitive but incorrect answers. Georganas et al. (2015) report that the subjects' CRT scores have moderate predictive power on their expected earnings and level-$k$ types.

The second test is the Wechsler Digit Span Test (Wechsler, 1939), which is designed to test short-term memory. In our experiment, this test contains eleven rounds. In each round, a subject needs to repeat a sequence of digits, which is displayed on the screen at the rate of one digit every second. The maximum length of the digit sequence a subject can memorize reflects the subject's short-term memory capacity.[13] Devetag and Warglien (2003) find a positive correlation between individual short-term memory and strategic reasoning ability.

After completing the tasks, the subjects played the ring games in two different scenarios, the *Robot Treatment* and the *History Treatment*. To balance out potential spillover effects from one treatment to another, we alternated the order of the two scenarios across sessions. In each scenario, each subject played the two four-player, three-action ring games (G1 and G2 in Figure 1.1) in each position in each game once (for a total of eight rounds). Each subject was, in addition, assigned a neutral label (Member A, B, C, or D) before the ring games started. The label was only used for the explanation of an opponent's strategy in the History Treatment and did not reflect player position. To facilitate the cross-subject comparison, all the subjects played the games in the following fixed order: P1 in G1, P2 in G1, P3 in G1, P4 in G1, P1 in G2, P2 in G2, P3 in G2, and P4 in G2.[14] The order of payoff matrices was also fixed, with a subject's own payoff matrix being fixed at the leftmost side.[15] Note that the payoff structures of our ring games are the same as those in Kneeland (2015).[16] Adopting the same payoff structure facilitate the comparability between our and Kneeland's results.

In the Robot Treatment, the subjects played against fully rational computer players. Specifically, each subject in each round was matched with three robot players who only select the strategies that survive iterated dominance elimination (i.e., the equilibrium strategy). We informed the subjects of the presence of robot players that exhibit third-order rationality.[17] The instructions for the robot strategy are as follows:[18]

[13]The length of the digit sequence increases from three digits to thirteen digits round by round.

[14]Note that Player 4 has a dominant strategy in the ring game. We have our subjects play in each position in the reverse order of the IEDS procedure to mitigate potential framing effects resulting from the hierarchical structure.

[15]This feature is adopted in Jin (2021) and the main treatment of Kneeland (2015). Kneeland (2015) perturbs the order of payoff matrices in a robust treatment and finds no significant effects on subject behavior.

[16]The only difference is that we alter the order of the (row) payoffs for Player 4 in G1 to avoid coinciding, predicted actions given by a standard level-$k$ model. The actions $a, b, c$ in G1 (Figure 1.1) for Player 4 correspond to $c, a, b$, respectively, in the G1 in Kneeland (2015).

[17]Since we are not able to identify a subject's order of rationality above four in a four-player ring game, incorporating a third-order rational computer player is sufficient for the identification.

[18]Our instructions are adapted from the experiment instructions of Study 2 of Johnson et al. (2002). The original instructions are as follows: "In generating your offers, or deciding whether to accept or reject offers, assume the following: 1. You will be playing against a computer which is

*When you start each new round, you will be grouped with three other participants who are in different roles. The three other participants will be computers that are programmed to take the following strategy:*

1. *The computers aim to earn as much payoff as possible for themselves.*
2. *A computer believes that every participant will try to earn as much payoff as one can.*
3. *A computer believes that every participant believes "the computers aim to earn as much payoff as possible for themselves."*

The first line of a robot's decision rule ("The computers aim to...") implies that a robot never plays strictly dominated strategies and thus exhibits first-order rationality. The second line (along with the first line) indicates that a robot holds the belief that other players are (first-order) rational and best responds to such belief, which implies a robot's second-order rationality. The third line (along with the first and second lines) implies that, applying the same logic, a robot exhibits third-order rationality.

In the History Treatment, the subjects played against the data drawn from their decisions in the previous scenario. Specifically, in each round, a subject was matched with three programmed players who adopt actions chosen in the Robot Treatment by three other subjects. Every subject was informed that other human participants' payoffs would not be affected by her choices at this stage. By having the subjects play against past decision data, we can exclude the potential confounding effect of other-regarding preferences on individual actions.

After the ring games, the subjects played the two-person guessing games (in the order of $p = \frac{2}{3}, \frac{1}{3}, \frac{1}{2}$) in both the Robot Treatment and the History Treatment. Instead of being matched with three opponents, a subject was matched with only one player in the guessing games. The instructions for the guessing games in both treatments are revised accordingly.

At the last section of the experiment, we introduce an individual task developed by Bone et al. (2009)—the *farsightedness task*—to measure a subject's ability to do backward induction, or to anticipate her own future action and make the best choice

programmed to make as much money as possible for itself in each session. The computer does not care how much money you make. 2. The computer program expects you to try to make as much money as you can, and the program realizes that you have been told, in instruction (1) above, that it is trying to earn as much money as possible for itself" (p. 44-45).

accordingly. By comparing a subject's decisions in the ring games/guessing games and the farsightedness task, we can evaluate the correlation between one's depth of reasoning in a simultaneous game and that in a sequential decision task. We describe the details of the farsightedness task in the next subsection (Section 1.5).

There was a 180-second time limit on every subject's decisions in the ring games, guessing games, and farsightedness task. A subject who did not confirm her choice within 180 seconds earned zero payoff (for that round).[19]

The subjects were paid based on the payoffs (in ESC, Experimental Standard Currency) they received throughout the experiment. In addition to the payoff in the farsightedness task, one round in the ring games and one round in the guessing games were randomly chosen for payment. A subject also got three ESC for each correct answer in the CRT, and one ESC for each correct answer in the Digit Span Test.

**Farsightedness Task**

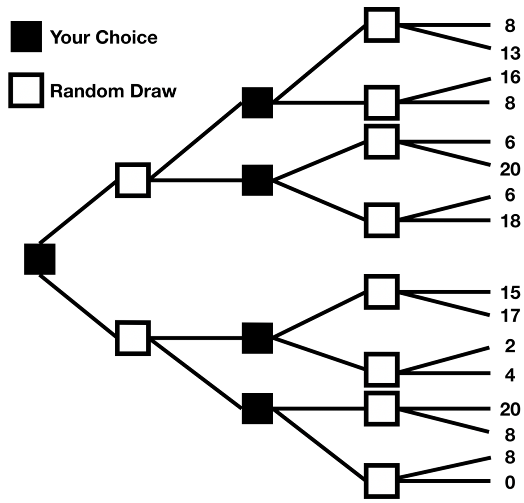


Figure 1.3: The Farsightedness Task in Bone et al. (2009)

The farsightedness task (Bone et al., 2009) is a sequential task that involves two sets of decision nodes and two sets of chance nodes (see the decision tree in Figure 1.3). The first and third sets of nodes are the decision nodes where a decision maker is going to take an action (up or down). The second and fourth sets of nodes are the chance nodes where the decision maker is going to be randomly assigned an action (with equal probability).

[19] Jin (2021) sets a 60-second time limit on decisions in the ring games and finds little effect on type classification.

Notice that there is one dominant action, in the sense of first-order stochastic dominance, at each of the third set of nodes (i.e., the second set of decision nodes). Anticipating the dominant actions at the second set of decision nodes, the decision maker also has a dominant action (down) at the first node. However, if a payoff maximizer lacks farsightedness and anticipates that each payoff will be reached with equal chance, then the dominated action (up) at the first node will become the dominant option from this decision maker's perspective. Therefore, a farsighted payoff-maximizer is expected to choose down, but a myopic one is expected to choose up, at the first move (and choose the dominant actions at the second moves).

**Laboratory Implementation**

We conducted 41 sessions between August 31, 2020 and January 28, 2021 at the Taiwan Social Sciences Experiment Laboratory (TASSEL) in National Taiwan University (NTU). Each session lasted about 140 minutes, and all participants were NTU students recruited through ORSEE (Greiner, 2015). A total of 299 subjects participated in the experiment, among whom 136 subjects played the Robot Treatment before the History Treatment in both families of games (RH Order) and 157 subjects played the History Treatment first (HR Order).[20] The experiment was programmed with the software zTree (Fischbacher, 2007) and instructed in Chinese. Including a show-up fee of NT$200 (approximately $7 in USD in 2020), the earnings in the experiment ranged between NT$303 and NT$554, with an average of NT$430.[21]

## 1.6 Hypotheses

The Robot Treatment is designed to convince subjects that the computer opponents they face are the most sophisticated players they could encounter. Consequently, if our Robot Treatment is effectively implemented, it should prompt subjects to employ a strategy at the highest achievable level $k$, i.e., $k_i(\gamma, \tau_i = Robot) = c_i(\gamma)$ for all $\gamma$ and $i$. (Recall that $k_i$ and $c_i$ denote subject $i$'s realized level and capacity, respectively.) This observation gives rise to the first hypothesis we aim to evaluate.

**Hypothesis 1 (Bounded capacity).** $k_i(\gamma, \tau_i') \leq k_i(\gamma, \tau_i = Robot)$ *for all* $\tau_i'$ *and* $\gamma$.

In words, we test whether subjects' rationality levels against robots capture individual strategic reasoning capacity. If Hypothesis 1 holds, then we can evaluate several

[20] Six subjects were dropped due to computer crashes.

[21] The exchange rate was 1 ESC for NT$4, and the foreign exchange rate was around US$1 = NT$29.4.

possible restrictions on $c_i$ by forming hypotheses on $k_i(\gamma, Robot)$. In evaluating Hypothesis 2 and 3, we study if we can observe any stable patterns of (revealed) individual depth of reasoning across games.

**Hypothesis 2** (**Constant capacity**). *$k_i(\gamma, Robot) = k_i(\gamma', Robot)$ for all $\gamma, \gamma'$.*

**Hypothesis 3** (**Constant ordering of capacity**). *For every $i, j \in N$, $k_i(\gamma, Robot) \geq k_j(\gamma, Robot)$ for some $\gamma$ implies $k_i(\gamma', Robot) \geq k_j(\gamma', Robot)$ for all $\gamma' \in \Gamma$.*

We first examine if a player's reasoning depth is constant across games, which is the strictest restriction on the stability of individual rationality levels (i.e., Hypothesis 2). If Hypothesis 2 does not hold, we can examine a weaker restriction and test whether the ranking of players (in terms of rationality levels) remains the same across games (i.e., Hypothesis 3). In other words, we examine if playing against robots can give us a measure of one's *absolute* level of depth of reasoning by evaluating Hypothesis 2, and a measure of *relative* level by evaluating Hypothesis 3.

The last restriction we will evaluate is whether the ordering of games (in terms of a player's rationality level) remain the same across players:

**Hypothesis 4** (**Consistent ordering of games**). *For every $\gamma, \gamma' \in \Gamma$, $k_i(\gamma, Robot) \geq k_i(\gamma', Robot)$ for some $i$ implies $k_j(\gamma, Robot) \geq k_j(\gamma', Robot)$ for all $j \in N$.*

In words, we examine if playing against robots can give us a measure of game difficulty (in terms of players' depth of reasoning) by evaluating Hypothesis 4.[22]

**Discussion**

An implicit assumption behind Hypothesis 1 is that a subject has an incentive to play a strategy with the maximum level she can achieve when encountering fully rational opponents that play at their highest reasoning level. This statement is trivially true for an equilibrium-type subject since she knows her opponents will play the equilibrium strategy and is able to best respond to it. However, it may or may not be true for a bounded rational player. If one believes that an iterative reasoning model describes an individual's actual decision-making process, there are two possible reasons that a player is only able to perform $k$ steps of iterative reasoning. First, she may incorrectly believe that other players can exhibit (at most) $(k-1)$th-order of rationality and best respond to such belief. Second, she may correctly perceive that

[22] Our Hypothesis 2, 3, and 4 correspond to Restriction 2, 3, and 5 in Georganas et al. (2015), respectively (see p. 377).

other players can exhibit (at least) $k$th-order of rationality but fail to best respond to it. While our statement regarding incentive compatibility still holds in the first case, it becomes unclear how a bounded rational player would respond when confronted with a player exhibiting an order of rationality above $k$.

Nevertheless, we argue that this case will not be a problem under the identification strategy of the revealed rationality approach. Notice that a player who exhibits $k$th-order rationality would also exhibit $m$th-order rationality for all $m \leq k$. Thus, a level-$k$ individual $i$ who perceives that other players are exhibiting (at least) $k$th-order rationality would also perceive that they are exhibiting $(k-1)$th-order rationality. That is, she knows that her robot opponents' strategies will survive $k-1$ rounds of IEDS. Thus, a payoff-maximizing player $i$ who is able to perform $k$ steps of iterative reasoning will choose some strategy in $R_i^k(\cdot)$, which contains all the undominated strategies after $k-1$ rounds of IEDS. Under the revealed rationality approach, player $i$ will then be classified as a $k$th-order revealed-rational player.

## 1.7 Experiment Results

### Data Description

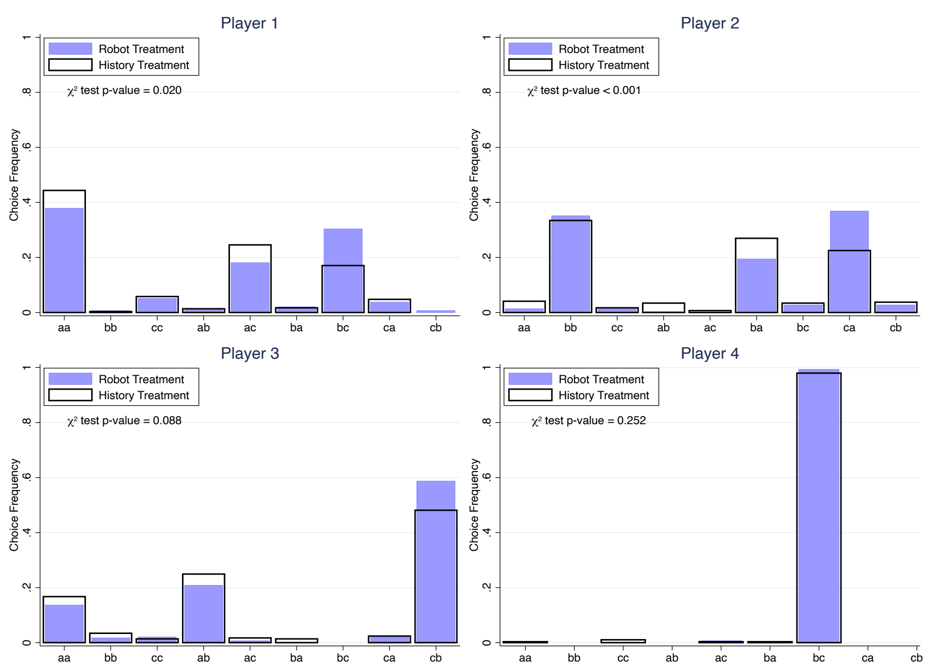


Figure 1.4: Ring Game Choice Frequency at Each Position

Before delving into the main results, we begin by summarizing the subjects' choice frequencies in the ring games (Figure 1.4) and guessing games (Figure 1.5). Figure

1.4 reports the subjects' choice frequencies in the two ring games (G1 and G2, see Figure 1.1) at each player position. It is noteworthy that, first, over 97% of subjects choose the strictly dominant strategy ((action in G1, action in G2) = ($b$, $c$) at the Player 4 position) in both treatments, which suggests a clear understanding of the payoff structure of ring games and the ability to recognize strict dominance. Second, at each player position except P4, the Robot Treatment exhibits a 10 to 15 percentage points higher frequency of subjects choosing the equilibrium strategy (($b$, $c$), ($c$, $a$), ($c$, $b$) at P1, P2, P3, resp.) compared to the History Treatment, which suggests that the Robot Treatment effectively prompts subjects to display higher levels of rationality. Third, at each player position except P4, a notable proportion of subjects choose the action that maximizes the minimum possible payoff among the three available actions ($a$ at P1, $b$ at P2, $a$ at P3). It is also worth noting that in G1, the minimum possible payoffs of equilibrium actions (except at P4) are all 0. As reported in Figure 1.4, there is a higher proportion of subjects choosing the equilibrium actions in G2 but maxmin actions in G1 (($a$, $c$), ($b$, $a$), ($a$, $b$) at P1, P2, P3, resp.) in the History Treatment compared to the Robot Treatment. This evidence suggests that when players have uncertainty about their opponents' reasoning and strategic behavior, some players may opt for a non-equilibrium strategy to avoid the possibility of experiencing the worst possible payoff.

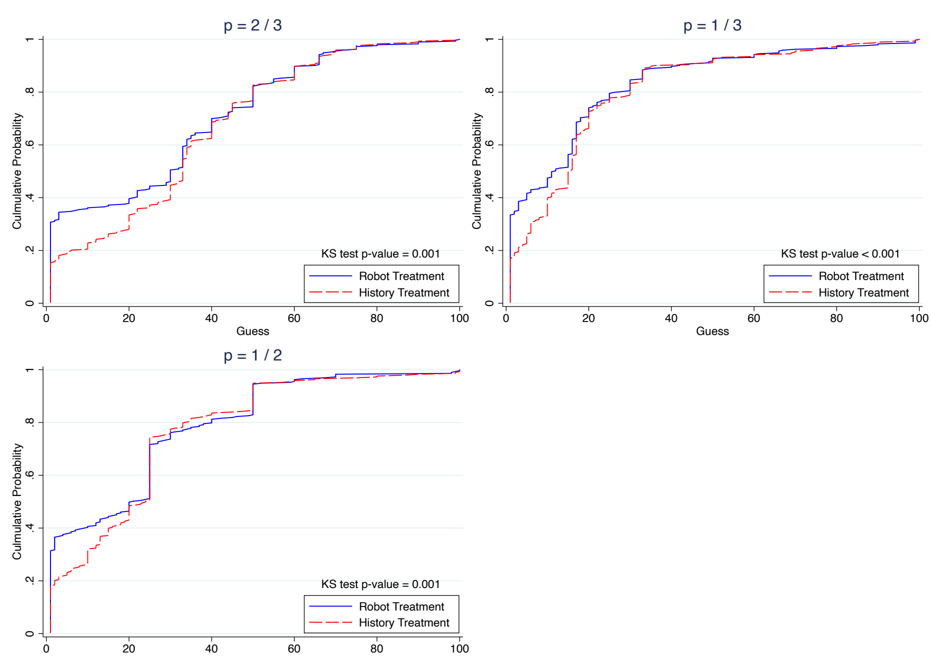


Figure 1.5: Cumulative Distribution of Guesses

Figure 1.5 summarizes the subjects' guesses in the three guessing games, depicted in terms of cumulative distribution. Consistent with the observations from the ring games, we observe a 13 to 17 percentage points higher proportion of subjects making the equilibrium guess (i.e., choosing 1) across the three guessing games in the Robot Treatment compared to the History Treatment. This difference in proportions results in the (first-order stochastic) dominance of the cumulative distribution of guesses in the Robot Treatment over that in the History Treatment, suggesting a higher level of rationality exhibited by the subjects in the Robot Treatment for the guessing games as well. In the subsequent subsection, we will describe our approach for classifying individual rationality levels and perform statistical tests to assess whether subjects demonstrate higher orders of rationality when playing against robots.

**Type Classification**

We adopt the revealed rationality approach to classify subjects into different rationality levels. Specifically, let $s_i = (s_i^\gamma)$ be the vector which collects player $i$'s actions in each family of games $\gamma$, where $\gamma \in \{Ring, Guessing\}$. In the ring games, we classify subjects based on the classification rule shown in Table 1.1. In both the Robot Treatment and the History Treatment, if a subject's action profile matches one of the predicted action profiles of type R0–R4 exactly, then the subject is assigned that type. Therefore, we can obtain each subject's type in the Robot Treatment and the History Treatment, which are denoted as $k_i(Ring, Robot)$ and $k_i(Ring, History)$, respectively.

Similarly, for the guessing games, we classify subjects based on the rule outlined in Table 1.2. In both treatments, each subject makes three guesses (at $p = \frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$). If a subject is categorized into different types in different guessing games, we assign her the lower type. Thus, we can obtain the types in both treatments, denoted as $k_i(Guessing, Robot)$ and $k_i(Guessing, History)$, respectively. Following this rationale, we construct the overall distribution of individual rationality levels for each treatment by assigning each subject the lower type she exhibits across the two classes of games, i.e., $k_i(\tau_i) = \min\{k_i(Ring, \tau_i), k_i(Guessing, \tau_i)\}$.

**Type Distributions**

Figure 1.6 reports the distributions of rationality levels in the two treatments for both ring games and guessing games. As shown in the top of Figure 1.6, subjects tend to be classified into higher types when playing against robots. There are more R1 and R2 players but fewer R3 and R4 players in the History Treatment than in the Robot Treatment. To examine if a subject's rationality depth is bounded by her revealed rationality level in the Robot Treatment (Hypothesis 1), at the aggregate level, we conduct the two-sample Kolmogorov-Smirnov test to compare the distributions of rationality levels in the two treatments. If Hypothesis 1 holds, we should observe either no difference in the two distributions or the distribution in the Robot Treatment dominating the distribution in the History Treatment. Our results show that the underlying distribution of individual rationality levels in the Robot Treatment stochastically dominates the one in the History Treatment (K-S test $p = 0.015$ for ring games and $p = 0.001$ for guessing games), and thus provide supporting evidence for Hypothesis 1.

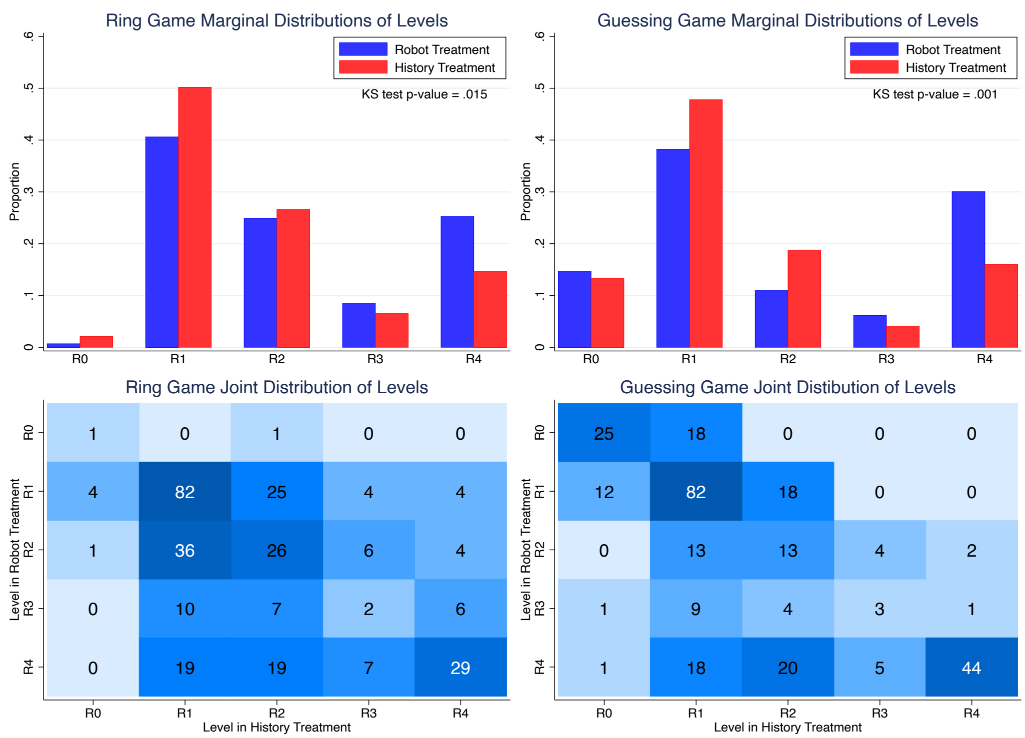


Figure 1.6: Frequency of Rationality Levels in Ring Games (Left) and Guessing Games (Right)

Moreover, our within-subject design gives us paired data of individual rationality types across treatments (as summarized in the bottom of Figure 1.6), which gives us another way to test Hypothesis 1. Overall, 72 percent of subjects (211/293) exhibit (weakly) higher rationality levels in the Robot Treatment than in the History Treatment in both families of games. In contrast, only fewer than four percent of subjects (11/293) consistently exhibit strictly lower rationality levels in the Robot Treatment across games. We further conduct the Wilcoxon signed-rank test to examine whether the subjects' rationality levels in the Robot Treatment are significantly greater than the History Treatment. Consistent with Hypothesis 1, we observe higher rationality levels in the Robot Treatment (Wilcoxon test $p < 0.0001$ for both ring games and guessing games). Therefore, we conclude that the rationality levels in the Robot Treatment for a game can serve as a proxy of individual strategic reasoning capacity in that game.

In the guessing games, our classification results display a typical distribution pattern of estimated levels as documented in Costa-Gomes and Crawford (2006) and Georganas et al. (2015). First, the modal type is R1 (Level 1), with more than 35 percent of subjects classified as R1 players in both treatments (Robot: 38.23%; History: 47.78%; Costa-Gomes and Crawford: 48.86%; Georganas et al.: 50.00%). In particular, the proportion of R1 players reported in the History treatment of our guessing games is very close to the proportion of level-1 players reported in Costa-Gomes and Crawford (2006) and Georganas et al. (2015). Second, R3 (Level 3) represents the least frequently observed category among the rational types (i.e., R1–R4), with fewer than 10 percent of subjects classified as R3 players in both treatments, a proportion that aligns with findings in the literature. (Robot: 6.14%; History: 4.10%; Costa-Gomes and Crawford: 3.41%; Georganas et al.: 10.34%). Third, the percentage of R4 players in our History Treatment falls within the range of equilibrium-type player proportions reported in Costa-Gomes and Crawford (2006) and Georganas et al. (2015) (Robot: 30.03%; History: 16.04%; Costa-Gomes and Crawford: 15.91%; Georganas et al.: 27.59%). Noticeably, in our Robot Treatment, we observe a relatively high frequency of R4 players compared to previous literature.[23] This finding underscores the significant impact of non-equilibrium belief about opponents on non-equilibrium behavior.

It is noteworthy that, contrary to previous findings, we observe very few R0 players in the ring games in both treatments (Robot: 0.68%; History: 2.04%).[24] In our experiment, the subjects do not interact with each other in both treatments. Thus, our observation suggests that, when human interactions exist, social preferences may play some roles in a ring game and lead to (seemingly) irrational behavior,

[23] For instance, Arad and Rubinstein (2012) also note that, in their 11–20 game, the percentage of subjects employing more than three steps of iterative reasoning does not exceed 20 percent. This aligns with the proportion of R4 players identified in our History Treatment but is lower than that in our Robot Treatment.

[24] Kneeland (2015) observes 6 percent of R0 players (with the ER approach) and Cerigioni et al. (2019) observe more than 15 percent of R0 players (with the revealed rationality approach) in their experiments.

though we cannot exclude the possibility that this discrepancy in the prevalence of R0 players is due to different samples.

### Constant Absolute Rationality Levels

In this subsection, we evaluate the hypothesis that an individual has constant strategic reasoning capacity across games in the Robot Treatment (i.e., Hypothesis 2). Figure 1.7 reports how frequently a ring game player with a rationality depth is classified into the same or another type in the guessing games. If the observed individual rationality level is the same across games, then every diagonal entry of each transition matrix in Figure 1.7 will be 100(%). Alternatively, if subjects' rationality orders in the ring games and guessing games are uncorrelated, every row in a transition matrix will be the same and equals the overall distribution in the guessing games.

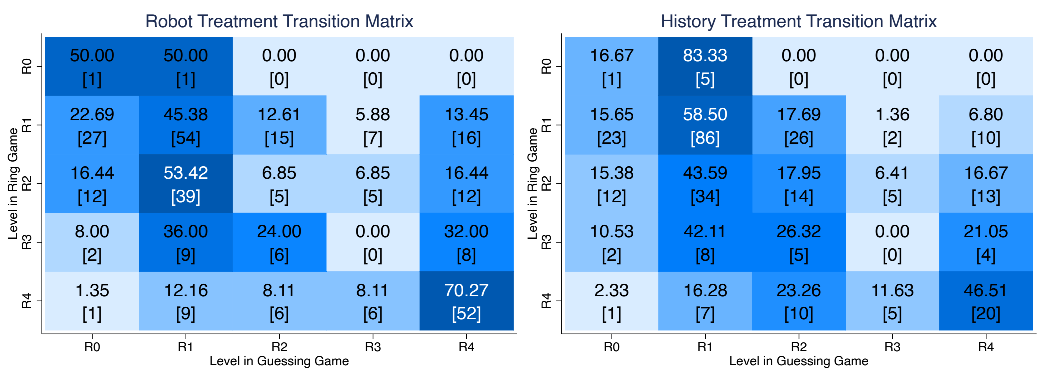


Figure 1.7: Transition Matrix for Rationality Levels in Both Treatments (from Ring to Guessing Games)

The transition matrices for both treatments show that most R1 and R4 players in the ring games remain as the same type in the guessing games. Most R2 ring game players, however, only exhibit first-order rationality in the guessing games. We do not observe any subjects consistently classified into R3 for both ring and guessing games, possibly because we have relatively low numbers of R3 subjects in either games. Overall, there is a relatively high proportion of subjects that exhibit the same rationality depth across games in both treatments (Robot: 38.23%; History: 40.27%).[25] Note that in the Robot Treatment, we observe a relatively high proportion (52/293 = 17.74%) of subjects classified as R4 players in both games,[26] suggesting

[25] Georganas et al. (2015) report that only 27.3% of their subjects play at the same level across two families of games.

[26] In the History Treatment, constant R4 players across games constitute only 6.82% (20/293) of the subjects.

that subjects in our experiment understand the instruction for robots' decision rules and try to play the best response to such rules.

To test if the high proportion of constant-level players actually results from independent type distributions, we generate 10,000 random samples of 293 pairs of levels, independently drawn from the empirical distribution of rationality levels in the Robot Treatment. The simulated datasets provide a distribution of the frequency with which a subject plays at the same level in both game families. The mean frequency is found to be 32.86%, with a 95 percent confidence interval ranging from 27.65% to 38.23%. The observed frequency is 38.23%, rejecting the null hypothesis that the subjects' rationality depths are independently distributed across game families in terms of absolute rationality depths, at a significance level close to 5% ($p = 0.057$). Thus, we conclude that the hypothesis stating an individual exhibits constant depths of rationality across games has predictive power (despite not being perfectly accurate) regarding experimental subjects' actions under proper belief control.

To establish a baseline for comparison, we utilize the same Monte Carlo simulation and statistical test outlined above to investigate whether the restriction of constant rationality depth can be applied to modeling subjects' actions when they face human opponents (choice data) instead of robots in the History Treatment. Upon examining the pooled data in the History Treatment, we find that the null hypothesis of independently distributed rationality levels cannot be rejected despite the seemingly high proportion of constant-level players. The simulated samples generated from the data in the History Treatment exhibit an average of 40.27% constant-level players (95% CI = $[34.81\%, 45.73\%]$), and the observed frequency in the actual data is 41.30% ($p = 0.768$). This finding suggests that the high stability observed in individual (absolute) rationality depth within our experiment does not result from the specific games selected. Moreover, it indicates that unifying subjects' beliefs about opponents' strategic reasoning depth effectively stabilizes the individual revealed order of rationality across games.

**Constant Ordering of Rationality Levels**

In this subsection, we evaluate the hypothesis that the ranking of individual strategic reasoning capacity between two players in different game families is the same (i.e., Hypothesis 3). Table 1.3 reports the switch frequency, non-switch frequency, and switch ratio observed in the actual data and computed under the null hypothesis of independently distributed rationality levels. The *switch frequency*, as defined in Georganas et al. (2015), represents the proportion of player pairs in which the player who exhibits a strictly higher level in one game becomes the player with a strictly lower level in another game. On the other hand, the *non-switch frequency* corresponds to the proportion of player pairs in which the player with a strictly higher level in one game maintains that higher level in another game.[27] The *switch ratio* is calculated by dividing the switch frequency by the non-switch frequency. If the relative rationality levels are preserved across games, the switch ratio will be zero. Alternatively, if the rationality levels are independently drawn, we expect to observe a switch ratio of one.

| | Pooled Data ($n$ = 293) | | RH Order ($n$ = 136) | | HR Order ($n$ = 157) | |
|---|---|---|---|---|---|---|
| Ring Game vs. Guessing Game | Empirical Data | Null Hypothesis | Empirical Data | Null Hypothesis | Empirical Data | Null Hypothesis |
| Robot Treatment | | | | | | |
| Switch frequency: | 12.3% | 22.5% | 11.9% | 19.8% | 12.5% | 24.0% |
| Non-switch frequency: | 41.3% | 22.5% | 37.7% | 19.8% | 42.3% | 24.1% |
| Switch ratio: | 0.30 | 1.01 | 0.32 | 1.03 | 0.29 | 1.02 |
| $p$-value: | < 0.0001 | | < 0.0001 | | < 0.0001 | |
| History Treatment | | | | | | |
| Switch frequency: | 12.9% | 17.9% | 11.0% | 21.2% | 14.8% | 14.5% |
| Non-switch frequency: | 34.5% | 17.8% | 40.3% | 21.3% | 28.1% | 14.5% |
| Switch ratio: | 0.37 | 1.02 | 0.27 | 1.02 | 0.53 | 1.04 |
| $p$-value: | < 0.0001 | | < 0.0001 | | 0.019 | |

Table 1.3: Switch Ratio for the Robot and History Treatment

The results presented in Table 1.3 provide compelling evidence of stable rankings of individual rationality levels. Our pooled data show that non-switching occurs three times more frequently than switching in the Robot Treatment (Non-switching: 41.30%; Switching: 12.28%). The switch ratio of 0.30, derived from the switch and non-switch frequencies, is lower than any switch ratio obtained from our 10,000-sample simulated data. Additionally, these results consistently hold across different treatment orders. Whether the Robot Treatment is played first or second, the observed switch ratios remain around 0.30, both of which are lower than any switch ratio obtained from the simulated data. Consequently, we reject the null hypothesis of independently distributed levels in terms of relative rationality depths, with a $p$-value less than 0.0001.

[27]The sum of the switch frequency and non-switch frequency may not be one since the paired players who exhibit the same level in one game are excluded.

We also calculate the switch and non-switch frequencies in the History Treatment to investigate whether the rankings of individual rationality levels remain stable when subjects' beliefs about others' rationality depths are not controlled. In the History Treatment, the null hypothesis of independently distributed levels in terms of relative rationality depths is also rejected ($p < 0.0001$), with the pooled data showing switch and non-switch frequencies of 12.89% and 34.47%, respectively, resulting in a switch ratio of 0.37. However, it is noteworthy that the switch ratio in the History Treatment is 23% higher than that in the Robot Treatment, and this difference increases to 66% when focusing solely on the Robot and History Treatments that are played first by subjects (Robot: 0.32; History: 0.53).[28] This result mainly stems from the fact that the non-switch frequency in the Robot Treatment is substantially higher than that in the History Treatment. These findings once again suggest that unifying subjects' beliefs about the strategic reasoning capability of their opponents can significantly improve the stability observed in individual rationality levels across games.

To summarize, our Robot Treatment reveals stability in both the subjects' absolute and relative rationality depths. These findings indicate that strategic reasoning ability could be an inherent personal characteristic that can be inferred from choice data when participants interact with robot players.

**Persistence of Ordering of Games**

In this subsection, we evaluate the hypothesis that the ranking of games in terms of individual strategic reasoning capacity is the same (i.e., Hypothesis 4). Table 1.4 reports the change-in-same-direction frequency, change-in-opposite-directions frequency, and the opposite/same ratio computed based on actual data and simulated data generated from independently-drawn levels. The *change-in-same-direction frequency* represents the proportion of player pairs in which both players exhibit a strictly higher level in the same game. On the other hand, the *change-in-opposite-directions frequency* refers to the proportion of player pairs in which the two players exhibit a strictly higher level in different games (Georganas et al., 2015).[29] The

[28] We conduct a statistical comparison by contrasting a switch ratio of 0.32 with the switch ratios obtained from 10,000 random samples of independently drawn levels from the empirical distribution of rationality levels in the History Treatment under HR Order. Our analysis reveals that, when focusing exclusively on the data from treatments played first, we can reject the null hypothesis that the observed rationality levels in the Robot Treatment are drawn from the same distribution of levels as in the History Treatment, in terms of switch ratios ($p = 0.027$).

[29] Again, the sum of the change-in-same-direction frequency and change-in-opposite-directions frequency may not be one since a pair of players is excluded if one of them exhibits the same level

*opposite/same ratio* is calculated by dividing the change-in-opposite-directions frequency by the change-in-same-direction frequency. In the case of a constant ranking of games across players, the opposite/same ratio would be zero. Conversely, if the rationality levels are independently drawn, we would expect the opposite/same ratio to be one.

| | Pooled Data ($n$ = 293) | | RH Order ($n$ = 136) | | HR Order ($n$ = 157) | |
|---|---|---|---|---|---|---|
| Ring Game vs. Guessing Game | Empirical Data | Null Hypothesis | Empirical Data | Null Hypothesis | Empirical Data | Null Hypothesis |
| Robot Treatment | | | | | | |
| Change in opposite direction: | 17.5% | 22.6% | 18.7% | 19.8% | 16.5% | 24.1% |
| Change in same direction: | 20.6% | 22.6% | 20.2% | 19.8% | 20.8% | 24.1% |
| Opposite/same ratio: | 0.85 | 1.00 | 0.93 | 1.00 | 0.79 | 1.00 |
| $p$-value: | < 0.0001 | | 0.079 | | 0.0004 | |
| History Treatment | | | | | | |
| Change in opposite direction: | 16.3% | 17.9% | 17.1% | 21.2% | 15.6% | 14.5% |
| Change in same direction: | 18.1% | 17.9% | 18.2% | 21.3% | 17.8% | 14.5% |
| Opposite/same ratio: | 0.90 | 1.00 | 0.94 | 1.00 | 0.88 | 1.00 |
| $p$-value: | 0.002 | | 0.112 | | 0.025 | |

Table 1.4: Opposite/same Ratio for the Robot and History Treatment

In the Robot Treatment, the frequency with which two paired players change their rationality levels in the same direction (20.58%) is 3 percentage points higher than the frequency of changing in the opposite directions (17.58%), as shown in Table 1.4 (the column of Pooled Data). The observed opposite/same ratio of 0.85 significantly deviates from the mean of the simulated datasets (1.00 with a 95 percent confidence interval of 0.96 to 1.01), leading to the rejection of the null hypothesis of independently distributed levels in terms of the ordering of games ($p < 0.0001$). This result remains robust regardless of the order of treatments, although it only reaches marginal significance when the analysis is limited to the subjects who played the Robot Treatment first (RH Order: $p = 0.079$; HR Order: $p = 0.0004$). Consequently, our findings suggest that an individual's strategic reasoning level, measured under an environment where a player's belief is well controlled, could serve as a reliable proxy for the (relative) complexity or difficulty of a game.

In the History Treatment, we find a similar result but with weaker evidence. The simulated datasets generated from the History Treatment data yield a mean opposite/same ratio of 1.00, with a 95 percent confidence interval of 0.95 to 1.01. The actual ratio of 0.90, which is 6% higher than that in the Robot Treatment, still rejects

across games.

the null hypothesis of independently distributed levels with a significance level of $p = 0.002$. However, this result becomes less robust when considering the order of treatments. We cannot reject the null hypothesis when examining only the subjects who played against robots before playing against human choice data (RH Order: $p = 0.112$; HR Order: $p = 0.025$). Without controlling for a subject's belief about her opponents' strategic thinking abilities, the observed rationality level could reflect either the complexity of the environment or how a subject believes others would perceive the complexity of the environment, and thus has weaker predictive power on other players' (revealed) order of rationality.

**Cognitive Tests, Farsightedness, and Strategic Thinking**

If individual strategic sophistication is persistent across games, a natural next question will be whether a subject's performance in other cognitive tests or strategic tasks can predict her strategic reasoning ability. Accordingly, we run regressions of revealed rationality levels on subjects' CRT scores, short-term memory task scores, and farsightedness task scores (see Figure 1.8). We cluster the regression standard errors at the session level.

The definitions of the independent variables are as follows: *CRT Score* (ranging from 0 to 3) represents the number of correct answers a subject gets in the three CRT questions. *Memory Score* (ranging from 0 to 11) is defined as the number of correct answers a subject provides before making the first mistake. *Farsightedness* is an indicator variable that equals one if a subject chooses to go down at the first move in the farsightedness task (see Section 1.5). Last, the dependent variable is the individual rationality level (ranging from 0 to 4) revealed in each type of games and each treatment.

Figure 1.8 presents a coefficient plot summarizing the OLS regression results of revealed rationality levels. The analysis demonstrates a positive association between a subject's performance in the CRT and her revealed rationality depth across all types of games and treatments. Overall, the CRT score has a stronger predictive power on subjects' rationality levels in the guessing games and in the Robot Treatment. In the Robot Treatment, each additional correct answer on the CRT is associated with an average increase of 0.298 ($p < 0.001$) in the individual's revealed rationality level for ring games and 0.566 ($p < 0.001$) for guessing games. Comparatively, in the History Treatment, each additional correct answer on the CRT is associated with a relatively smaller average increase of 0.239 ($p = 0.002$) in the individual's revealed

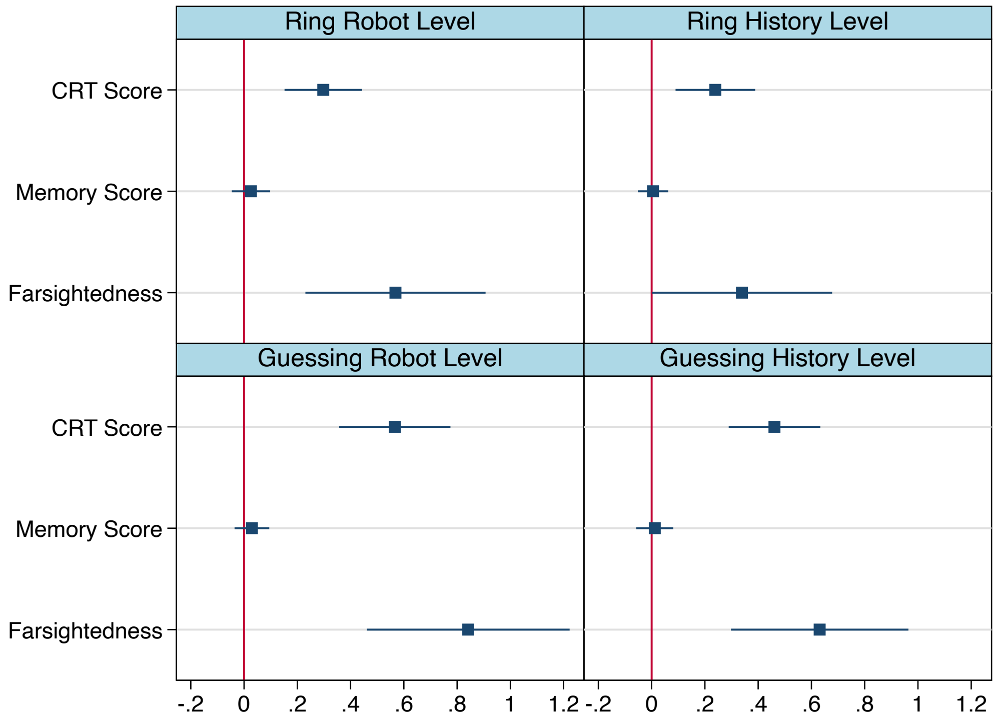


Figure 1.8: Coefficient Plot for OLS Regressions with Different Dependent Variables

Note: Error bars show 95% CI. Standard errors are clustered at the session level.

rationality level for ring games and 0.461 ($p < 0.001$) for guessing games, both approximately 0.8 times the coefficient sizes reported under the Robot Treatment.

In contrast to the previous finding, our results show no significant correlation between short-term memory and strategic sophistication. The coefficient estimates of *Memory Score* are all below 0.03, and all the corresponding $p$-values are above 0.3. Notably, these findings are in line with those of Georganas et al. (2015), who also observe that CRT scores hold some predictive power over subjects' strategic thinking types, whereas short-term memory capacity does not.

A subject's choice in the farsightedness task also holds significant predictive power over her revealed rationality depth across all types of games and treatments. Similar to the CRT score, we observe a stronger association between farsightness and individual rationality levels in the guessing games and in the Robot Treatment. In the Robot Treatment, a farsighted subject's revealed rationality level is, on average, 0.569 ($p = 0.002$) and 0.842 ($p < 0.001$) levels higher than that of a myopic subject when playing ring games and guessing games, respectively. Comparatively, in the History Treatment, a farsighted subject's revealed rationality level is, on average, 0.339 ($p = 0.050$) and 0.631 ($p < 0.001$) levels higher than that of a myopic subject when playing ring games and guessing games, respectively. Both of these coefficients are smaller in size compared to the estimates reported for the Robot Treatment. In summary, these results indicate a strong correlation between an important strategic thinking skill in a dynamic game——backward induction ability––and the strategic reasoning ability in a one-shot interaction.

## 1.8 Concluding Remarks

This study delves into the cognitive capacity of individuals in strategic interactions. To examine their ability to engage in multi-step reasoning, we conduct an experiment designed to elicit and identify each subject's "rationality bound," while controlling for a subject's belief about her opponent's rationality depth. Following the revealed rationality approach, we use two classes of dominance solvable games, ring games and guessing games, as the base games in our experiment for identifying a subject's order of rationality. More importantly, to disentangle the confounding impact of beliefs, we introduce equilibrium-type computer players that are programmed to exhibit infinite order of rationality into the experiment. Under the theoretical framework of Georganas et al. (2015), which formalizes the idea of individual capacity of strategic reasoning, we then test (1) whether a subject's order of rationality is (weakly) higher in the Robot Treatment and (2) whether the observed order of rationality in the Robot Treatment exhibits any stable pattern across games.

Overall, our results offer compelling evidence that matching subjects with robot players to elicit and identify individual strategic reasoning ability is an effective approach. First, subjects exhibit a higher order of rationality in the Robot Treatment compared to the History Treatment, supporting the hypothesis that a subject plays at her highest achievable rationality level (i.e., her capacity bound) in the Robot Treatment. Second, the observed absolute and relative order of rationality in the Robot Treatment remains stable across different types of games, a rare finding in previous literature. Additionally, we find a positive association between a subject's rationality level and her CRT score and backward induction ability, while no significant correlation is observed with short-term memory. These findings indicate that strategic reasoning ability may represent an inherent personal characteristic that is distinct from other cognitive abilities and can be reliably inferred from choice data when subjects' beliefs about others are properly controlled.

Considering that the revealed rationality bound identified in the Robot Treatment can serve as a reliable proxy for an individual's strategic thinking ability, we can independently implement dominance-solvable games, such as ring games and guessing games, with human subjects playing against fully rational computer opponents to effectively elicit and identify human players' strategic capacity, either before or after any lab experiment. By matching human players with computer players, their revealed strategic sophistication is not confounded by their endogenous beliefs about each other's level of sophistication. Furthermore, the robot approach eliminates the need for multiple players to identify a single player's $k$th-order rationality in a game, allowing for an individual task that efficiently elicits and identifies a subject's higher-order rationality. Additionally, as the interactions with computer players are independent of the interactions with human players, the two experiences are expected to have minimal influence on each other. Consequently, the measurement of strategic reasoning ability could remain distinct from the behavioral patterns observed in the main experiment session, thereby avoiding any potential contamination between the two.

In conclusion, we believe that such experiment protocol, particularly the robot approach, has the potential to become a standard tool for measuring a player's actual strategic sophistication, analogous to the usage of the established method (for eliciting risk attitude) in Holt and Laury (2002) but applied to the domain of strategic reasoning. By utilizing this tool, we can gain a better understanding of whether non-equilibrium behavior observed in the main experiment can be attributed to bounded strategic thinking capability or other factors, such as non-equilibrium beliefs and social preferences.

## References

Agranov, M., Potamites, E., Schotter, A., & Tergiman, C. (2012). Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior*, *75*(2), 449–463.

Alaoui, L., Janezic, K. A., & Penta, A. (2020). Reasoning about others' reasoning. *Journal of Economic Theory*, *189*, 105091.

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

Arad, A., & Rubinstein, A. (2012). The 11–20 money request game: A level-k reasoning study. *American Economic Review*, *102*(7), 3561–3573.

Bayer, R. C., & Renou, L. (2016). Logical omniscience at the laboratory. *Journal of Behavioral and Experimental Economics*, *64*, 41–49.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, *52*(4), 1007–1028.

Bone, J., Hey, J. D., & Suckling, J. (2009). Do people plan? *Experimental Economics*, *12*, 12–25.

Bosch-Rosa, C., & Meissner, T. (2020). The one player guessing game: A diagnosis on the relationship between equilibrium play, beliefs, and best responses. *Experimental Economics*, *23*, 1129–1147.

Brandenburger, A., Danieli, A., & Friedenberg, A. (2019). Identification of reasoning about rationality. *Mimeo*.

Burchardi, K. B., & Penczynski, S. P. (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior*, *84*, 39–57.

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

Cerigioni, F., Germano, F., Rey-Biel, P., & Zuazo-Garin, P. (2019). Higher orders of rationality and the structure of games. *Mimeo*.

Chen, C.-T., Huang, C.-Y., & Wang, J. T.-y. (2018). A window of cognition: Eyetracking the reasoning process in spatial beauty contest games. *Games and Economic Behavior*, *111*, 143–158.

Costa-Gomes, M., & Crawford, V. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, *96*(5), 1737–1768.

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.

Devetag, G., & Warglien, M. (2003). Games and phone numbers: Do short-term memory bounds affect strategic behavior? *Journal of Economic Psychology*, *24*(2), 189–202.

Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Friedenberg, A., Kets, W., & Kneeland, T. (2018). Is bounded rationality driven by limited ability? *Unpublished paper*.

Georganas, S., Healy, P. J., & Weber, R. A. (2015). On the persistence of strategic sophistication. *Journal of Economic Theory*, *159*, 369–400.

Grehl, S., & Tutić, A. (2015). Experimental evidence on iterated reasoning in games. *PloS one*, *10*(8), e0136524.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association*, *1*(1), 114–125.

Grosskopf, B., & Nagel, R. (2008). The two-person beauty contest. *Games and Economic Behavior*, *62*(1), 93–99.

Hanaki, N., Jacquemet, N., Luchini, S., & Zylbersztejn, A. (2016). Cognitive ability and the effect of strategic uncertainty. *Theory and Decision*, *81*(1), 101–121.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, *92*(4), 1062–1069.

Jin, Y. (2021). Does level-k behavior imply level-k thinking? *Experimental Economics*, *24*, 330–353.

Johnson, E. J., Camerer, C., Sen, S., & Rymon, T. (2002). Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, *104*(1), 16–47.

Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, *83*(5), 2065–2079.

Lim, W., & Xiong, S. (2016). On identifying higher-order rationality. *Mimeo*.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, 102426.

Meijering, B., Van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PloS one*, *7*(9), e45961.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, *85*(5), 1313–1326.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, *52*(4), 1029–1050.

Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, *25*(3), 309–327.

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.

Van den Bos, W., Li, J., Lau, T., Maskin, E., Cohen, J. D., Montague, P. R., & McClure, S. M. (2008). The value of victory: Social origins of the winner's curse in common value auctions. *Judgment and Decision Making*, *3*(7), 483.

Wechsler, D. (1939). *The measurement of adult intelligence*. Williams & Wilkins.

*Chapter 2*

# CURSED SEQUENTIAL EQUILIBRIUM

## 2.1 Introduction

Cursed equilibrium (CE) proposed by Eyster and Rabin (2005) is a leading behavioral equilibrium concept that was developed to explain the "winner's curse" and related anomalies in applied game theory. The basic idea behind CE is that individuals do not fully take account of the dependence of other players' strategic actions on private information. Cursed behavior of this sort has been detected in a variety of contexts. Capen et al. (1971) first noted that in oil-lease auctions, "the winner tends to be the bidder who most overestimates the reserves potential" (Capen et al., 1971, p. 641). Since then, this observation of overbidding relative to the Bayesian equilibrium benchmark, which can result in large losses for the winning bidder, has been widely documented in laboratory auction experiments (Bazerman & Samuelson, 1983; Camerer et al., 2016; Dyer et al., 1989; Forsythe et al., 1989; Ivanov et al., 2010; Kagel & Levin, 1986, 2009; Kagel et al., 1989; Lind & Plott, 1991). In addition, the neglect of the connection between the opponents' actions and private information is also found in non-auction environments, such as bilateral bargaining games (J. Carrillo & Palfrey, 2009; J. D. Carrillo & Palfrey, 2011; Holt & Sherman, 1994; Samuelson & Bazerman, 1985), zero-sum betting games with asymmetric information (Rogers et al., 2009; Søvik, 2009), and voting and jury decisions (Guarnaschelli et al., 2000).

While CE provides a tractable alternative to Bayesian Nash equilibrium and can explain some anomalous behavior in games with a winner's-curse structure, a significant limitation is that it is only developed as a strategic form concept for simultaneous-move Bayesian games. Thus, when applying the standard CE to dynamic games, the CE analysis is carried out on the strategic form representation of the game, implying that CE cannot distinguish behavior across dynamic games that differ in their timing of moves but have the same strategic form. That is, players are assumed to choose type-dependent contingent strategies simultaneously and *not* update their beliefs as the history of play unfolds. A further limitation implied by the strategic form approach is that CE and standard Bayesian Nash equilibrium make identical predictions in games with a private-values information structure (Eyster and Rabin (2005), Proposition 2). In this paper we extend the CE in a simple and natural way to multi-stage games of incomplete information. We call the new equilibrium concept *Cursed Sequential Equilibrium* (CSE).

In Section 2.2, we present the framework and our extension of cursed equilibrium to dynamic games. We consider the framework of multi-stage games with observed actions, introduced by Fudenberg and Tirole, 1991b, where players' private information is represented by types, with the assumption that the set of available actions is independent of their types at each public history. Our new solution concept is in the same spirit of the cursed equilibrium—in our model, at each stage, players will (partially) neglect the dependence of the other players' behavioral strategies on their types, by placing some weight on the incorrect belief that all types adopt the average *behavioral* strategy. Specifically, at each public history, this corresponds to the average distribution of actions *given the current belief about others' types at that stage*. Therefore, as players update their beliefs about others' private information via Bayes' rule, but with incorrect beliefs about the other players' behavioral strategies, in later stages this can lead them to have *incorrect* beliefs about the other players' average distribution of actions.

Following Eyster and Rabin (2005)'s notion of cursedness, we parameterize the model by a single parameter $\chi \in [0, 1]$ which captures the degree of cursedness and define fully cursed ($\chi = 1$) CSE analogously to fully cursed ($\chi = 1$) CE. Recall that in a fully cursed ($\chi = 1$) CE, each type of each player chooses a best reply to expected (cursed) equilibrium distribution of other players' actions, averaged over the type-conditional strategies of the other players, with this average distribution calculated using the prior belief on types. Loosely speaking, a player best responds to the average CE strategy of the others. In a $\chi$-CE, players are only partially cursed, in the sense that each player best responds to a $\chi$-weighted linear combination of the *average* $\chi$-CE strategy of the others and the *true* (type-dependent) $\chi$-CE strategy of the others.

The extension of this definition to multi-stage games with observed actions is different from $\chi$-CE in two essential ways: (1) the game is analyzed with behavioral strategies; and (2) we impose sequential rationality and Bayesian updating. In a fully cursed ($\chi = 1$) CSE, (1) implies at every stage $t$ and each public history at $t$, each type of each player $i$ chooses a best reply to the expected (cursed) equilibrium distribution of other players' stage-$t$ actions, averaged over the type-conditional stage-$t$ behavioral strategies of other players, with this average distribution calculated using

$i$'s current belief about types at stage $t$. That is, player $i$ best responds to the average stage-$t$ CSE strategy of others. Moreover, (2) requires that each player's belief at each public history is derived by Bayes' rule wherever possible, and best replies are with respect to the continuation values computed by using the fully cursed beliefs about the behavioral strategies of the other players in current and future stages.

A $\chi$-CSE, for $\chi < 1$, is then defined in analogously to $\chi$-CE, except for using a $\chi$-weighted linear combination of the *average* $\chi$-CSE *behavioral strategies* of others and the *true* (type-dependent) $\chi$-CSE *behavioral strategies* of others. Thus, similar to the fully cursed CE, in a fully cursed ($\chi = 1$) CSE, each player believes other players' actions at each history are *independent* of their private information. On the other hand, $\chi = 0$ corresponds to the standard sequential equilibrium where players have correct perceptions about other players' behavioral strategies and are able to make correct Bayesian inferences.[1]

After defining the equilibrium concept, in Section 2.3 we explore some general properties of the model. We first prove the existence of a cursed sequential equilibrium in Proposition 1. Intuitively speaking, CSE mirrors the standard sequential equilibrium. The only difference is that players have incorrect beliefs about the other players' behavioral strategies at each stage since they fail to fully account for the correlation between others' actions and types at every history. We prove in Proposition 2 that the set of CSE is upper hemi-continuous with respect to $\chi$. Consequently, every limit point of a sequence of $\chi$-CSE points as $\chi$ converges to 0 is a sequential equilibrium. This result bridges our behavioral solution concept with the standard equilibrium theory. Finally, we also show in Proposition 4 that $\chi$-CSE is equivalent to $\chi$-CE for one-stage games, demonstrating the connection between the two behavioral solutions.

In multi-stage games, cursed beliefs about behavioral strategies will distort the evolution of a player's beliefs about the other players' types. As shown in Proposition 3, a direct consequence of the distortion is that in $\chi$-CSE players tend to update their beliefs about others' types too passively. That is, there is some persistence in beliefs in the sense that at each stage $t$, each $\chi$-cursed player's belief about any type profile is at least $\chi$ times the belief about that type profile at stage $t-1$. Among other things,

[1]For the off-path histories, similar to the idea of Kreps and Wilson (1982), we impose the $\chi$-*consistency* requirement (see Definition 2) so the assessment is approachable by a sequence of totally mixed behavioral strategies. The only difference is that players' beliefs are incorrectly updated by assuming others play the $\chi$-cursed behavioral strategies. Hence, in our approach if $\chi = 0$, a CSE is a sequential equilibrium.

this implies that if the *prior* belief about the types is full support and $\chi > 0$, the full support property will persist at all histories, and players will (possibly incorrectly) believe every profile of others' types is possible at every history.

This dampened updating property plays an important role in our framework. Not only does it contribute to the difference between CSE and the standard CE through the updating process, but it also implies additional restrictions on off-path beliefs. The effect of dampened updating is starkly illustrated in the pooling equilibria of signaling games where every type of sender behaves the same everywhere. In this case, Proposition 5 shows if an assessment associated with a pooling equilibrium is a $\chi$-CSE, then it also a $\chi'$-CSE for all $\chi' \leq \chi$, but it is not necessarily a pooling equilibrium for all $\chi' > \chi$. This contrasts with one of the main results about CE, that if a pooling equilibrium is a $\chi$-CE for some $\chi$, then it is a $\chi'$-CE for all $\chi' \in [0, 1]$ (Eyster and Rabin, 2005, Proposition 3).

This suggests that perhaps the dampened updating property is an equilibrium selection device that eliminates some pooling equilibrium, but actually this is *not* a general property. As we demonstrate later, the $\chi$-CE and $\chi$-CSE sets can be non-overlapping, which we illustrate with a variety of applications. The intuition is that in CSE, players generally do not have correct beliefs about the opponents' average behavioral strategies. The pooling equilibrium is just a special case where players have correct beliefs.

In Section 2.4 we explore the implications of cursed sequential equilibrium with five applications in economics and political science. The first subsection of Section 2.4 analyzes the $\chi$-CSE of signaling games. Besides studying the theoretical properties of pooling $\chi$-CSE, we also analyze two simple signaling games that were studied in a laboratory experiment (Brandts & Holt, 1993). We show how varying the degree of cursedness can change the set of $\chi$-CSE in these two signaling games in ways that are consistent with the reported experimental findings. Next, we turn to the exploration of how sequentially cursed reasoning can influence strategic communication. To this end, we analyze the $\chi$-CSE for a public goods game with communication (Palfrey & Rosenthal, 1991; Palfrey et al., 2017) in the second subsection of Section 2.4, finding that $\chi$-CSE predicts there will be less effective communication when players are more cursed.

Next, in the third subsection of Section 2.4 we apply $\chi$-CSE to the centipede game studied experimentally by McKelvey and Palfrey (1992) where one of the players believes the other player might be an "altruistic" player who always passes. This is a simple reputation-building game, where selfish types can gain by imitating altruistic types in early stages of the game. The public goods application and the centipede game are both private-values environments, so these two applications clearly demonstrate how CSE departs from CE and the Bayesian Nash equilibrium, and shows the interplay between sequentially cursed reasoning and the learning of types in private-value models.

In strategic voting applications, conditioning on "pivotality"—the event where your vote determines the final outcome—plays a crucial role in understanding equilibrium voting behavior. To illustrate how cursedness distorts the pivotal reasoning, in the fourth subsection of Section 2.4 we study the three-voter two-stage agenda voting game introduced by Ordeshook and Palfrey (1988). Since this is a private value game, the predictions of the $\chi$-CE and the Bayesian Nash equilibrium coincide for all $\chi$. That is, cursed equilibrium predicts no matter how cursed the voters are, they are able to correctly perform pivotal reasoning. On the contrary, our CSE predicts that cursedness will make the voters less likely to vote strategically. This is consistent with the empirical evidence about the prevalence of sincere voting over sequential agendas when inexperienced voters have incomplete information about other voters' preferences (Levine and Plott, 1977; Plott and Levine, 1978; Eckel and Holt, 1989).

Finally, in the last subsection of Section 2.4 we study the relationship between cursedness and epistemic reasoning by considering the two-person dirty faces game previously studied by Weber (2001) and Bayer and Chan (2007). In this game, $\chi$-CSE predicts cursed players are, to some extent, playing a "coordination" game where they coordinate on a specific learning speed about their face types. Therefore, from the perspective of CSE, the non-equilibrium behavior observed in experiments can be interpreted as possibly due to a coordination failure resulting from cognitive limitations.

The cursed sequential equilibrium extends the concept of cursed equilibrium from static Bayesian games to multi-stage games with observed actions. This generalization preserves the spirit of the original cursed equilibrium in a simple and tractable way, and provides additional insights about the effect of cursedness in dynamic games. A contemporaneous working paper by Cohen and Li, 2023 is closely related to our paper. That paper adopts an approach based on the coarsening of information sets to define sequential cursed equilibrium (SCE) for extensive form games with perfect recall. The SCE model captures a different kind of cursedness[2] that arises if a player neglects the dependence of other players' unobserved (i.e., either future or simultaneous) actions on the history of play in the game, which is different from the dependence of other players' actions on their type (as in CE and CSE). In the terminology of Eyster and Rabin (2005) (p. 1665), the cursedness is with respect to endogenous information, i.e., what players observe about the path of play. The idea is to treat the unobserved actions of other players in response to different histories (endogenous information) similarly to how cursed equilibrium treats players' types. A two-parameter model of partial cursedness is developed, and a series of examples demonstrate that for plausible parameter values, the model is consistent with some experimental findings related to the failure of subjects to fully take account of unobserved hypothetical events, whereas behavior is "more rational" if subjects make decisions after directly observing such events. A more detailed discussion of the differences between CSE and SCE can be found Fong et al. (2023). At a more conceptual level, our paper is related to several other behavioral solution concepts developed for dynamic games, such as agent quantal response equilibrium (AQRE) (McKelvey & Palfrey, 1998), dynamic cognitive hierarchy theory (DCH) (Lin, 2022; Lin & Palfrey, 2022), and the analogy-based expectation equilibrium (ABEE) (Jehiel, 2005; Jehiel & Koessler, 2008), all of which modify the requirements of sequential equilibrium in different ways than cursed sequential equilibrium.

### 2.2 The Model

Since CSE is a solution concept for dynamic games of incomplete information, in this paper we will focus on the framework of multi-stage games with observed actions (Fudenberg & Tirole, 1991b). The first subsection of Section 2.2 defines the structure of multi-stage games with observed actions, followed by the second subsection of Section 2.2, where the $\chi$-cursed sequential equilibrium is formally developed.

#### Multi-Stage Games with Observed Actions

Let $N = \{1, \ldots, n\}$ be a finite set of players. Each player $i \in N$ has a *type* $\theta_i$ drawn from a finite set $\Theta_i$. Let $\theta \in \Theta \equiv \times_{i=1}^{n} \Theta_i$ be the type profile and $\theta_{-i} \in \Theta_{-i} \equiv \times_{j \neq i} \Theta_j$ be the type profile without player $i$. All players share a common (full support) prior distribution $\mathcal{F}(\cdot) : \Theta \to (0, 1)$. Therefore, for every player $i$, the belief of other

[2] We illustrate some implications of these differences in the application to signaling games in Section 2.4.

players' types conditional on his own type is

$$\mathcal{F}(\theta_{-i}|\theta_i) = \frac{\mathcal{F}(\theta_{-i}, \theta_i)}{\sum_{\theta'_{-i} \in \Theta_{-i}} \mathcal{F}(\theta'_{-i}, \theta_i)}.$$

At the beginning of the game, players observe their own types, but not the other players' types. That is, each player's type is his own private information.

The game is played in stages $t = 1, 2, \ldots, T$. In each stage, players simultaneously choose actions, which will be revealed at the end of the stage. The feasible set of actions can vary with histories, so games with alternating moves are also included. Let $\mathcal{H}^{t-1}$ be the set of all possible histories at stage $t$, where $\mathcal{H}^0 = \{h_\emptyset\}$ and $\mathcal{H}^T$ is the set of terminal histories. Let $\mathcal{H} = \cup_{t=0}^{T} \mathcal{H}^t$ be the set of all possible histories of the game, and $\mathcal{H} \backslash \mathcal{H}^T$ be the set of non-terminal histories.

For every player $i$, the available information at stage $t$ is in $\mathcal{H}^{t-1} \times \Theta_i$. Therefore, player $i$'s information sets can be specified as $\mathcal{I}_i \in \mathcal{Q}_i = \{(h, \theta) : h \in \mathcal{H} \backslash \mathcal{H}^T, \theta_i \in \Theta_i\}$. That is, a type $\theta_i$ player $i$'s information set at the public history $h^t$ can be defined as $\bigcup_{\theta_{-i} \in \Theta_{-i}} (h^t, \theta_i, \theta_{-i})$. With a slight abuse of notation, it will be denoted as $(h^t, \theta_i)$. For the sake of simplicity, we assume that, at each history, the feasible set of actions for every player is independent of their type and use $A_i(h^{t-1})$ to denote the feasible set of actions for player $i$ at history $h^{t-1}$. Let $A_i = \times_{h \in \mathcal{H} \backslash \mathcal{H}^T} A_i(h)$ denote player $i$'s feasible actions in all histories of the game and $A = A_1 \times \cdots \times A_n$. In addition, we assume $A_i$ is finite for all $i \in N$ and $|A_i(h)| \geq 1$ for all $i \in N$ and any $h \in \mathcal{H} \backslash \mathcal{H}^T$.

A behavioral strategy for player $i$ is a function $\sigma_i : \mathcal{Q}_i \to \Delta(A_i)$ satisfying $\sigma_i(h^{t-1}, \theta_i) \in \Delta(A_i(h^{t-1}))$. Furthermore, we use $\sigma_i(a_i^t | h^{t-1}, \theta_i)$ to denote the probability player $i$ chooses $a_i^t \in A_i(h^{t-1})$. We use $a^t = (a_1^t, \ldots, a_n^t) \in \times_{i=1}^n A_i(h^{t-1}) \equiv A(h^{t-1})$ to denote the action profile at stage $t$ and $a_{-i}^t$ to denote the action profile at stage $t$ without player $i$. If $a^t$ is the action profile realized at stage $t$, then $h^t = (h^{t-1}, a^t)$. Finally, each player $i$ has a payoff function $u_i : \mathcal{H}^T \times \Theta \to \mathbb{R}$, and we let $u = (u_1, \ldots, u_n)$ be the profile of payoff functions. A multi-stage game with observed actions, $\Gamma$, is defined by the tuple $\Gamma = \langle T, A, N, \mathcal{H}, \Theta, \mathcal{F}, u \rangle$.

**Cursed Sequential Equilibrium**

In a multi-stage game with observed actions, a solution is defined by an "assessment," which consists of a (behavioral) strategy profile $\sigma$, and a belief system $\mu$. Since action profiles will be revealed to all players at the end of each stage, the belief system specifies, for each player, a conditional distribution over the set of type profiles conditional on each history. Consider an assessment $(\mu, \sigma)$. Following the spirit of the cursed equilibrium, for player $i$ at stage $t$, we define the *average behavioral strategy profile of the other players* as:

$$\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i) = \sum_{\theta_{-i}\in\Theta_{-i}} \mu_i(\theta_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i})$$

for any $i \in N$, $\theta_i \in \Theta_i$ and $h^{t-1} \in \mathcal{H}^{t-1}$.

In CSE, players have incorrect perceptions about other players' behavioral strategies. Instead of thinking they are using $\sigma_{-i}$, a $\chi$-cursed[3] type $\theta_i$ player $i$ would believe the other players are using a $\chi$-weighted average of the average behavioral strategy and the true behavioral strategy:[4]

$$\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i}, \theta_i) = \chi\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i) + (1-\chi)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i}).$$

The beliefs of player $i$ about $\theta_{-i}$ are updated in the $\chi$-CSE via Bayes' rule, whenever possible, assuming other players are using the $\chi$-cursed behavioral strategy rather than the true behavioral strategy. We call this updating rule the *$\chi$-cursed Bayes' rule*. Specifically, an assessment satisfies the $\chi$-cursed Bayes' rule if the belief system is derived from the Bayes' rule while perceiving others are using $\sigma^{\chi}_{-i}$ rather than $\sigma_{-i}$.

**Definition 1.** $(\mu, \sigma)$ *satisfies $\chi$-cursed Bayes' rule if the following is applied to update the posterior beliefs whenever* $\sum_{\theta'_{-i}\in\Theta_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i}, \theta_i) > 0$*:*

$$\mu_i(\theta_{-i}|h^t, \theta_i) = \frac{\mu_i(\theta_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i}, \theta_i)}{\sum_{\theta'_{-i}\in\Theta_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i}, \theta_i)}.$$

Let $\Sigma^0$ be the set of totally mixed behavioral strategy profiles, and let $\Psi^{\chi}$ be the set of assessments $(\mu, \sigma)$ such that $\sigma \in \Sigma^0$ and $\mu$ is derived from $\sigma$ using $\chi$-cursed Bayes' rule.[5] Lemma 1 below shows that another interpretation of the $\chi$-cursed Bayes' rule is that players have correct perceptions about $\sigma_{-i}$ but are unable to make perfect Bayesian inference when updating beliefs. From this perspective, player $i$'s cursed belief is simply a linear combination of player $i$'s cursed belief at the beginning of that stage (with $\chi$ weight) and the Bayesian posterior belief (with $1-\chi$ weight). Because $\sigma$ is totally mixed, there are no off-path histories.

[3] We assume throughout the paper that all players are equally cursed, so there is no $i$ subscript on $\chi$. The framework is easily extended to allow for heterogeneous degrees of cursedness.

[4] If $\chi = 0$, players have correct beliefs about other players' behavioral strategies at every stage.

[5] In the following, we will use $\mu^{\chi}(\cdot)$ to denote the belief system derived under $\chi$-cursed Bayes' Rule. Also, note that both $\sigma^{\chi}_{-i}$ and $\mu^{\chi}$ are induced by $\sigma$; that is, $\sigma^{\chi}_{-i}(\cdot) = \sigma^{\chi}_{-i}[\sigma](\cdot)$ and $\mu^{\chi}(\cdot) = \mu^{\chi}[\sigma](\cdot)$. For the ease of exposition, we drop $[\sigma]$ when it does not cause confusion.

**Lemma 1.** *For any* $(\mu, \sigma) \in \Psi^{\chi}$, $i \in N$, $h^t = (h^{t-1}, a^t) \in \mathcal{H}\backslash\mathcal{H}^T$ *and* $\theta \in \Theta$,

$$\mu_i(\theta_{-i}|h^t, \theta_i) = \chi\mu_i(\theta_{-i}|h^{t-1}, \theta_i) + (1-\chi)\left[\frac{\mu_i(\theta_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i})}{\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i})}\right]$$

*Proof.* See Appendix A.1. □

This is analogous to Lemma 1 of Eyster and Rabin (2005). Another insight provided by Lemma 1 is that even if player types are independently drawn, i.e., $\mathcal{F}(\theta) = \Pi_{i=1}^n \mathcal{F}_i(\theta_i)$, players' cursed beliefs about other players' types are generally *not* independent across players. That is, in general, $\mu_i(\theta_{-i}|h^t, \theta_i) \neq \Pi_{j\neq i}\mu_{ij}(\theta_j|h^t, \theta_i)$. The belief system will preserve the independence only when the players are either fully rational ($\chi = 0$) or fully cursed ($\chi = 1$).

Finally, we place a consistency restriction, analogous to consistent assessments in sequential equilibrium, on how $\chi$-cursed beliefs are updated off the equilibrium path, i.e., when

$$\sum_{\theta'_{-i}\in\Theta_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i}, \theta_i) = 0.$$

An assessment satisfies $\chi$-consistency if it is in the closure of $\Psi^{\chi}$.

**Definition 2.** $(\mu, \sigma)$ *satisfies* $\chi$*-consistency if there is a sequence of assessments* $\{(\mu_k, \sigma_k)\} \subseteq \Psi^{\chi}$ *such that* $\lim_{k\to\infty}(\mu_k, \sigma_k) = (\mu, \sigma)$.

For any $i \in N$, $\chi \in [0, 1]$, $\sigma$, and $\theta \in \Theta$, let $\rho^{\chi}_i(h^T|h^t, \theta, \sigma^{\chi}_{-i}, \sigma_i)$ be $i$'s perceived conditional realization probability of terminal history $h^T \in \mathcal{H}^T$ at history $h^t \in \mathcal{H}\backslash\mathcal{H}^T$ if the type profile is $\theta$ and $i$ uses the behavioral strategy $\sigma_i$ whereas perceives other players' using the cursed behavioral strategy $\sigma^{\chi}_{-i}$. At every non-terminal history $h^t$, a $\chi$-cursed player in $\chi$-CSE will use $\chi$-cursed Bayes' rule (Definition 1) to derive the posterior belief about the other players' types. Accordingly, a type $\theta_i$ player $i$'s conditional expected payoff at history $h^t$ is:

$$\mathbb{E}u_i(\sigma|h^t, \theta_i) = \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu_i(\theta_{-i}|h^t, \theta_i)\rho^{\chi}_i(h^T|h^t, \theta, \sigma^{\chi}_{-i}, \sigma_i)u_i(h^T, \theta_i, \theta_{-i}).$$

**Definition 3.** *An assessment* $(\mu^*, \sigma^*)$ *is a* $\chi$*-cursed sequential equilibrium if it satisfies* $\chi$*-consistency and* $\sigma^*_i(h^t, \theta_i)$ *maximizes* $\mathbb{E}u_i(\sigma^*|h^t, \theta_i)$ *for all* $i$, $\theta_i$, $h^t \in \mathcal{H}\backslash\mathcal{H}^T$.

## 2.3 General Properties of $\chi$-CSE

In this section, we characterize some general theoretical properties of $\chi$-CSE. The first result is the existence of the $\chi$-CSE. The definition of $\chi$-CSE mirrors the definition of the sequential equilibrium by Kreps and Wilson (1982)—the only difference is that players in $\chi$-CSE update their beliefs by $\chi$-cursed Bayes' rule and best respond to $\chi$-cursed (behavioral) strategies. Therefore, one can prove the existence of $\chi$-CSE in a similar way as in the standard argument of the existence of sequential equilibrium.

**Proposition 1.** *For any $\chi \in [0,1]$ and any finite multi-stage game with observed actions, there is at least one $\chi$-CSE.*

*Proof.* We briefly sketch the proof here, and the details can be found in Appendix A.1. Fix any $\chi \in [0,1]$. For any $i \in N$ and any $\mathcal{I}_i = (h^{t-1}, \theta_i)$, player $i$ has to choose every action $a_i^t \in A_i(h^{t-1})$ with probability at least $\epsilon$. Since there are no off-path histories, the belief system is uniquely pinned down by $\chi$-cursed Bayes' rule and a $\chi$-CSE exists in this $\epsilon$-constrained game. We denote this $\chi$-CSE as $(\mu^\epsilon, \sigma^\epsilon)$. By compactness, there is a converging sub-sequence of assessments such that $(\mu^\epsilon, \sigma^\epsilon) \to (\mu^*, \sigma^*)$ as $\epsilon \to 0$, which is a $\chi$-CSE. □

Let $\Phi(\chi)$ be the correspondence that maps $\chi \in [0,1]$ to the set of $\chi$-CSE. Proposition 1 guarantees $\Phi(\chi)$ is non-empty for any $\chi \in [0,1]$. Because $\chi$-cursed Bayes' rule changes continuously in $\chi$, we further prove that $\Phi(\chi)$ is an upper hemi-continuous correspondence.

**Proposition 2.** *$\Phi(\chi)$ is upper hemi-continuous with respect to $\chi$.*

*Proof.* The proof follows a standard argument. See Appendix A.1 for details. □

As shown in Corollary 1, a direct consequence of upper hemi-continuity is that every limit point of a sequence of $\chi$-CSE when $\chi \to 0$ is a sequential equilibrium. This result bridges our behavioral equilibrium concept with standard equilibrium theory.

**Corollary 1.** *Every limit point of a sequence of $\chi$-CSE with $\chi$ converging to 0 is a sequential equilibrium.*

*Proof.* By Proposition 2, we know $\Phi(\chi)$ is upper hemi-continuous at 0. Consider a sequence of $\chi$-CSE. As $\chi \to 0$, the limit point remains a CSE, which is a sequential equilibrium at $\chi = 0$. This completes the proof. □

Finally, by a similar argument to Kreps and Wilson (1982), for any $\chi \in [0, 1]$, $\chi$-CSE is also upper hemi-continuous with respect to payoffs. In other words, our $\chi$-CSE preserves the continuity property of sequential equilibrium.

The next result is the characterization of a necessary condition for $\chi$-CSE. As seen from Lemma 1, players update their beliefs more passively in $\chi$-CSE than in the standard equilibrium—they put $\chi$-weight on their beliefs formed in previous stage. To formalize this, we define the *$\chi$-dampened updating property* in Definition 4. An assessment satisfies this property if at *any* non-terminal history, the belief puts at least $\chi$ weight on the belief in previous stage—both on and off the equilibrium path. In Proposition 3, we show that $\chi$-consistency implies the $\chi$-dampened updating property.

**Definition 4.** *An assessment $(\mu, \sigma)$ satisfies the $\chi$-dampened updating property if for any $i \in N$, $\theta \in \Theta$ and $h^t = (h^{t-1}, a^t) \in \mathcal{H}\backslash\mathcal{H}^T$,*

$$\mu_i(\theta_{-i}|h^t, \theta_i) \geq \chi\mu_i(\theta_{-i}|h^{t-1}, \theta_i).$$

**Proposition 3.** *$\chi$-consistency implies $\chi$-dampened updating for any $\chi \in [0, 1]$.*

*Proof.* See Appendix A.1. □

It follows that if assessment $(\mu, \sigma)$ satisfies the $\chi$-dampened updating property, then for any player $i$, any history $h^t$ and any type profile $\theta$, player $i$'s belief about $\theta_{-i}$ is bounded by

$$\chi\mu_i(\theta_{-i}|h^{t-1}, \theta_i) \leq \mu_i(\theta_{-i}|h^t, \theta_i) \leq 1 - \chi \sum_{\theta'_{-i} \neq \theta_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i).$$

One can see from this condition that when $\chi$ increases, the feasible range of $\mu_i(\theta_{-i}|h^t, \theta_i)$ shrinks, and the restriction on the belief system becomes more stringent. Moreover, if the history $h^t$ is an off-path history of $(\mu, \sigma)$, then this condition characterizes the feasible set of off-path beliefs, which shrinks as $\chi$ increases.

An important implication of this observation is that $\Phi(\chi)$ is not lower hemi-continuous with respect to $\chi$. The intuition is that for some $\chi$-CSE that contains off-path histories, the off-path beliefs to support the equilibrium might not be $\chi$-consistent for sufficiently large $\chi$. In this case, the $\chi$-CSE is not attainable by a sequence of $\chi_k$-CSE where $\chi_k$ converges to $\chi$ from above, causing the lack of lower hemi-continuity.[6]

[6] An example is provided in Section 2.4 (see Footnote 7).

Lastly, another implication of $\chi$-dampened updating property is that for each player $i$, history $h^t$ and type profile $\theta$, the belief $\mu_i(\theta_{-i}|h^t, \theta_i)$ has a lower bound that is *independent* of the strategy profile. The lower bound is characterized in Corollary 2. This result implies that when $\chi > 0$, $\mathcal{F}(\theta_{-i}|\theta_i) > 0$ implies $\mu_i(\theta_{-i}|h^t, \theta_i) > 0$ for all $h^t$, so that if prior beliefs are bounded away from zero, beliefs are always bounded away from 0 as well. In other words, when $\chi > 0$, because of the $\chi$-dampened updating, beliefs will always have full support even if at off-path histories.

**Corollary 2.** *For any $\chi$-consistent assessment $(\mu, \sigma)$, $i \in N$, $\theta \in \Theta$ and $h^t \in \mathcal{H}\backslash\mathcal{H}^T$,*

$$\mu_i(\theta_{-i}|h^t, \theta_i) \geq \chi^t \mathcal{F}(\theta_{-i}|\theta_i)$$

*Proof.* See Appendix A.1. □

If the game has only one stage, then the dampened updating property has no effect, in which case $\chi$-CSE and $\chi$-CE are equivalent solution concepts. This is formally stated and proved in Proposition 4.

**Proposition 4.** *For any one-stage game and for any $\chi$, $\chi$-CSE and $\chi$-CE are equivalent.*

*Proof.* For any one-stage game, the only public history is the initial history $h_\emptyset$. Thus, in any $\chi$-CSE, for each player $i \in N$ and type profile $\theta \in \Theta$, player $i$'s belief about other players' types at this history is

$$\mu_i(\theta_{-i}|h_\emptyset, \theta_i) = \mathcal{F}(\theta_{-i}|\theta_i).$$

Since the game has only one stage, the outcome is simply $a^1 = (a_1^1, \ldots, a_n^1)$, the action profile at stage 1. Moreover, given any behavioral strategy profile $\sigma$, player $i$ believes $a^1$ will be the outcome with probability

$$\sigma_i(a_i^1|h_\emptyset, \theta_i) \times \left[\chi \bar{\sigma}_{-i}(a_{-i}^1|h_\emptyset, \theta_i) + (1-\chi)\sigma_{-i}(a_{-i}^1|h_\emptyset, \theta_{-i})\right].$$

Therefore, if $\sigma$ is the behavioral strategy profile of a $\chi$-CSE in an one-stage game, then for each player $i$, type $\theta_i \in \Theta_i$ and each $a_i^1 \in A_i(h_\emptyset)$ such that $\sigma_i(a_i^1|h_\emptyset, \theta_i) > 0$,

$$a_i^1 \in \arg\max_{a_i^{1\prime} \in A_i(h_\emptyset)} \sum_{\theta_{-i} \in \Theta_{-i}} \mathcal{F}(\theta_{-i}|\theta_i) \times \left\{ \sum_{a_{-i}^1 \in A_{-i}(h_\emptyset)} \left[\chi \bar{\sigma}_{-i}(a_{-i}^1|h_\emptyset, \theta_i) + (1-\chi)\sigma_{-i}(a_{-i}^1|h_\emptyset, \theta_{-i})\right] \right\} u_i(a_i^{1\prime}, a_{-i}^1, \theta_i, \theta_{-i}),$$

which coincides with the maximization problem of $\chi$-CE. This completes the proof. □

From the proof of Proposition 4, one can see that in one-stage games players have *correct* perceptions about the average strategy of others. Therefore, the maximization problem of $\chi$-CSE coincides with the problem of $\chi$-CE. For general multi-stage games, because of the $\chi$-dampened updating property, players will update beliefs incorrectly and thus their perceptions about other players' future moves can also be distorted.

### 2.4 Applications

In this section, we will explore $\chi$-CSE in five applications of multi-stage games with observed actions, in order to illustrate the range of effects it can have and to show how it is different from the $\chi$-CE and sequential equilibrium. The omitted proofs in this section can be found in Appendix A.2.

Our first application is the sender-receiver signaling game, which is practically the simplest possible multi-stage game. From our analysis, we will see both the theoretical and empirical implications of our $\chi$-CSE.

**Pooling Equilibria in Signaling Games**

We first make a general observation about pooling equilibria in multi-stage games. Player $j$ follows a *pooling strategy* if for every non-terminal history, $h^t$, all types of player $j$ take the same action $a_j^{t+1} \in A_j(h^t)$. Conceptually, since every type of player $j$ takes the same action, players other than $j$ cannot make any inference about $j$'s type from $j$'s actions. A *pooling $\chi$-CSE* is a $\chi$-CSE where every player follows a pooling strategy. Hence, every player has correct beliefs about any other player's future move because every type of every player chooses the same action.

Since in any pooling $\chi$-CSE, players can correctly anticipate other players' future moves no matter how cursed they are, one may naturally conjecture that a pooling $\chi$-CSE is also a $\chi'$-CSE for any $\chi' \in [0, 1]$. As shown by Eyster and Rabin (2005), this is true for one-stage Bayesian games: if a pooling strategy profile is a $\chi$-cursed equilibrium, then it is also a $\chi'$-cursed equilibrium for any $\chi' \in [0, 1]$. Surprisingly, this result does not extend to multi-stage games. Proposition 5 shows if a pooling behavioral strategy profile is a $\chi$-CSE, then it remains a $\chi'$-CSE only for $\chi' \leq \chi$, which is a weaker result than Eyster and Rabin (2005).

This result is driven by the $\chi$-dampened updating property which restricts the set of off-path beliefs. As discussed above, when $\chi$ gets larger, the set of feasible off-path beliefs shrinks, eliminating some pooling $\chi$-CSE.

**Proposition 5.** *A pooling $\chi$-CSE is a $\chi'$-CSE for $\chi' \leq \chi$.*

*Proof.* See Appendix A.1. □

The proof strategy is similar to the one in Eyster and Rabin (2005) Proposition 3. Given a $\chi$-CSE behavioral strategy profile, we can separate the histories into on-path and off-path histories. For on-path histories in a pooling equilibrium, since all types of players make the same decisions, players cannot make any inference about other players' types. Therefore, for on-path histories, their beliefs are the prior beliefs, which are independent of $\chi$. On the other hand, for off-path histories, as shown in Proposition 3, a necessary condition for $\chi$-CSE is that the belief system has to satisfy the $\chi$-dampened updating property. As $\chi$ gets larger, this requirement becomes more stringent, and hence some pooling $\chi$-CSE may break down.

Example 1 is a signaling game where the sender has only two types and two messages, and the receiver has only two actions. This example demonstrates the implication of Proposition 5 and shows the lack of lower hemi-continuity; i.e., it is possible for a pooling behavioral strategy profile to be a $\chi$-CSE, but not a $\chi'$-CSE for $\chi' > \chi$. We will also use this example to illustrate how the notion of cursedness in sequential cursed equilibrium proposed by Cohen and Li, 2023 departs from CSE.

**Example 1.** The sender has two possible types drawn from the set $\Theta = \{\theta_1, \theta_2\}$ with $\Pr(\theta_1) = 1/4$. The receiver does not have any private information. After the sender's type is drawn, the sender observes his type and decides to send a message $m \in \{A, B\}$, or any mixture between the two. After that, the receiver decides between action $a \in \{L, R\}$ or any mixture between the two, and the game ends. The game tree is illustrated in Figure 2.1.

If we solve for the $\chi$-CE of the game (or the sequential equilibria), we find that there are two pooling equilibria for every value of $\chi$. In the first pooling $\chi$-CE, both sender types choose $A$; the receiver chooses $L$ in response to $A$ and $R$ at the off-path history $B$. In the second pooling $\chi$-CE, both sender types pool at $B$ and the receiver chooses $R$ at both histories. By Proposition 3 of Eyster and Rabin (2005), these two equilibria are pooling $\chi$-CE for all $\chi \in [0, 1]$. The intuition is that in

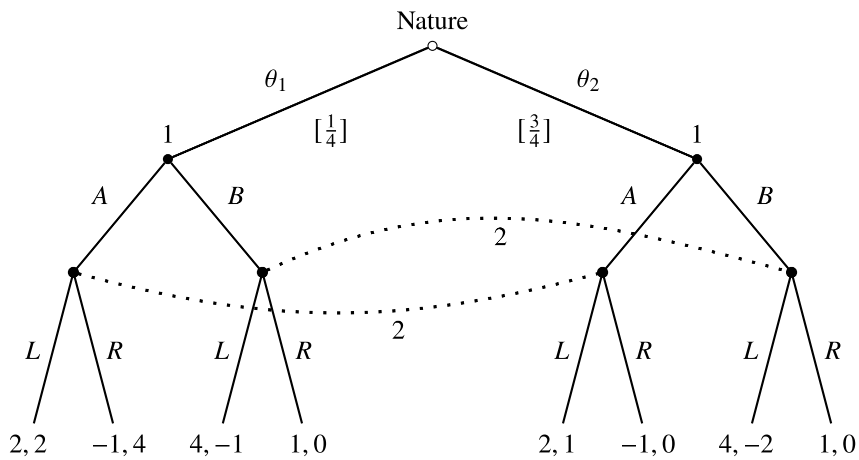


Figure 2.1: Game Tree for Example 1

a pooling $\chi$-CE, players are not able to make any inference about other players' types from their actions because the average normal form strategy is the same as the type-conditional normal form strategy. Therefore, their beliefs are independent of $\chi$, and hence a pooling $\chi$-CE will still be an equilibrium for any $\chi \in [0, 1]$.

However, as summarized in Claim 1 below, the $\chi$-CSE imposes *stronger* restrictions than $\chi$-CE in this example, in the sense that when $\chi$ is sufficiently large, the second pooling equilibrium cannot be supported as a $\chi$-CSE. The key reason is that when the game is analyzed in its normal form, the $\chi$-dampened updating property shown in Proposition 3 does not have any bite, allowing both pooling equilibria to be supported as a $\chi$-CE for any value of $\chi$. Yet, in the $\chi$-CSE analysis, the additional restriction of $\chi$-dampened updating property eliminates some extreme off-path beliefs, and hence, eliminates the second pooling $\chi$-CSE equilibrium for sufficiently large $\chi$. For simplicity, we use a four-tuple $[(m(\theta_1), m(\theta_2)); (a(A), a(B))]$ to denote a behavioral strategy profile.

**Claim 1.** *In this example, there are two pure pooling $\chi$-CSE, which are:*

1. $[(A, A); (L, R)]$ *is a pooling $\chi$-CSE for any $\chi \in [0, 1]$.*
2. $[(B, B); (R, R)]$ *with $\mu_2(\theta_1|A) \in \left[\frac{1}{3}, 1 - \frac{3}{4}\chi\right]$ is a pooling $\chi$-CSE if and only if $\chi \leq 8/9$.*

From previous discussion, we know in general, the sets of $\chi$-CSE and $\chi$-CE are non-overlapping because of the nature of sequential distortion of beliefs in $\chi$-CSE. Yet, a pooling $\chi$-CSE is an exception. In a pooling $\chi$-CSE, players can correctly anticipate others' future moves, so a pooling $\chi$-CSE will mechanically be a pooling $\chi$-CE. In cases such as this, we can find that $\chi$-CSE is a *refinement* of $\chi$-CE.[7]

**Remark.** This game is useful for illustrating some of the differences between the notions of "cursedness" in $\chi$-CSE and the sequential cursed equilibrium ($(\chi_S, \psi_S)$-SCE) proposed by Cohen and Li, 2023.[8] The first distinction is that the $\chi$ and $\chi_S$ parameters capture substantively different sources of distortion in a player's beliefs about the other players' strategies. In $\chi$-CSE, the degree of cursedness, $\chi$, captures how much a player neglects the dependence of the other players' behavioral strategies on those players' (exogenous) *private information*, i.e, types, drawn by nature, and as a result, mistakenly treats different types as behaving the same with probability $\chi$. In contrast, in $(\chi_S, \psi_S)$-SCE, the cursedness parameter, $\chi_S$, captures how much a player neglects the dependence of the other players' strategies on future moves of the others, or current moves that are unobserved because of simultaneous play. Thus, it is a neglect related to endogenous information. If player $i$ *observes* a previous move by some other player $j$, then player $i$ correctly accounts for the dependence of player $j$'s chosen action on player $j$'s private type, as would be the case in $\chi$-CSE only at the boundary where $\chi = 0$.

In the context of pooling equilibria in sender-receiver signaling games, if $\chi_S = 1$, then in SCE the sender believes the receiver will respond the same way both on and off the equilibrium path. This distorts how the sender perceives the receiver's future action in response to an off-equilibrium path message. In $\chi$-CSE, cursedness does not hinder the sender from correctly perceiving the receiver's strategy since the receiver only has one type. Take the strategy profile $[(A, A); (L, R)]$ for example, which is a pooling $\chi$-CSE equilibrium for all $\chi \in [0, 1]$. However, with $(\chi_S, \psi_S)$-SCE, a sender misperceives that the receiver, upon receiving the off-path message $B$, will, with probability $\chi_S$, take the same action ($L$) as when receiving the on-path message $A$. If $\chi_S$ is sufficiently high, the sender will deviate to send $B$, which implies that $[(A, A); (L, R)]$ cannot be supported as an equilibrium when $\chi_S$ is sufficiently large ($\chi_S > 1/3$). The distortion induced by $\chi_S$ also creates an additional SCE if $\chi_S$ is sufficiently large: $[(B, B); (L, R)]$. To see this, if $\chi_S = 1$, then a sender

[7]Note that the $\chi$-CSE correspondence $\Phi(\chi)$ is not lower hemi-continuous with respect to $\chi$. To see this, we consider a sequence of $\{\chi_k\}$ where $\chi_k = \frac{8}{9} + \frac{1}{9k}$ for $k \geq 1$. From the analysis of Claim 1, we know $[(B, B); (R, R)] \notin \Phi(\chi_k)$ for any $k \geq 1$. However, in the limit where $\chi_k \to 8/9$, $[(B, B); (R, R)]$ with $\mu_2(\theta_1|A) = 1/3$ is indeed a CSE. That is, $[(B, B); (R, R)]$ is not approachable by this sequence of $\chi_k$-CSE.

[8]To avoid confusion, we will henceforth add the subscript "S" to the parameters of SCE.

incorrectly believes that the receiver will continue to choose $R$ if the sender deviates to $A$, rather than switching to $L$, and hence $B$ is optimal for both sender types. However, $[(B, B); (L, R)]$ is *not* a $\chi$-CSE equilibrium for any $\chi \in [0, 1]$, or a $\chi$-CE in the sense of Eyster and Rabin (2005), or a sequential equilibrium.

In the two possible pooling equilibria analyzed in the last paragraph, the second SCE parameter, $\psi_S$, does not have any effect, but the role of $\psi_S$ can be illustrated in the context of the $[(B, B); (R, R)]$ sequential equilibrium. This second SCE parameter, $\psi_S$, is introduced to accommodate a player's possible failure to fully account for the informational content from *observed* events. The larger $(1 - \psi_S)$ is, the greater extent a player neglects the informational content of observed actions. Although the parameter $\psi_S$ has a similar flavor to $1 - \chi$ in $\chi$-CSE, it is different in a number of ways. In particular this parameter only has an effect via its interaction with $\chi_S$ and thus does not independently arise. In the two parameter model, the overall degree of cursedness is captured by the product, $\chi_S(1 - \psi_S)$, and thus any cursedness effect of $\psi_S$ is shut down when $\chi_S = 0$. For instance, under our $\chi$-CSE, the strategy profile $[(B, B); (R, R)]$ can only be supported as an equilibrium when $\chi$ is sufficiently small. However, $[(B, B); (R, R)]$ can be supported as a $(\chi_S, \psi_S)$-SCE even when $(1 - \psi_S) = 1$ as long as $\chi_S$ is sufficiently small. In fact, when $\chi_S = 0$, a $(\chi_S, \psi_S)$-SCE is equivalent to sequential equilibrium regardless of the value of $\psi_S$. See Fong et al. (2023) for a more detailed discussion.

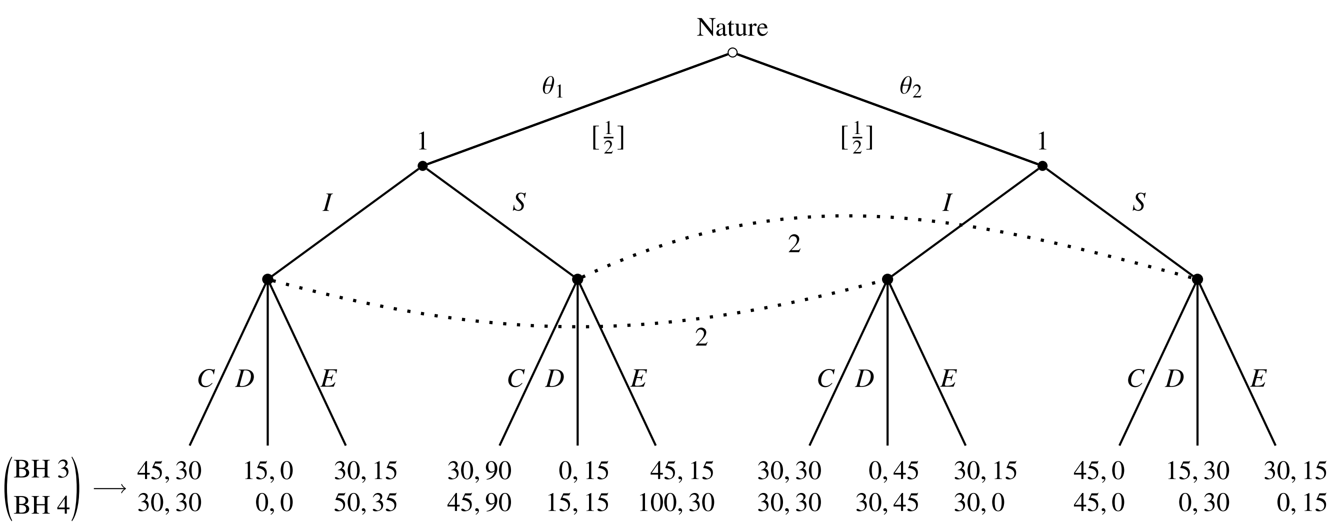


Figure 2.2: Game Tree for *BH 3* and *BH 4* in Brandts and Holt, 1993

**Example 2.** Here we analyze two signaling games that were studied experimentally by Brandts and Holt (1993) (*BH 3* and *BH 4*) and show that $\chi$-CSE can help explain some of their findings. In both Game *BH 3* and *BH 4*, the sender has two possible types $\{\theta_1, \theta_2\}$ which are equally likely. There are two messages $m \in \{I, S\}$ available to the sender.[9] After seeing the message, the receiver chooses an action from $a \in \{C, D, E\}$. The game tree and payoffs for both games are summarized in Figure 2.2.

In both games, there are two pooling sequential equilibria. In the first equilibrium, both sender types send message *I*, and the receiver will choose *C* in response to *I* and *D* in response to *S*. In the second equilibrium, both sender types send message *S*, and the receiver will choose *D* in response to *I* while choose *C* in response to *S*. Both are sequential equilibria, in both games, but only the first equilibrium where the sender sends *I* satisfies the intuitive criterion proposed by Cho and Kreps (1987).

Since the equilibrium structure is similar in both games, the sequential equilibrium and the intuitive criterion predict the behavior should be the same in both games. However, this prediction is strikingly rejected by the data. Brandts and Holt (1993) report that in the later rounds of the experiment, almost all type $\theta_1$ senders send *I* in Game *BH 3* (97 %), and yet all type $\theta_1$ senders send *S* in Game *BH 4* (100%). In contrast, type $\theta_2$ senders behave similarly in both games—46.2% and 44.1% of type $\theta_2$ senders send *I* in Games *BH 3* and *BH 4*, respectively. Qualitatively speaking, the empirical pattern reported by Brandts and Holt (1993) is that *sender type $\theta_1$ is more likely to send I in Game BH 3 than Game BH 4 while sender type $\theta_2$'s behavior is insensitive to the change of games.*

To explain this finding, Brandts and Holt (1993) propose a descriptive story based on *naive receivers*. A naive receiver will think both sender types are equally likely, regardless of which message is observed. This naive reasoning will lead the receiver to choose *C* in both games. Given this naive response, a type $\theta_1$ sender has an incentive to send *I* in Game *BH 3* and choose *S* in Game *BH 4*. (Brandts and Holt, 1993, p. 284 – 285)

In fact, their story of naive reasoning echoes the logic of $\chi$-CSE. When the receiver is fully cursed (or naive), he will ignore the correlation between the sender's action and type, causing him to not update the belief about the sender's type. Proposition 6 characterizes the set of $\chi$-CSE of both games. Following the previous notation, we use a four-tuple $[(m(\theta_1), m(\theta_2)); (a(I), a(S))]$ to denote a behavioral strategy profile.

[9] *I* stands for "**I**ntuitive" and *S* stands for "**S**equential but not intuitive", corresponding to the two pooling sequential equilibria of the two games.

**Proposition 6.** *The set of $\chi$-CSE of Game BH 3 and BH 4 are characterized as below.*

- *In Game BH 3, there are three pure $\chi$-CSE:*
  1. $[(I, I); (C, D)]$ *is a pooling $\chi$-CSE if and only if* $\chi \leq 4/7$.
  2. $[(S, S); (D, C)]$ *is a pooling $\chi$-CSE if and only if* $\chi \leq 2/3$.
  3. $[(I, S); (C, C)]$ *is a separating $\chi$-CSE if and only if* $\chi \geq 4/7$.
- *In Game BH 4, there are three pure $\chi$-CSE:*
  1. $[(I, I); (C, D)]$ *is a pooling $\chi$-CSE if and only if* $\chi \leq 4/7$.
  2. $[(S, S); (D, C)]$ *is a pooling $\chi$-CSE if and only if* $\chi \leq 2/3$.
  3. $[(S, S); (C, C)]$ *is a pooling $\chi$-CSE for any* $\chi \in [0, 1]$.

As noted earlier for Example 1, by Proposition 3 of Eyster and Rabin (2005), pooling equilibria (1) and (2) in games BH 3 and BH 4 survive as $\chi$-CE for *all* $\chi \in [0, 1]$. Hence, Proposition 6 implies that $\chi$-CSE refines the $\chi$-CE pooling equilibria for larger values of $\chi$. Moreover, $\chi$-CSE actually eliminates *all* pooling equilibria in BH 3 if $\chi > 2/3$. Proposition 6 also suggests that for any $\chi \in [0, 1]$, sender type $\theta_2$ will behave similarly in both games, which is qualitatively consistent with the empirical pattern. In addition, $\chi$-CSE predicts that a highly cursed ($\chi > 2/3$) type $\theta_1$ sender will send different messages in different games—highly cursed type $\theta_1$ senders will send $I$ and $S$ in Games *BH 3* and *BH 4*, respectively. This is consistent with the empirical data.

### A Public Goods Game with Communication

Our second application is a threshold public goods game with private information and pre-play communication, variations of which have been studied in laboratory experiments (Palfrey and Rosenthal, 1991; Palfrey et al., 2017). Here we consider the "unanimity" case where there are $N$ players and the threshold is also $N$.

Each player $i$ has a private cost parameter $c_i$, which is independently drawn from a uniform distribution on $[0, K]$ where $K > 1$. After each player's $c_i$ is drawn, each player observes their own cost, but not the others' costs. Therefore, $c_i$ is player $i$'s private information and corresponds to $\theta_i$ in the general formulation.[10] The game

[10] This application has a continuum of types. The framework of analysis developed for finite types is applied in the obvious way.

consists of two stages. After the profile of cost parameters is drawn, the game will proceed to stage 1 where each player simultaneously broadcasts a public message $m_i \in \{0, 1\}$ without any cost or commitment. After all players observe the message profile from this first stage, the game proceeds to stage 2 which is a unanimity threshold public goods game. Player $i$ has to pay the cost $c_i$ if he contributes, but the public good will be provided only if all players contribute. The public good is worth a unit of payoff for every player. Thus, if the public good is provided, each player's payoff will be $1 - c_i$.

If there is no communication stage, the unique Bayesian Nash equilibrium is that no player contributes, which is also the unique $\chi$-CE for any $\chi \in [0, 1]$. In contrast, with the communication stage, there exists an efficient sequential equilibrium where each player $i$ sends $m_i = 1$ if and only if $c_i \leq 1$ and contributes if and only if all players send 1 in the first stage.[11] Since this is a private value game, the standard cursed equilibrium has no bite, and this efficient sequential equilibrium is also a $\chi$-CE for all values of $\chi$, by Proposition 2 of Eyster and Rabin (2005). In the following, we demonstrate that the prediction of $\chi$-CSE is different from CE (and sequential equilibrium).

To analyze the $\chi$-CSE, consider a set of "cutoff" costs, $\{C_c^\chi, C_0^\chi, C_1^\chi, \ldots, C_N^\chi\}$. In the communication stage, each player communicates the message $m_i = 1$ if and only if $c_i \leq C_c^\chi$. In the second stage, if there are exactly $0 \leq k \leq N$ players sending $m_i = 1$ in the first stage, then such a player would contribute in the second stage if and only if $c_i \leq C_k^\chi$. A $\chi$-CSE is a collection of these cost cutoffs such that the associated strategies are a $\chi$-CSE for the public goods game with communication. The most efficient sequential equilibrium identified above for $\chi = 0$ corresponds to cutoffs with $C_0^0 = C_1^0 = \cdots = C_{N-1}^0 = 0$ and $C_c^0 = C_N^0 = 1$.

There are in fact multiple equilibria in this game with communication. In order to demonstrate how the cursed belief can distort players' behavior, here we will focus on the $\chi$-CSE that is similar to the most efficient sequential equilibrium identified above, where $C_0^\chi = C_1^\chi = \cdots = C_{N-1}^\chi = 0$ and $C_c^\chi = C_N^\chi$. The resulting $\chi$-CSE is given in Proposition 7.

**Proposition 7.** *In the public goods game with communication, there is a $\chi$-CSE where*

[11] One can think of the first stage as a poll, where players are asked the following question: "Are you willing to contribute if everyone else says they are willing to contribute?". The message $m_i = 1$ corresponds to a "yes" answer and the message $m_i = 0$ corresponds to a "no" answer.

1. $C_0^\chi = C_1^\chi = \cdots = C_{N-1}^\chi = 0$*, and*

2. *there is a unique* $C^*(N, K, \chi) \leq 1$ *s.t.* $C_c^\chi = C_N^\chi = C^*(N, K, \chi)$ *that solves:*

$$C^*(N, K, \chi) - \chi \left[ \frac{C^*(N, K, \chi)}{K} \right]^{N-1} = 1 - \chi.$$

To provide some intuition, we sketch the proof by analyzing the two-person game, where the $\chi$-CSE is characterized by four cutoffs $\{C_c^\chi, C_0^\chi, C_1^\chi, C_2^\chi\}$, with $C_0^\chi = C_1^\chi = 0$ and $C_c^\chi = C_2^\chi$. If players use the strategy that they would send message 1 if and only if the cost is less than $C_c^\chi$, then by Lemma 1, at the history where both players send 1, player $i$'s cursed posterior belief density would be

$$\mu_i^\chi(c_{-i}|\{1,1\}) = \begin{cases} \chi \cdot \left(\frac{1}{K}\right) + (1-\chi) \cdot \left(\frac{1}{C_c^\chi}\right) & \text{if } c_{-i} \leq C_c^\chi \\ \chi \cdot \left(\frac{1}{K}\right) & \text{if } c_{-i} > C_c^\chi. \end{cases}$$

Notice that cursedness leads a player to put some probability weight on a type that is not compatible with the history. Namely, for $\chi$-cursed players, when seeing another player sending 1, they still believe the other player might have $c_{-i} > C_c^\chi$. When $\chi$ converges to 1, the belief simply collapses to the prior belief as fully cursed players never update their beliefs. On the other hand, when $\chi$ converges to 0, the belief converges to $1/C_c^\chi$, which is the correct Bayesian inference.

Given this cursed belief density, the optimal cost cutoff to contribute, $C_2^\chi$, solves

$$C_2^\chi = \int_0^{C_2^\chi} \mu_i^\chi(c_{-i}|\{1,1\}) dc_{-i}.$$

Finally, at the first stage cutoff equilibrium, the $C_c^\chi$ type of player would be indifferent between sending 1 and 0 at the first stage. Therefore, $C_c^\chi$ satisfies

$$0 = \left(\frac{C_c^\chi}{K}\right) \left\{ -C_c^\chi + \int_0^{C_2^\chi} \mu_i^\chi(c_{-i}|\{1,1\}) dc_{-i} \right\}.$$

After substituting $C_c^\chi = C_2^\chi$, we obtain the $\chi$-CSE satisfies $C_c^\chi = C_2^\chi = (K - K\chi)/(K - \chi)$.

From this expression, one can see that the cutoff $C_c^\chi$ (as well as $C_2^\chi$) is decreasing in $\chi$ and $K$. When $\chi \to 0$, $C_c^\chi$ converges to 1, which is the cutoff of the sequential equilibrium. On the other hand, when $\chi \to 1$, $C_c^\chi$ converges to 0, so there is no possibility for communication when players are fully cursed. Similarly, when

$K \to 1$, $C_c^{\chi}$ converges to 1, which is the cutoff of the sequential equilibrium, while $\lim_{K\to\infty} C_c^{\chi} = 1 - \chi$.

These comparative statics results with respect to $\chi$ and $K$ are not just a special property of the $N = 2$ case, but hold for all $N > 1$. Furthermore, there is a similar effect of increasing $N$ that results in a lower cutoff (less effective communication). These properties of $C^*(N, K, \chi)$ are summarized in Corollary 3.

**Corollary 3.** *The efficient $\chi$-CSE predicts for all $N \geq 2$ and $K > 1$:*

1. $C^*(N, K, 0) = 1$ *and* $C^*(N, K, 1) = 0$.

2. $C^*(N, K, \chi)$ *is strictly decreasing in $N$, $K$, and $\chi$ for any* $\chi \in (0, 1)$.

3. *For all* $\chi \in [0, 1]$, $\lim_{N\to\infty} C^*(N, K, \chi) = \lim_{K\to\infty} C^*(N, K, \chi) = 1 - \chi$.

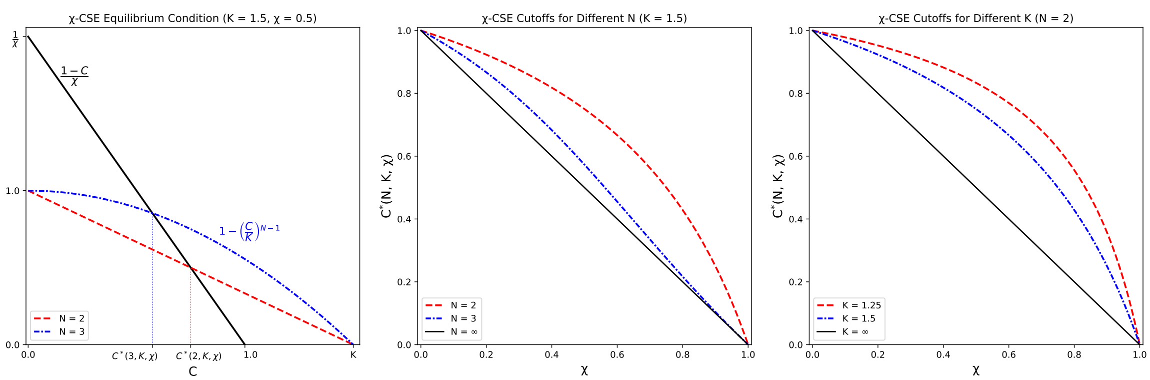


Figure 2.3: (Left) Illustration of the $\chi$-CSE equilibrium condition when $K = 1.5$ and $\chi = 0.5$. (Middle) The $\chi$-CSE cutoff $C^*(N, K, \chi)$ for $N = 2, 3$ and for $N \to \infty$ when $K = 1.5$. (Right) The $\chi$-CSE cutoff $C^*(N, K, \chi)$ for $K = 1.25, 1.5$ and for $K \to \infty$ when $N = 2$.

These properties are illustrated in Figure 2.3. The left panel illustrates the equilibrium condition for $C^*$ in a graph where the horizontal axis is $C \in [0, K]$. We can rewrite the characterization of $C^*(N, K, \chi)$ in Proposition 7 as a solution for $C$ to the following equation:

$$\frac{1 - C}{\chi} = 1 - \left[\frac{C}{K}\right]^{N-1}.$$

The left panel displays the LHS of this equation, $\frac{1-C}{\chi}$, as the downward sloping line that connects the points $(0, \frac{1}{\chi})$ and $(1, 0)$. The RHS is displayed for $N = 2$ and $N = 3$ by the two curves that connect the points $(0, 1)$ and $(K, 0)$. The equilibrium,

$C^*(N, K, \chi)$, is given by the (unique) intersection of the LHS and RHS curves. It is easy to see that $C^*(N, K, \chi)$ is strictly decreasing in $N$, $K$, and $\chi$. When $N$ increases, the RHS increases for all $C \in (0, K)$, resulting in an intersection at a lower value of $C$. When $K$ increases, again the RHS increases for all $C \in (0, K)$, and also the intercept of the RHS on the horizontal axis increases, leading to a similar effect; and when $\chi$ increases, the intercept of the LHS on the horizontal axis decreases, resulting in an intersection at a lower value of $C$. In addition, when $N$ grows without bound, the RHS approaches to 1 for $C < K$, resulting in a limiting intersection at $C^*(\infty, K, \chi) = 1 - \chi$. This is illustrated in the middle panel of Figure 2.3, which graphs $C^*(2, 1.5, \cdot)$, $C^*(3, 1.5, \cdot)$, and $C^*(\infty, 1.5, \cdot)$. A similar effect occurs for $K \to \infty$, illustrated in the right panel of Figure 2.3, which displays $C^*(2, 1.25, \cdot)$, $C^*(2, 1.5, \cdot)$, and $C^*(2, \infty, \cdot)$.

An interesting takeaway of this analysis is that in the public goods game with communication, *cursedness limits information transmission*: $\chi$-CSE predicts when players are more cursed (higher $\chi$), it will be harder for them to effectively communicate in the first stage for efficient coordination in the second stage. Moreover, Corollary 3 shows this $\chi$-CSE varies systematically with *all three parameters of the model*: *$N, K$, and $\chi$*. In contrast, in the standard $\chi$-CE, players best respond to the *average type-contingent strategy* rather than the average behavioral strategy. Since it is a private value game, players do not care about the distribution of types, only the distribution of actions. Thus, the prediction of standard CE coincides with the equilibrium prediction *for all $N, K$, and $\chi$*. This seems behaviorally implausible and is also suggestive of an experimental design that varies the two parameters $N$ and $K$, since the qualitative effects of changing these parameters are identified.

**Reputation Building: The Centipede Game with Altruists**

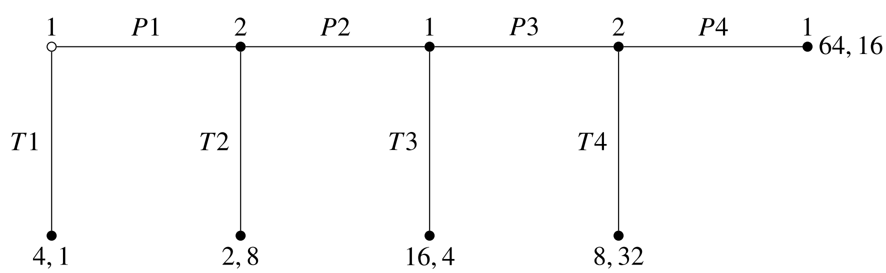


Figure 2.4: Four-stage Centipede Game

In order to further demonstrate the difference between $\chi$-CE and $\chi$-CSE, in this section we consider a variation of the centipede game with private information, as analyzed in McKelvey and Palfrey (1992) and Kreps, 1990. This game is an illustration of reputation-building, where a selfish player imitates an altruistic type in order to develop a reputation for passing, which in turn entices the opponent to pass and leads to higher payoffs.

There are two players and four stages, and the game tree is shown in Figure 2.4. In stage 1, player one can choose either Take ($T1$) or Pass ($P1$). If she chooses $T1$, the game ends and the payoffs to players one and two are 4 and 1, respectively. If she chooses the action $P1$, the game continues and player two has a choice between take ($T2$) and pass ($P2$). If he chooses $T2$, the game ends and the payoffs to player one and two are 2 and 8, respectively. If he chooses $P2$, the game continues to the third stage where player one chooses between $T3$ and $P3$. Similar to the previous stages, if she chooses $T3$, the payoffs to player one and two are 16 and 4, respectively. If she chooses $P3$, the game proceeds to the last stage where player two chooses between $T4$ and $P4$. If player two chooses $T4$ the payoffs are 8 and 32, respectively. If player two alternatively chooses $P4$, the payoffs are 64 and 16, respectively.

Player one has two types, selfish and altruistic. Selfish types are assumed to have a utility function that is linear in their own payoff. Altruistic types instead have a utility function that is linear in the sum of the two payoffs. For the sake of simplicity, we assume that player two has only one type, selfish. The common probability that player one is altruistic is $\alpha$. Player one knows her own type, but player two does not. Thus, player one's type is her private information. In the following, we focus on the interesting case where $\alpha \leq 1/7$.[12]

Because this is a game of incomplete information with *private values,* the standard $\chi$-CE is equivalent to the Bayesian Nash equilibrium of the game for all $\chi \in [0, 1]$, and yields the same take probabilities as the Bayesian equilibrium. Since altruistic player one wants to maximize the sum of the payoffs, it is optimal for her to always pass. The equilibrium behavior is summarized in Claim 2.

**Claim 2.** *In the Bayesian Nash equilibrium, selfish player one will choose $P1$ with probability $\frac{6\alpha}{1-\alpha}$ and choose $T3$ with probability 1; player two will choose $P2$ with probability $\frac{1}{7}$ and choose $T4$ with probability 1.*

It is useful to see exactly why, in this example (and more generally) the standard

[12]If $\alpha > 1/7$, player two always chooses $P2$ in the second stage since the probability of encountering altruistic player one is sufficiently high. Selfish player one would thus chooses $P1$ in the first stage and choose $T3$ in the third stage.

$\chi$-CE is the same as the perfect Bayesian equilibrium. In particular, why it is not the case that cursed beliefs will change player two's updating process after observing $P1$ at stage one. Belief updating is *not* a property of the standard $\chi$-CE as the analysis is in the strategic form, and thus is solved as a BNE of the game in the reduced normal form.[13] Table 2.1 summarizes the payoff matrices in the reduced normal form of centipede game for selfish and altruistic type.

Table 2.1: Reduced Normal Form Centipede Game Payoff Matrix

| selfish $(1-\alpha)$ | $T_2$ | $P_2T_4$ | $P_2P_4$ |
|---|---|---|---|
| $T_1$ | 4, 1 | 4, 1 | 4, 1 |
| $P_1T_3$ | 2, 8 | 16, 4 | 16, 4 |
| $P_1P_3$ | 2, 8 | 8, 32 | 64, 16 |

| altruistic $(\alpha)$ | $T_2$ | $P_2T_4$ | $P_2P_4$ |
|---|---|---|---|
| $T_1$ | 5, 1 | 5, 1 | 5, 1 |
| $P_1T_3$ | 10, 8 | 20, 4 | 20, 4 |
| $P_1P_3$ | 10, 8 | 40, 32 | 80, 16 |

It is easily verified that at the Bayesian Nash equilibrium, selfish player one would choose $T_1$ with probability $(1-7\alpha)/(1-\alpha)$ and choose $P_1T_3$ with probability $6\alpha/(1-\alpha)$, while player two would choose $T_2$ with probability $6/7$.

To solve the standard $\chi$-CE, let selfish player one choose $T_1$ with probability $p$ and $P_1T_3$ with probability $1-p$. Let player two choose $T_2$ with probability $q$ and $P_2T_4$ with probability $1-q$. Notice that for player two, $P_2P_4$ is a dominated strategy and given this, it is also sub-optimal for selfish player one to choose $P_1P_3$. In this case, selfish player one would choose $T_1$ if and only if $4 \geq 2q + 16(1-q) \iff q \geq 6/7$, implying that selfish player one's best response correspondence in the standard cursed analysis coincides with the Bayesian Nash equilibrium analysis. On the other hand, to solve for player two's best responses we need to first solve for the *perceived* strategy. When player two is $\chi$-cursed, he would think that player one is using $\sigma_1^{\chi}(a|\theta)$ where $a \in \{T_1, P_1T_3, P_1P_3\}$ and $\theta \in \{\text{selfish, altruistic}\}$. Player one's true strategy is given in Table 2.2.

Table 2.2: Player one's True Strategy

| | player one's type | |
|---|---|---|
| $\sigma_1(a|\theta)$ | selfish | altruistic |
| $T_1$ | $p$ | 0 |
| $P_1T_3$ | $1-p$ | 0 |
| $P_1P_3$ | 0 | 1 |

In this case, player one's *average strategy* is simply:

$$\bar{\sigma}_1(T_1) = (1-\alpha)p, \quad \bar{\sigma}_1(P_1T_3) = (1-\alpha)(1-p), \quad \bar{\sigma}_1(P_1P_3) = \alpha.$$

[13]The analysis is similar for the unreduced normal form.

By definition, $\sigma_1^\chi(a|\theta) = \chi\bar{\sigma}_1(a) + (1-\chi)\sigma_1(a|s)$ and hence $\sigma_1^\chi(a|\theta)$ is given in Table 2.3.

Table 2.3: Cursed Perception of Player one's Strategy

| | player one's type | |
|---|---|---|
| $\sigma_1^\chi(a\vert\theta)$ | selfish | altruistic |
| $T_1$ | $p(1-\chi\alpha)$ | $p\chi(1-\alpha)$ |
| $P_1T_3$ | $(1-p)(1-\chi\alpha)$ | $(1-p)\chi(1-\alpha)$ |
| $P_1P_3$ | $\chi\alpha$ | $1-\chi+\chi\alpha$ |

From player two's perspective, given any action profile, player two's expected payoff is not affected by whether player one is selfish or altruistic. Hence, player two only cares about the *marginal* distribution of player one's actions. In this case, $\chi$-cursed player two believes player one will choose $a \in \{T_1, P_1T_3, P_1P_3\}$ with probability $\bar{\sigma}_1(a)$. Therefore, it is optimal for player two to choose $T_2$ if and only if

$$\bar{\sigma}_1(T_1) + 8\left[1-\bar{\sigma}_1(T_1)\right] \geq \bar{\sigma}_1(T_1) + 4\bar{\sigma}_1(P_1T_3) + 32\bar{\sigma}_1(P_1P_3) \iff p \leq \frac{1-7\alpha}{1-\alpha},$$

implying player two's best responses in the standard cursed analysis also coincides with the Nash best responses. As a result, one concludes that standard $\chi$-CE would make exactly the same prediction as the Bayesian Nash equilibrium regardless how cursed the players are.

In contrast, the $\chi$-CSE will exhibit distortions to the *conditional beliefs* of player two, given that player one has passed, because player two incorrectly takes into account how player one's choice to pass depended on player one's private information. In particular, it is harder to build a reputation, since a selfish type will have to imitate altruists in such a way that the true posterior on altruistic type conditional on a pass is higher than in the perfect Bayesian equilibrium, because the updating by player two about player one's type is dampened relative to this true posterior due to cursedness. This distorted belief updating will result in *less* passing by player one compared to the Bayesian equilibrium. Formally, the $\chi$-CSE is described in Proposition 8.

**Proposition 8.** *In the $\chi$-CSE, selfish player one will choose $P1$ with probability $q_1^\chi$ and choose $T3$ with probability 1; player two will choose $P2$ with probability $q_2^\chi$ and choose $T4$ with probability 1 where*

$$q_1^\chi = \begin{cases} \left[\frac{7\alpha-7\alpha\chi}{1-7\alpha\chi} - \alpha\right] \Big/ (1-\alpha) & \text{if } \chi \leq \frac{6}{7(1-\alpha)} \\ 0 & \text{if } \chi > \frac{6}{7(1-\alpha)} \end{cases} \quad \text{and} \quad q_2^\chi = \begin{cases} 1/7 & \text{if } \chi \leq \frac{6}{7(1-\alpha)} \\ 0 & \text{if } \chi > \frac{6}{7(1-\alpha)}. \end{cases}$$

In order to see how the cursedness affects the equilibrium behavior, here we focus on the case of $\chi \leq \frac{6}{7(1-\alpha)}$ where selfish player one and player two will both mix at stage one and two. Given selfish player one chooses $P1$ with probability $q_1^\chi$, by Lemma 1, we know when the game reaches stage two, player two's belief about player one being altruistic becomes

$$\mu^\chi = \chi\alpha + (1-\chi)\left[\frac{\alpha}{\alpha + (1-\alpha)q_1^\chi}\right].$$

Here we see that when $\chi$ is larger, player two will update his belief more slowly. Therefore, in order to maintain indifference at the mixed equilibrium, selfish player one has to pass with *lower* probability so that $P1$ is a more informative signal to player two. Thus, to make player two indifferent between $T2$ and $P2$, the following condition must hold at the equilibrium:

$$\mu^\chi = \frac{1}{7} \iff q_1^\chi = \left[\frac{7\alpha - 7\alpha\chi}{1 - 7\alpha\chi} - \alpha\right] \Big/ (1-\alpha).$$

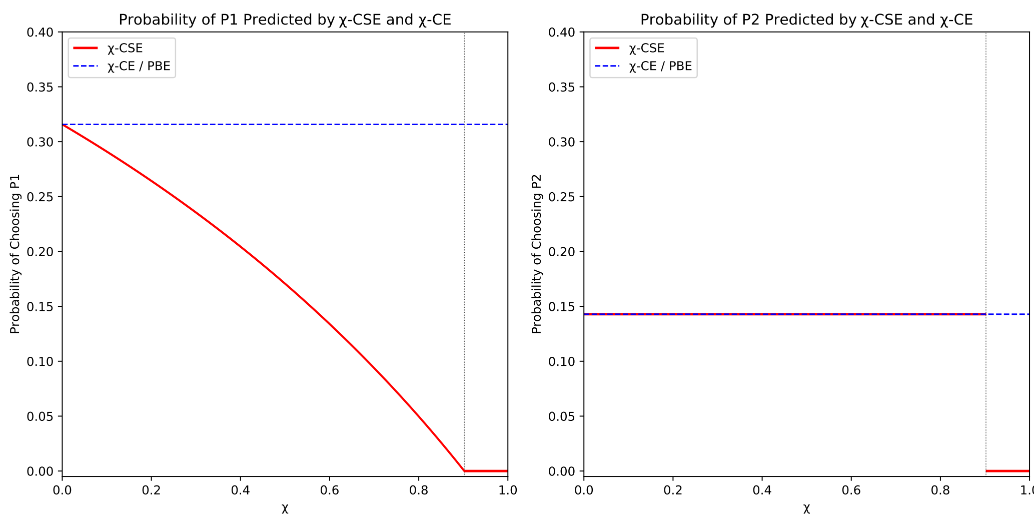


Figure 2.5: $\chi$-CSE of the centipede game with altruistic players ($\alpha = 0.05$)

To conclude this section, in Figure 2.5, we plot the probabilities of choosing $P1$ and $P2$ at $\chi$-CSE when there is a five percent chance that player one is an altruist (i.e., $\alpha = 0.05$). From our analysis above, we can find that both the standard equilibrium theory and $\chi$-CE predict selfish player one chooses $P1$ with probability 0.32 and player two chooses $P2$ with probability 0.14. Moreover, these probabilities are independent of $\chi$. However, $\chi$-CSE predicts when players are more cursed, selfish player one is less likely to choose $P1$. When players are sufficiently cursed

($\chi \geq 0.91$), selfish player one and player two will never pass—i.e., behave as if there were no altruistic players.

**Sequential Voting over Binary Agendas**

In this section, we apply the concept of $\chi$-CSE to the model of strategic binary amendment voting with incomplete information studied by Ordeshook and Palfrey (1988). Let $N = \{1, 2, 3\}$ denote the set of voters. These three voters will vote over three possible alternatives in $X = \{a, b, c\}$. Voting takes place in a two-stage agenda. In the first stage, voters vote between $a$ and $b$. In the second stage, voters vote between $c$ and the majority rule winner of the first stage. The majority rule winner of the second stage is the outcome.

Each voter $i$ has three possible private-value types where $\Theta \in \{\theta_1, \theta_2, \theta_3\}$ is the set of possible types. Each voter's type is independently drawn from a common prior distribution of types, $p$. In other words, the probability of a voter being type $\theta_k$ is $p_k$. Each voter's type is their own private information. Each voter has the same type-dependent payoff function, which is denoted by $u(x|\theta)$ for any $x \in X$ and $\theta \in \Theta$. We summarize the payoff function with the following table.

| | | $x$ | | |
|---|---|---|---|---|
| | $u(x\|\theta)$ | $a$ | $b$ | $c$ |
| | $\theta_1$ | 1 | $v$ | 0 |
| $\theta$ | $\theta_2$ | 0 | 1 | $v$ |
| | $\theta_3$ | $v$ | 0 | 1 |

Notice that $v \in (0, 1)$ is a parameter that measures the *intensity* of the second ranked outcome relative to the top ranked outcome. This intensity parameter, $v$, is assumed to be the same for all types of all voters. Because this is a game of private values, the standard $\chi$-CE and the Bayesian Nash equilibrium coincide.

We use $a_i^1(\theta)$ to denote type $\theta$ voter $i$'s action at stage 1. As is standard in majority voting games we will focus on the analysis of symmetric pure-strategy equilibria where voters do not use weakly dominated strategies. In other words, we will consider $a_i^t(\cdot) = a_j^t(\cdot)$ for all $i, j \in N$, and will drop the subscript.

In this PBE (and $\chi$-CE) all voters will vote sincerely in equilibrium *except for* type $\theta_1$ voters at stage 1. To see this, first note that voting insincerely in the last stage is dominated and thus eliminated, so all types of voters vote for their preferred alternative on the last ballot. Second, voting sincerely in both stages is a dominant strategy for a type $\theta_2$ voter, who prefers any lottery between $b$ and $c$ to either $a$ or $c$. Third, voting sincerely in both stages is also dominant for a type $\theta_3$ voter in the sense that, in the event that neither of the other two voters are type $\theta_3$, then any lottery between $a$ and $c$ is better than a vote between $b$ and $c$ since $b$ (i.e., type $\theta_3$'s least preferred alternative) will win.[14] The PBE (and $\chi$-CE) prediction about a type $\theta_1$ voter's strategy at stage 1 is summarized in the following claim.

**Claim 3.** *The symmetric (undominated pure) PBE strategy for type $\theta_1$ voters in the first stage can be characterized as follows.*

1. $a^1(\theta_1) = b$ *is a PBE strategy if and only if* $v \geq \frac{p_1}{p_1+p_2}$.
2. $a^1(\theta_1) = a$ *is a PBE strategy if and only if* $v \leq \frac{p_1}{p_1+p_3}$.

*Proof.* See Ordeshook and Palfrey (1988). □

Claim 3 shows that, if $v$ is relatively large, only type $\theta_1$ voting sophisticatedly for $b$ instead of sincerely for $a$ can be supported by a PBE. Conditional on being pivotal, voting for $b$ in the first stage guarantees an outcome of $b$ and thus guarantees getting $v$, while voting for $a$ leads to a lottery between $a$ and $c$. Thus, when $v$ is sufficiently high, a type $\theta_1$ voter will have an incentive to strategically vote for $b$ to avoid the risk of having $c$ elected stage 2.

The analysis of a cursed sequential equilibrium is different from the standard cursed equilibrium in strategic form because the cursedness affects belief updating over the stages of the game, and players anticipate future play of the game. Due to the dynamics and the anticipation of future cursed behavior, such cursed behavior at later stages of a game can feedback and affect strategic behavior earlier in the game.

In the context of the two-stage binary amendment strategic voting model, cursed behavior and belief updating mean that voters in the first stage use the expected cursed beliefs in the second stage to compute the continuation values in the two continuation games of the second stage, either a vote between $a$ and $c$ or a vote between $b$ and $c$. Because they have a cursed understanding about the relationship between types and the voting behavior in the first stage, this affects their predictions about which alternative wins in the second stage, conditional on which alternative wins in the first stage.

[14] When there is another type $\theta_3$ voter, the first ballot does not matter since their most preferred alternative $c$ will always win in the second stage.

It is noteworthy that, given any $\chi \in [0,1]$, all voters will still vote sincerely in $\chi$-CSE *except for* type $\theta_1$ voters at stage 1. As implied by Proposition 4, a voter in the last stage would act as if solving a maximization problem of $\chi$-CE but under an (incorrectly) updated belief. Therefore, we can follow the same arguments as solving for the undominated Bayesian equilibrium and conclude that type $\theta_2$ and $\theta_3$ voters as well as type $\theta_1$ voters at stage 2 will vote sincerely under a $\chi$-CSE.

Proposition 9 establishes that the set of parameters $v$ and $p$ that can support a $\chi$-CSE in which type $\theta_1$ voters vote sophisticatedly for $b$ shrinks as $\chi$ increases.

**Proposition 9.** *If $a^1(\theta_1) = b$ can be supported by a symmetric $\chi$-CSE, then it can also be supported by a symmetric $\chi'$-CSE for all $\chi' \leq \chi$.*

The intuition behind strategic voting over agendas mainly comes from the information content of *hypothetical* pivotal events. However, a cursed voter does not (fully) take such information into consideration, and thus becomes overly optimistic about his favorite alternative $a$ being elected in the second stage. Therefore, a type $\theta_1$ voter has a stronger incentive to deviate from sophisticated voting to sincere voting in stage 1 as $\chi$ increases.

Interestingly, the set of $v$ and $p$ that can support a $\chi$-CSE in which type $\theta_1$ voters vote sincerely for $a$ does not necessarily expand as the level of cursedness becomes higher, as characterized in Proposition 10.

**Proposition 10.** *Given $p$ and $v \in (0,1)$, there exists $\tilde{\chi}(p,v)$ such that*

1. *If $v > \frac{p_1}{p_1+p_3}$, then $a^1(\theta_1) = a$ is a $\chi$-CSE strategy if and only if $\chi \geq \tilde{\chi}(p,v)$;*
2. *If $v < \frac{p_1}{p_1+p_3}$, then $a^1(\theta_1) = a$ is a $\chi$-CSE strategy if and only if $\chi \leq \tilde{\chi}(p,v)$.*

Thus, Proposition 10 shows that, when $\chi$ is sufficiently large, there are some values of $(v,p)$ that cannot support sincere voting for type $\theta_1$ voters under PBE (and $\chi$-CE) but can support it under $\chi$-CSE. Alternatively, there also exist some values of $(v,p)$ that can support sincere voting under PBE but fail to support it under $\chi$-CSE when $\chi$ is large.

To illustrate this, Figure 2.6 plots the set of $p$ (fixing $v = 0.7$) that can support a $\chi$-CSE for type $\theta_1$ voters at stage 1 to vote sophisticatedly for $b$ and sincerely for $a$. The left panel of Figure 2.6 shows that a sophisticated voting $\chi$-CSE becomes

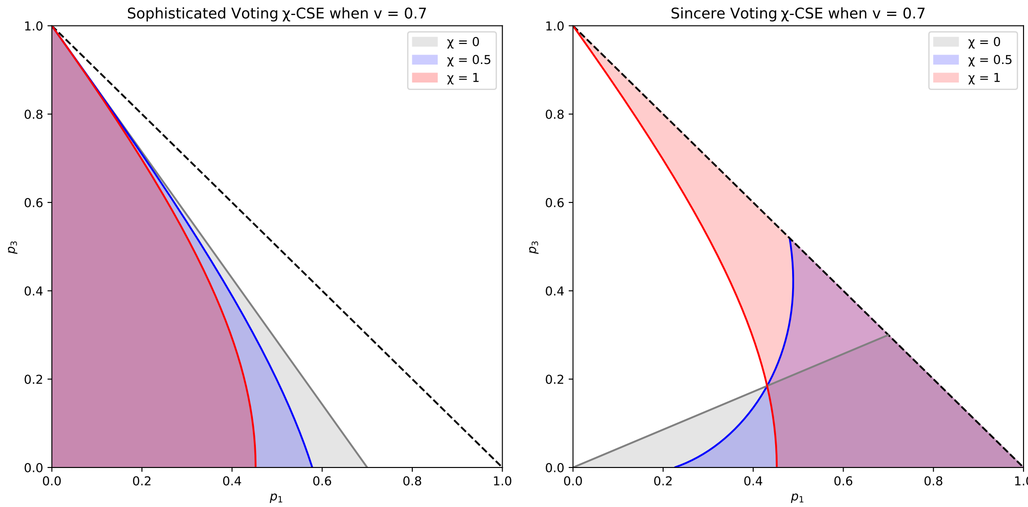


Figure 2.6: $\chi$-CSE for Sophisticated (left) and Sincere (right) Voting When $v = 0.7$

harder to be supported as $\chi$ increases, as indicated by Proposition 9. For example, when $p \equiv (p_1, p_2, p_3) = (0.6, 0.3, 0.1)$, type $\theta_1$ voters will not vote for second preferred alternative $b$ if $\chi > 0.18$.

On the other hand, the right panel of Figure 2.6 shows that, while type $\theta_1$ voters who sincerely vote for $a$ at stage 1 cannot be supported under PBE when $p_3$ is large, they may emerge in a $\chi$-CSE with sufficiently high $\chi$. Also note that when $p_2$ is large, sincere voting by type $\theta_1$ voters is no longer a $\chi$-CSE with high $\chi$. In such a sincere voting equilibrium, a fully rational type $\theta_1$ voter knows there will be only one type $\theta_2$ voter among the other two voters when being pivotal. As a result, whether to sincerely vote for $a$ is determined by the ratio of $p_1$ to $p_3$. When $p_3$ is large, sincere voting at stage 1 will likely lead to zero payoff for type $\theta_1$ voters and thus cannot be a PBE strategy. However, cursed type $\theta_1$ voters will take the possibility of having two type $\theta_2$ voters into account since they are not correctly conditioning on pivotality. As a result, when $p_2$ is large, sincere voting at stage 1 will likely lead to zero payoff for type $\theta_1$ voters, and thus cannot be a $\chi$-CSE strategy with high $\chi$, while voting sophisticatedly for $b$ can likely secure a payoff of $v$.

### The Dirty Faces Game

The dirty faces game was first described by Littlewood (1953) to study the relationship between common knowledge and behavior.[15] There are several different

[15]The dirty faces game has also been reframed as the "cheating wives puzzle" (Gamow & Stern, 1958), the "cheating husbands puzzle" (Moses et al., 1986), the "muddy children puzzle" (Barwise, 1981) and (Halpern & Moses, 1990), and the "red hat puzzle" (Hardin & Taylor, 2008).

variants of this game, but here we focus on a simplified version, the two-person dirty faces game, which was theoretically analyzed by Fudenberg and Tirole, 1991a and Lin (2022) and was experimentally studied by Weber (2001) and Bayer and Chan (2007).

Let $N = \{1, 2\}$ be the set of players. For each $i \in N$, let $x_i \in \{O, X\}$ represent whether player $i$ has a clean face ($O$) or a dirty face ($X$). Each player's face type is independently and identically determined by a commonly known probability $p = \Pr(x_i = X) = 1 - \Pr(x_i = O)$. Once the face types are drawn, each player $i$ can observe the other player's face $x_{-i}$ but not their own face.[16] If there is at least one player with a dirty face, a public announcement of this fact is broadcast to both players at the beginning of the game. Let $\omega \in \{0, 1\}$ denote whether there is an announcement or not. If there is an announcement ($\omega = 1$), all players are informed there is at least one dirty face but not the identities. When $\omega = 0$, it is common knowledge to both players that their faces are clean and the game becomes trivial. Hence, in the following, we will focus only on the interesting case where $\omega = 1$.

There are a finite number of $T \geq 2$ stages. In each stage, each player $i$ simultaneously chooses $s_i \in \{U, D\}$. The game ends as soon as either player (or both) chooses $D$, or at the end of stage $T$ in case neither player has chosen $D$. Actions are revealed at the end of each stage. Payoffs depend on own face types and action. If a player chooses $D$, he will get $\alpha > 0$ if he has a dirty face while receive $-1$ if he has a clean face. We assume that

$$p\alpha - (1-p) < 0 \iff 0 < \bar{\alpha} \equiv \frac{\alpha}{(1-p)(1+\alpha)} < 1, \tag{2.1}$$

where $p\alpha - (1 - p)$ is the expected payoff of $D$ when the belief of having a dirty face is $p$. Thus, Assumption (2.1) guarantees it is strictly dominated to choose $D$ at stage 1 when observing a dirty face. In other words, players will be rewarded when correctly inferring the dirty face but penalized when wrongly claiming the dirty face.

The payoffs are discounted with a common discount factor $\delta \in (0, 1)$. To summarize, conditional on reaching stage $t$, each player's payoff function (which depends on their own face and action) can be written as:

$$u_i(s_i|t, x_i = X) = \begin{cases} \delta^{t-1}\alpha & \text{if } s_i = D \\ 0 & \text{if } s_i = U \end{cases} \quad \text{and} \quad u_i(s_i|t, x_i = O) = \begin{cases} -\delta^{t-1} & \text{if } s_i = D \\ 0 & \text{if } s_i = U. \end{cases}$$

[16] To fit into the framework, each player's "type" (their own private information) can be specified as "other players' faces." That is, $\theta_i = x_{-i}$.

Therefore, a two-person dirty faces game is defined by a tuple $\langle p, T, \alpha, \delta \rangle$.

Since the game ends as soon as some player chooses $D$, the information sets of the game can be specified by the face type the player observes and the stage number. Thus a behavioral strategy can be represented as:

$$\sigma : \{O, X\} \times \{1, \dots, T\} \to [0, 1],$$

which is a mapping from information sets to the probability of choosing $D$, where $\{O, X\}$ corresponds to a player's observation of the other player's face.

There is a unique Nash equilibrium. When observing a clean face, a player would immediately know his face is dirty. Hence, it is strictly dominant to choose $D$ at stage 1 in this case. On the other hand, when observing a dirty face, because of Assumption (2.1), it is optimal for the player to choose $U$ at stage 1. However, if the game proceeds to stage 2, the player would know his face is dirty because the other player would have chosen $D$ at stage 1 if his face were clean and the game would not have reached stage 2. This result is independent of the payoffs, the timing, the discount factor, and the (prior) probability of having a dirty face. The only assumption for this argument is common knowledge of rationality.

Alternatively, when players are "cursed," they are not able to make perfect inferences from the other player's actions. Specifically, since a cursed player has incorrect perceptions about the relationship between the other player's actions and their private information after seeing the other player choose $U$ in stage 1, a cursed player does not believe they have a dirty face for sure. At the extreme when $\chi = 1$, fully cursed players never update their beliefs. In the following, we will compare the predictions of the standard $\chi$-CE and the $\chi$-CSE. A surprising result is that there is always a *unique* $\chi$-CE, but there can be multiple $\chi$-CSE.

For the sake of simplicity, we focus on the characterization of pure strategy equilibrium in the following analysis. Since the game ends when some player chooses $D$, we can equivalently characterize a stopping strategy as a mapping from the observed face type to a stage in $\{1, 2, \dots, T, T + 1\}$ where $T + 1$ corresponds to the strategy of never stopping. Furthermore, both $\chi$-CE and $\chi$-CSE will be symmetric because if players were to stop at different stages, least one of the players would have a profitable deviation. Finally, we use $\hat{\sigma}^{\chi}(x_{-i})$ and $\tilde{\sigma}^{\chi}(x_{-i})$ to denote the equilibrium stopping strategies of $\chi$-CE and $\chi$-CSE, respectively.

We characterize the $\chi$-CE in Proposition 11. Since $\chi$-CE is defined for simultaneous move Bayesian games, to solve for the $\chi$-CE, we need to look at the corresponding normal form where players simultaneously choose $\{1, 2, \ldots, T, T+1\}$ given the observed face type.

**Proposition 11.** *The $\chi$-cursed equilibrium can be characterized as follows.*

1. *If $\chi > \bar{\alpha}$, the only $\chi$-CE is that both players choose:*

$$\hat{\sigma}^{\chi}(O) = 1 \quad \text{and} \quad \hat{\sigma}^{\chi}(X) = T + 1.$$

2. *If $\chi < \bar{\alpha}$, the only $\chi$-CE is that both players choosing*

$$\hat{\sigma}^{\chi}(O) = 1 \quad \text{and} \quad \hat{\sigma}^{\chi}(X) = 2.$$

Proposition 11 shows that $\chi$-CE makes an extreme prediction—when observing a dirty face, players would either choose $D$ at stage 2 (the equilibrium prediction) or never choose $D$. In addition, the prediction of $\chi$-CE is unique for $\chi \neq \bar{\alpha}$. As characterized in Proposition 12, for extreme values of $\chi$, the prediction of $\chi$-CSE coincides with $\chi$-CE. But for intermediate values of $\chi$, there can be multiple $\chi$-CSE.

**Proposition 12.** *The pure strategy $\chi$-CSE can be characterized as follows.*

1. $\tilde{\sigma}^{\chi}(O) = 1$ *for all* $\chi \in [0, 1]$.
2. *Both players choosing $\tilde{\sigma}^{\chi}(X) = T + 1$ is a $\chi$-CSE if and only if $\chi \geq \bar{\alpha}^{\frac{1}{T+1}}$.*
3. *Both players choosing $\tilde{\sigma}^{\chi}(X) = 2$ is a $\chi$-CSE if and only if $\chi \leq \bar{\alpha}$.*
4. *For any $3 \leq t \leq T$, both players choosing $\tilde{\sigma}^{\chi}(X) = t$ is a $\chi$-CSE if and only if*

$$\left(\frac{1 - \kappa(\chi)}{1 - p}\right)^{\frac{1}{t-2}} \leq \chi \leq \bar{\alpha}^{\frac{1}{t-1}} \quad \textit{where}$$

$$\kappa(\chi) \equiv \frac{[(1+\alpha)(1+\delta\chi) - \alpha\delta] - \sqrt{[(1+\alpha)(1+\delta\chi) - \alpha\delta]^2 - 4\delta\chi(1+\alpha)}}{2\delta\chi(1+\alpha)}.$$

**Illustrative Example**

In order to illustrate the sharp contrast between the predictions of $\chi$-CE and $\chi$-CSE, here we consider an illustrative example where $\alpha = 1/4$, $\delta = 4/5$, $p = 2/3$ and the horizon of the game is $T = 5$. As characterized by Proposition 11, $\chi$-CE predicts players will choose $\hat{\sigma}^{\chi}(X) = 2$ if $\chi \leq \bar{\alpha} = 0.6$; otherwise, they will choose

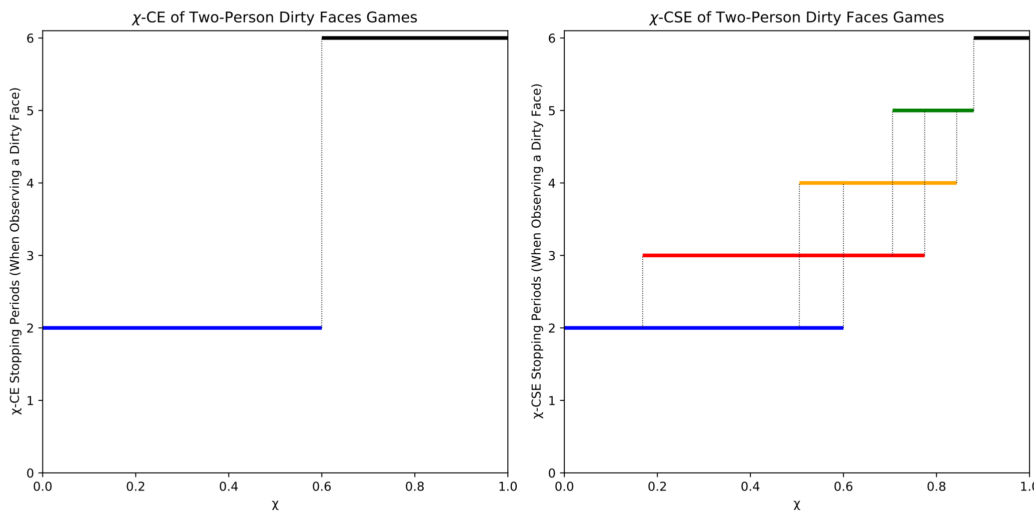


Figure 2.7: $\chi$-CE vs. $\chi$-CSE When $(\alpha, \delta, p, T) = \left(\frac{1}{4}, \frac{4}{5}, \frac{2}{3}, 5\right)$

$\hat{\sigma}^{\chi}(X) = 6$, i.e., they never choose $D$ when observing a dirty face. As demonstrated in the left panel of Figure 2.7, $\chi$-CE is (generically) unique and it predicts players will either behave extremely sophisticated or unresponsive to the other's action at all.

In contrast, as characterized by Proposition 12, there can be multiple $\chi$-CSE. As shown in the right panel of Figure 2.7, when $\chi \leq \bar{\alpha} = 0.6$, both players stopping at stage 2 is still an equilibrium, but it is not unique except for very low values of $\chi$. For $0.168 \leq \chi \leq 0.505$, both players stopping at stage 3 is also a $\chi$-CSE, and for $0.505 \leq \chi \leq 0.6$, there are three pure strategy $\chi$-CSE where both players stop at stage 2, 3, or 4, respectively.

The existence of multiple $\chi$-CSE in which both players stop at $t > 2$ highlights a player's learning process in a multi-stage game, which does not happen in strategic form cursed equilibrium. In the strategic form, a player has no opportunity to learn about the other player's type in middle stages. Thus, when level of cursedness is not low enough to support a $\chi$-CE with stopping at stage 2, both players would never stop. However, in a $\chi$-CSE of the multi-stage game, a cursed player would still learn about his own face being dirty as the game proceeds, even though he might not be confident enough to choose $D$ at stage 2. If $\chi$ is not too large, the expected payoff of choosing $D$ would eventually become positive at some stage before the last stage $T$.[17] For some intermediate values of $\chi$, there might be multiple stopping stages

[17]The upper bound of the inequality in Proposition 12 characterizes the stages at which stopping yields positive expected payoffs.

which yield positive expected payoffs. In this case, the dirty faces game becomes a special type of coordination games where both players coordinate on stopping strategies, resulting in the existence of multiple $\chi$-CSE.[18]

### 2.5 Concluding Remarks

In this paper, we formally developed Cursed Sequential Equilibrium, which extends the strategic form cursed equilibrium (Eyster & Rabin, 2005) to multi-stage games, and illustrated the new equilibrium concept with a series of applications. While the standard CE has no bite in private value games, we show that cursed beliefs can actually have significant consequences for dynamic private value games. In the private value games we consider, our cursed sequential equilibrium predicts (1) under-contribution caused by under-communication in the public goods game with communication, (2) low passing rate in the presence of altruistic players in the centipede game, and (3) less sophisticated voting in the sequential two-stage binary agenda game. We also illustrate the distinction between CE and CSE in some non-private value games. In simple signaling games, $\chi$-CSE implies refinements of pooling equilibria that are not captured by traditional belief-based refinements (or $\chi$-CE), and are qualitatively consistent with some experimental evidence. Lastly, we examine the dirty face game, showing that the CSE further expands the set of equilibrium and predicts stopping in middle stages of the game. Our findings are summarized in Table 2.4.

The applications we consider are only a small sample of the possible dynamic games where CSE could be usefully applied. One prominent class of problems where it would be interesting to study the dynamic effects of cursedness is social learning. For example, in the standard information cascade model of Bikhchandani et al. (1992), we conjecture that the effect would be to delay the formation of an information cascade because players will partially neglect the information content of prior decision makers. Laboratory experiments report evidence that subjects underweight the information contained in prior actions relative to their own signal (Goeree et al., 2007). A related class of problems involves information aggregation through sequential voting and bandwagon effects (Callander, 2007; Ali et al., 2008; Ali and Kartik, 2012). A natural conjecture is that CSE will impede information transmission in committees and juries as later voters will under-appreciate the in-

[18]Note that players with low levels of cursedness would not coordinate on stopping at late stages since the discount factor shrinks the informative value of waiting (i.e., both choosing $U$). This result is characterized by the lower bound of the inequality in Proposition 12.

Table 2.4: Summary of Findings in Section 2.4

| | Private-Value Game | $\chi$-CE vs. BNE | $\chi$-CSE vs. $\chi$-CE |
|---|---|---|---|
| Signaling Games with Pooling Equilibrium | No | $\neq$ | $\chi$-CSE $\subset$ $\chi$-CE |
| Public Goods Game with Communication | Yes | $=$ | $\neq$ |
| Centipede Game with Altruists | Yes | $=$ | $\neq$ |
| Sequential Voting Game | Yes | $=$ | $\neq$ |
| Dirty Faces Game | No | $\neq$ | $\neq$ |

formation content of the decisions by early voters. This would dampen bandwagon effects. The centipede example we studied suggests that CSE might have broader implications for behavior in reputation-building games, such as the finitely repeated prisoner's dilemma or entry deterrence games such as the chain store paradox.

The generalization of CE to dynamic games presented in this paper is limited in several ways. First, the CSE framework is formally developed for finite multi-stage games with observed actions. We do not extend CSE for games with continuous types but we do provide one application that shows how such an extension is possible. However, a complete generalization to continuous types (or continuous actions) would require more technical development and assumptions. We also assume that the number of stages is finite, and extending this to infinite horizon multi-stage games would be a useful exercise. Extending CSE to allow for imperfect monitoring in the form of private histories is another interesting direction to pursue. The SCE approach in Cohen and Li, 2023 allows for cursedness with respect to both public and private endogenous information, which leads to some important differences from our CSE approach. In CSE, we find that subjects are limited in their ability to make correct inferences about hypothetical events, but the mechanism is different from SCE, which introduces a second free parameter that modulates cursedness with respect to hypothetical events. For a more detailed discussion of these and other differences and overlaps between CSE and SCE, see Fong et al. (2023).

As a final remark, our analysis of applications of $\chi$-CSE suggests some interesting experiments. For instance, $\chi$-CSE predicts in the public goods game with communication, when either the number of players ($N$) or the largest possible contribution cost ($K$) increases, pre-play communication will be less effective, while the prediction of sequential equilibrium and $\chi$-CE is independent of $N$ and $K$. In other words, in an experiment where $N$ and $K$ are manipulated, significant treatment effects in this direction would provide evidence supporting $\chi$-CSE over $\chi$-CE. Also, $\chi$-CSE makes qualitatively testable predictions in the sequential voting games and the dirty faces games, which have not been extensively studied in laboratory experiments. In the sequential voting game, it would be interesting to test how sensitive strategic (vs. sincere) voting behavior is to preference intensity ($v$) and the type distribution. In the dirty faces game, it would be interesting to design an experiment to identify the extent to which deviations from sequential equilibrium are related to the coordination problem that arises in $\chi$-CSE.

**References**

Ali, N., Goeree, J. K., Kartik, N., & Palfrey, T. R. (2008). Information aggregation in standing and ad hoc committees. *American Economic Review*, *98*(2), 181–186.

Ali, N., & Kartik, N. (2012). Herding with collective preferences. *Economic Theory*, *51*, 601–626.

Barwise, J. (1981). Scenes and other situations. *Journal of Philosophy*, *78*(7), 369–397.

Bayer, R. C., & Chan, M. (2007). *The dirty faces game revisited* (tech. rep.). University of Adelaide, School of Economics.

Bazerman, M. H., & Samuelson, W. F. (1983). I won the auction but don't want the prize. *Journal of Conflict Resolution*, *27*(4), 618–634.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Brandts, J., & Holt, C. A. (1993). Adjustment patterns and equilibrium selection in experimental signaling games. *International Journal of Game Theory*, *22*(3), 279–302.

Callander, S. (2007). Bandwagons and momentum in sequential voting. *Review of Economic Studies*, *74*, 653–684.

Camerer, C. F., Nunnari, S., & Palfrey, T. R. (2016). Quantal response and non-equilibrium beliefs explain overbidding in maximum value auctions. *Games and Economic Behavior*, *98*, 243–263.

Capen, E. C., Clapp, R. V., & Campbell, W. M. (1971). Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, *23*(06), 641–653.

Carrillo, J., & Palfrey, T. R. (2009). The compromise game: Two-sided adverse selection in the laboratory. *American Economic Journal: Microeconomics*, *1*(1), 151–181.

Carrillo, J. D., & Palfrey, T. R. (2011). No trade. *Games and Economic Behavior*, *71*(1), 66–87.

Cho, I.-K., & Kreps, D. M. (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, *102*(2), 179–221.

Cohen, S., & Li, S. (2023). Sequential cursed equilibrium. *arXiv preprint arXiv:2212.06025v3 First version posted December 13, 2022*.

Dyer, D., Kagel, J. H., & Levin, D. (1989). A comparison of naive and experienced bidders in common value offer auctions: A laboratory analysis. *Economic Journal*, *99*(394), 108–115.

Eckel, C. C., & Holt, C. A. (1989). Strategic voting in agenda controlled committee experiments. *American Economic Review*, *79*(3), 763–773.

Eyster, E., & Rabin, M. (2005). Cursed equilibrium. *Econometrica*, *73*(5), 1623–1672.

Fong, M.-J., Lin, P.-H., & Palfrey, T. R. (2023). A note on cursed sequential equilibrium and sequentia cursed equilibrium. *arXiv preprint arXiv:2304.05515*.

Forsythe, R. E., Isaac, R. M., & Palfrey, T. R. (1989). Theories and tests of blind bidding in sealed-bid auctions. *Rand Journal of Economics*, *20*(2), 214–238.

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

Fudenberg, D., & Tirole, J. (1991b). Perfect bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory*, *53*(2), 236–260.

Gamow, G., & Stern, M. (1958). Forty unfaithful wives. *Puzzle Math*, 20–23.

Goeree, J. K., Palfrey, T. R., Rogers, B. W., & McKelvey, R. D. (2007). Self-correcting information cascades. *Review of Economic Studies*, *74*(3), 733–762.

Guarnaschelli, S., McKelvey, R. D., & Palfrey, T. R. (2000). An experimental study of jury decision rules. *American Political Science Review*, *94*(2), 407–423.

Halpern, J. Y., & Moses, Y. (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM (JACM)*, *37*(3), 549–587.

Hardin, C. S., & Taylor, A. D. (2008). An introduction to infinite hat problems. *The Mathematical Intelligencer*, *30*(4), 20–25.

Holt, C. A., & Sherman, R. (1994). The loser's curse. *American Economic Review*, *84*(3), 642–652.

Ivanov, A., Levin, D., & Niederle, M. (2010). Can relaxation of beliefs rationalize the winner's curse. *Econometrica*, *78*(4), 1435–1452.

Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic Theory*, *123*(2), 81–104.

Jehiel, P., & Koessler, F. (2008). Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, *62*(2), 533–557.

Kagel, J. H., & Levin, D. (1986). The winner's curse and public information in common value auctions. *American Economic Review*, 894–920.

Kagel, J. H., & Levin, D. (2009). Common value auctions and the winner's curse. In *Common value auctions and the winner's curse*. Princeton University Press.

Kagel, J. H., Levin, D., Battalio, R. C., & Meyer, D. J. (1989). First-price common value auctions: Bidder behavior and the "winner's curse". *Economic Inquiry*, *27*(2), 241–258.

Kreps, D. M. (1990). *A course in microeconomic theory*. Princeton university press.

Kreps, D. M., & Wilson, R. (1982). Sequential equilibria. *Econometrica*, *50*(4), 863–894.

Levine, M. E., & Plott, C. R. (1977). Agenda influence and its implications. *Virginia Law Review*, 561–604.

Lin, P.-H. (2022). Cognitive hierarchies in multi-stage games of incomplete information. *arXiv preprint arXiv:2208.11190*.

Lin, P.-H., & Palfrey, T. R. (2022). Cognitive hierarchies in extensive form games. *Caltech Social Science Working Paper*.

Lind, B., & Plott, C. R. (1991). The winner's curse: Experiments with buyers and with sellers. *American Economic Review*, *81*(1), 335–346.

Littlewood, J. E. (1953). *A mathematician's miscellany*. London, England: Meuthen & Co. Ltd.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, 803–836.

McKelvey, R. D., & Palfrey, T. R. (1998). Quantal response equilibria for extensive form games. *Experimental Economics*, *1*(1), 9–41.

Moses, Y., Dolev, D., & Halpern, J. Y. (1986). Cheating husbands and other stories: A case study of knowledge, action, and communication. *Distributed Computing*, *1*(3), 167–176.

Ordeshook, P. C., & Palfrey, T. R. (1988). Agendas, strategic voting, and signaling with incomplete information. *American Journal of Political Science*, 441–466.

Palfrey, T. R., & Rosenthal, H. (1991). Testing for effects of cheap talk in a public goods game with private information. *Games and Economic Behavior*, *3*(2), 183–220.

Palfrey, T. R., Rosenthal, H., & Roy, N. (2017). How cheap talk enhances efficiency in threshold public goods games. *Games and Economic Behavior*, *101*, 234–259.

Plott, C. R., & Levine, M. E. (1978). A model of agenda influence on committee decisions. *American Economic Review*, *68*(1), 146–160.

Rogers, B. W., Palfrey, T. R., & Camerer, C. F. (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory*, *144*(4), 1440–1467.

Samuelson, W. F., & Bazerman, M. H. (1985). Negotiation under the winner's curse. *Research in experimental economics*, *3*, 105–138.

Søvik, Y. (2009). Strength of dominance and depths of reasoning—an experimental study. *Journal of Economic Behavior & Organization*, *70*(1-2), 196–205.

Weber, R. A. (2001). Behavior and learning in the "dirty faces" game. *Experimental Economics*, *4*(3), 229–242.

*Chapter 3*

# CONFORMITY AND CONFIRMATION BIAS

## 3.1 Introduction

Polarization has been a central focus for political scientists in recent years. According to a survey by Pew Research Center in 2014, the distribution of Republicans and Democrats on a 10-scale political values has become much more divided in past decades. From 1994 to 2014, the proportion of self-identified Republicans who were more conservative than the median Democrats increased from 64% to 92%, and the proportion of Democrats who were more liberal than the median Republican grew from 70% to 94%.[1] It is often thought that exposing people to the evidence or views that challenge their beliefs could alleviate political polarization by getting people out of their echo chambers. Nevertheless, some recent studies undermine this intuitive thought. Bail et al. (2018) show that, as its title suggests, people become more polarized when they are exposed to opposing view on Twitter. Nyhan and Reifler (2010) find that corrections in news stories can actually strengthen misperceptions about political issues among the subjects who are most committed to a false belief. In this paper, we develop a model that demonstrates how and when such backfire effect of information may arise in a strategic interaction environment that resembles interactions within a group.

We use a game theoretic framework to model how a decision maker would strategically interpret the evidence at hand. Specifically, we construct an environment with two states, two signals, two policy choices, and two decision makers (or players). The decision makers share a common prior that is in favor of one state. Before individually making a policy choice, each decision maker receives a conditionally independent signal that is correlated with the true state, and a decision maker will get a positive payoff if the policy chosen matches the state (and zero otherwise). In absence of strategic interaction, this setup replicates the standard decision-making model with incomplete information.

Our model departs from the above standard setup under incomplete information by incorporating two additional features. First, a player will experience a utility loss

[1] Pew, Report (2014). "Political Polarization in the American Public." Pew Research Center, https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/, retrieved June 21, 2021.

if her posterior belief about the state is away from her opponent's posterior belief. Second, a player can choose how to read the signal and then form her posterior belief and policy choice accordingly. Specifically, a player can interpret a piece of information as either challenging or confirming her view. These strategic features capture an individual's tendency to exhibit *confirmatory bias*, or to (mis-)interpret new information as in consistent with her current hypotheses about the world, driven by the motivation of a sense of belonging and/or peer pressure in the same ideological group.

We examine the conditions that support the following two types of (Bayesian Nash) equilibria: (i) Bayesian Updating Equilibrium (BUE), in which players always correctly interpret their signals, and (ii) Confirmatory Bias Equilibrium (CBE), in which players always interpret their signals as in favor of their prior belief. In particular, we investigate how the equilibrium conditions respond to changes in the strength of prior belief, the accuracy of a signal, and the payoff from policy choice. Overall, the equilibrium conditions demonstrate a trade-off between a higher expected payoff from policy choice for interpreting a signal correctly, and a lower expected utility loss from players' belief misalignment for always interpreting a signal as supporting the more likely state based on current belief.

Our results show that, in general, an individual is more likely to exhibit confirmatory bias when she currently holds a stronger belief about which state is more possible. A confirmatory bias equilibrium is easier to be sustained when the prior belief favors one state more strongly. In contrast, the range of parameters that supports a Bayesian updating equilibrium shrinks as the prior becomes more extreme. Moreover, we show that a BUE can only be sustained when the prior belief is close enough to 50/50; alternatively, if a signal is only moderately accurate, a CBE can only be sustained when the prior is extreme enough. The reason behind this finding is that, as the players' prior belief becomes stronger, the likelihood of having a different belief from the other and the disutility from it will eventually overwhelm the possible gain from choosing a better policy. Our results thus demonstrate that the motivation of conforming to the belief induced by a more likely signal is a possible explanation for the backfire effect of information observed in previous literature.

Different from the effect of the prior belief on the equilibrium conditions, we find that a signal's accuracy has an ambiguous impact on the occurrence of confirmatory bias. Specifically, if the benefit from policy choice is large enough, lower accuracy of a signal will facilitate the occurrence of confirmatory bias. However, if the gain from implementing the optimal policy is small, a CBE (BUE, resp.) may be easier to be sustained under *greater* (lower, resp.) accuracy of a signal; that is, higher accuracy of a signal may actually facilitate the occurrence of confirmatory bias. This finding, to some extent, violates the intuition that information from a source with high credibility is less susceptible to misinterpretations by readers. The reason behind the above finding is that, although a signal has a higher instrumental value when it is more accurate, the distance between different posterior beliefs induced by opposite signals (and thus the corresponding disutility) increases with a signal's accuracy as well. Our result reveals that, as a signal's accuracy is improved, an increase in the loss from two players' belief misalignment could dominate the increase in the signal's instrumental value related to policy choice if the consequence of policy choice is trivial.

After the derivation of the main results, we consider two possible extensions of the basic model. First, we examine the scenario in which there are more than two players and a player will suffer a utility loss if her posterior belief is away from the median (or the majority) posterior belief among the other players. We show that an increase in the size of players decreases the sustainability of a Bayesian updating equilibrium. Second, we discuss the robustness of our main results to the assumption that a player's belief about the other player's type, or the likelihood that the other player receives a specific signal, do not depend on the player's own signal. In our basic model, we assume that a player's belief about types is solely determined by the common prior. In the extension, we allow a player to update her belief about types based on the signal she receives, and we argue that our main results pertaining to the confirmatory bias equilibrium still hold.

The paper proceeds as follows. We provide a brief review of the literature in the next subsection. Section 3.2 describes the model setup. In Section 3.3, we characterize the conditions that support the BUE and CBE, prove the existence of the equilibria, and obtain the main results. In Section 3.4 we discuss two extensions of the basic model. Section 3.5 concludes. The details of the proofs of propositions are provided in Appendix B.

**Related Literature**

The seminal theoretical work on confirmation bias is Rabin and Schrag (1999), in which they show how an individual's confirmatory bias can lead to overconfidence. Rabin and Schrag (1999) take confirmatory bias as their model's primitive, assuming that a decision maker may misread a signal that contradicts her current belief as confirming evidence with an exogenous and fixed probability. In other words, they do not model the mechanism behind an agent's misinterpretation of signals. Our model complements their work by providing a micro foundation of confirmatory bias and characterizing the relationship between the strength the of prior and the occurrence of confirmatory bias.

A recent work by Fryer Jr et al. (2019) characterizes another possible mechanism behind misperceptions of new information and confirmation bias. In their model, in addition to the two signals that are correlated to the two states of nature, there is another signal that is ambiguous and (thus) open to interpretation. Fryer Jr et al. (2019) assume that an agent is more likely to interpret the ambiguous signal as supporting her current belief and stores such perception in memory, which in turn induces confirmatory bias and polarization in the long run. Our model departs from Fryer Jr et al. (2019) in that, instead of introducing a third type of signal, we follow the setup in Rabin and Schrag (1999) and assume that an agent may misinterpret conflicting evidence as supporting evidence. Moreover, our model is established on a strategic interaction environment instead of an individual decision-making problem. The motivation behind distorting the meaning of a signal in our model is that a decision maker would like to conform to the (posterior) belief that is more likely to be held by another decision maker. Without multiple agents, a decision maker would not exhibit confirmatory bias under our model assumptions.

Our model joins the literature that studies conformism, or "the inclination of an individual to change spontaneously (without any order or request by anyone) his judgements and (or) actions to conform to 'socially prevailing' judgements and (or) actions" (Luzzati 1999, p. 111). The models in this literature can be divided into two classes based on their methodology. A representative model of the first class is Bernheim (1994), in which conformity is derived endogenously. In particular, Bernheim (1994) assumes that individuals would like to be considered to be of a given status (i.e., type), and conformism arises endogenously in the form of pooling equilibria. The second class of models, our model included, treats conformism as primitive in the sense that an agent directly suffers disutility for deviating from exogenous social or group norms on actions (e.g., Akerlof, 1980; Jones, 1984; Luzzati, 1999). For example, Jones (1984) assumes that a worker obtains a utility loss from the distance between the worker's production and the average production level of all workers. In our model, we assume that a player suffers a utility loss if the player's and her opponent's interpretations of their signals diverge. A key feature of our model is that the size of disutility is determined by the distance between the players' posterior beliefs (induced by their interpretations of signals), not by their choices themselves.

Our model is also related to the coordination games with private information, in the sense that the players in our model have an incentive to interpret their signals in the same way. Previous literature in this strand mainly focuses on how the issue of multiple equilibria in a coordination game may be overcome via incomplete information (e.g., Carlsson and Van Damme, 1993) or via communication (e.g., Banks and Calvert, 1992). Our focus is, however, not the properties of a coordination game itself but how this framework can help us explain the emergence of confirmatory bias and the backfire effect of new information.

Our model also contributes to the literature on motivated reasoning (Kunda, 1990; Bénabou and Tirole, 2016), which studies how an agent may distort her belief updating process to achieve a desired conclusion for intrinsic motivation such as self-confidence (e.g., Bénabou and Tirole, 2002; Gottlieb, 2014), moral self-esteem (e.g., Bénabou and Tirole, 2011), anticipatory emotions (e.g., Caplin and Leahy, 2001; Bracha and Brown, 2012), and dissonance reduction (e.g., Chauvin, 2020). In our model, the motivation underlying a decision maker's non-standard Bayesian belief updating is her preference of sharing the same (posterior) belief as her opponent. A main distinction between ours and previous models on motivated reasoning is that most of prior studies focus on individual decision-making (or belief-updating) problem, whereas our model is built within a strategic interaction environment.[2] Another distinctive feature of our model is that our decision maker does not directly suffer a cost of belief distortion that is positively correlated with the distance between the objective Bayesian belief and the decision maker's subjective belief. Instead, the cost comes from a higher chance of choosing an unmatched policy based on the decision maker's subjective belief. This implies that two different distorted beliefs will lead to the same utility cost if the a decision maker chooses the same policy under both beliefs.

Last, our model belongs to the vast literature on biases in belief updating (see Benjamin 2019 for a review). It is noteworthy, however, that we do not incorporate cognitive biases into an agent's belief updating process as a primitive. While a player

[2] One exception we notice is Bénabou (2013), where a group of agents can choose how to interpret *public* signals about future prospects.

in our model can misread the objective signal she receives, she follows the Bayes' rule to form her posterior belief based on her perception, or subjective interpretation, of a signal.

## 3.2 Model

The players in this model, denoted as $i = 1, 2$, are two decision makers from the same ideological group. They need to individually make a policy choice, $a_i \in \{l, r\}$, in a world with an uncertain state. Specifically, the state of the world $\omega$ is either Left ($L$) or Right ($R$), with $\mathbb{P}(\omega = R) = \theta > \frac{1}{2}$. That is, the players share a common prior in favor of $R$. Both players, before making their policy choice, receive a private signal $s_i \in \{\hat{L}, \hat{R}\}$ that is positively correlated with the true state. Specifically, we assume that $\mathbb{P}(s_i = \hat{L}|\omega = L) = \mathbb{P}(s_i = \hat{R}|\omega = R) = q > \theta$. We further assume that the signals are conditionally independent.

A key assumption that distinguishes our model from the above standard setup is that a player can choose how to interpret the evidence at hand. Formally, let $\tilde{s}_i \in \{\hat{L}, \hat{R}\}$ denote Player $i$'s subjective signal, or her interpretation of the (objective) signal. In words, a decision maker can either perceive her objective signal as its literal meaning, or misread it in the opposite direction. The motivation behind possible misinterpretation of a signal is that members from the same group would like to share a similar outlook on the world, which is formalized in the following subsection.

### Utilities and the Maximization Problem

A player's utility in this game consists of two components. First, as in a standard model of individual decision making under incomplete information, we assume that a player gets a positive payoff when her policy choice matches the state and 0 otherwise, i.e., $u(a_i = l|\omega = L) = u(a_i = r|\omega = R) = \delta > 0$ and $u(a_i = l|\omega = R) = u(a_i = r|\omega = L) = 0$. Note that $\delta$ represents the significance of choosing the correct policy. Moreover, a player suffers a utility loss when her posterior belief, induced by her subjective signal $\tilde{s}_i$, is different from the other player's subjective posterior belief. Formally, given $j \in \{1, 2\}$ such that $j \neq i$, let $v_i(p_i[\tilde{s}_i], p_j) = -(p_i - p_j)^2$, where $p_i[\tilde{s}_i] = \mathbb{P}(\omega = R|s_i = \tilde{s}_i)$; that is, $p_i[\tilde{s}_i]$ denotes the (posterior) probability that $\omega = R$ induced by $\tilde{s}_i$. Player $i$ experiences a quadratic utility loss from the distance between the subjective posterior beliefs of the two players.

To capture the tension between the motivation for conformity and the motivation for being accurate behind belief updating, we impose a further assumption on a player's actions. We require a player's policy choice to be consistent with her subjective posterior belief. Formally, we assume that $a_i = \arg\max_{a \in \{l,r\}} [(1 - p_i[\tilde{s}_i])u(a|\omega = L) + p_i[\tilde{s}_i]u(a|\omega = R)]$ (given Player $i$'s subjective signal $\tilde{s}_i$). In words, a player's policy choice maximizes her expected payoff, calculated based on her subjective belief, at the final decision-making stage. Given the symmetric payoff structure, the induced policy $a_i$ would be $r$ when $p_i[\tilde{s}_i] \geq \frac{1}{2}$ or $\tilde{s}_i = \hat{R}$, and $a_i = l$ when $p_i[\tilde{s}_i] < \frac{1}{2}$ or $\tilde{s}_i = \hat{L}$. This assumption reflects the scenario when a decision maker can only remember her (mis)interpretation of the information she receives instead of the content of the information upon the timing of decision making, or when a decision maker incurs a huge mental cost if her choice cannot be rationalized by the belief she forms.

Given the above assumption and Player $j$'s strategy $\left(\tilde{s}_j(s_j = \hat{R}), \tilde{s}_j(s_j = \hat{L})\right)$, we specify Player $i$'s maximization problem as follows:

$$\max_{\tilde{s}_i \in \{\hat{L}, \hat{R}\}} - \mathbb{P}(s_j = \hat{R}|\theta) \left(p_i[\tilde{s}_i] - p_j[\tilde{s}_j(s_j = \hat{R})]\right)^2 - \mathbb{P}(s_j = \hat{L}|\theta) \left(p_i[\tilde{s}_i] - p_j[\tilde{s}_j(s_j = \hat{L})]\right)^2$$
$$+ \mathbb{P}(\omega = R|s_i) \cdot \delta \mathbb{I}_{\{p_i[\tilde{s}_i] \geq \frac{1}{2}\}} + \mathbb{P}(\omega = L|s_i) \cdot \delta \mathbb{I}_{\{p_i[\tilde{s}_i] < \frac{1}{2}\}} \qquad (\star)$$

The first two terms in the objective function describe a player's expected utility (loss) derived from the gap between two players' beliefs. We assume that a player's belief about the signal received by her partner is induced by her prior belief.[3] The idea is that, before interpreting a signal, an individual does not have information other than her current belief to update her evaluation of the likelihood of the other's signal. Thus, the expected utility are calculated using probabilities conditional only on the prior $\theta$ but not the objective signal $s_i$. Given that the prior is in favor of the state $R$, a player believes that her partner is more likely to receive signal $\hat{R}$. As such, if both players do not misread the signal that is aligned with their prior, this assumption would generate a direct incentive to exhibit confirmatory bias for being more likely to share the same belief.

The last two terms describe a player's expected payoff from policy choice, which corresponds to the standard individual optimization problem with incomplete information. This expected payoff is calculated using the objective posterior probabilities, and thus represents the cost of distorting the meaning of a signal. This captures the intuition that, when an individual makes an interpretation of a signal, she may (consciously or subconsciously) realize that misinterpretation can lower her payoff

[3] In this paper, we term a player's opponent *partner* to emphasize that two players come from the same social group.

since she will choose a policy that is less likely to be optimal. As a result, there is a tension between conforming to another group member's belief and interpreting a signal accurately. Note that the expected utility terms from belief distance are dominated when $\delta$ goes to infinity, and the maximization problem is essentially reduced to the standard scenario.

**Bayesian Game**

The above model setup can be summarized as a static Bayesian game (Gibbons, 1992):

- Players: $i \in \{1, 2\}$
- Action space: $\tilde{S}_i = \{\hat{L}, \hat{R}\}$.
- Type space: $S_i = \{\hat{L}, \hat{R}\}$.
- Beliefs (about $j$'s type): $Pr_i(s_j = \hat{R}) = \theta q + (1-\theta)(1-q)$; $Pr_i(s_j = \hat{L}) = \theta(1-q) + (1-\theta)q$.
- Utility:
  - $U_i(\tilde{s}_i = \hat{R}, \tilde{s}_j = \hat{R}; s_i = \hat{L}) = \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\delta$
  - $U_i(\tilde{s}_i = \hat{R}, \tilde{s}_j = \hat{L}; s_i = \hat{L}) = -(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2 + \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\delta$
  - $U_i(\tilde{s}_i = \hat{L}, \tilde{s}_j = \hat{L}; s_i = \hat{L}) = \frac{(1-\theta)q}{\theta(1-q)+(1-\theta)q}\delta$
  - $U_i(\tilde{s}_i = \hat{L}, \tilde{s}_j = \hat{R}; s_i = \hat{L}) = -(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2 + \frac{(1-\theta)q}{\theta(1-q)+(1-\theta)q}\delta$
  - $U_i(\tilde{s}_i = \hat{R}, \tilde{s}_j = \hat{R}; s_i = \hat{R}) = \frac{\theta q}{\theta q+(1-\theta)(1-q)}\delta$
  - $U_i(\tilde{s}_i = \hat{R}, \tilde{s}_j = \hat{L}; s_i = \hat{R}) = -(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2 + \frac{\theta q}{\theta q+(1-\theta)(1-q)}\delta$
  - $U_i(\tilde{s}_i = \hat{L}, \tilde{s}_j = \hat{L}; s_i = \hat{R}) = \frac{(1-\theta)(1-q)}{\theta q+(1-\theta)(1-q)}\delta$
  - $U_i(\tilde{s}_i = \hat{L}, \tilde{s}_j = \hat{R}; s_i = \hat{R}) = -(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2 + \frac{(1-\theta)(1-q)}{\theta q+(1-\theta)(1-q)}\delta$

Note that $\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q} = Pr(\omega = R|\hat{R}) - Pr(\omega = R|\hat{L})$. Thus, $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$ is the utility loss for a player not having the same interpretation of a signal as the other.

### 3.3 Bayesian Updating Equilibrium and Confirmatory Bias Equilibrium

Given the symmetry of the game, we focus on two types of symmetric (pure-strategy) Bayesian Nash Equilibria in the following analyses:

1. Bayesian Updating Equilibrium (BUE): $\tilde{s}_i(\hat{R}) = \tilde{s}_j(\hat{R}) = \hat{R}$ and $\tilde{s}_i(\hat{L}) = \tilde{s}_j(\hat{L}) = \hat{L}$. In this equilibrium, both players always correctly interpret a signal and thus act as Bayesian rational agents.
2. Confirmatory Bias Equilibrium (CBE): $\tilde{s}_i(\hat{R}) = \tilde{s}_j(\hat{R}) = \tilde{s}_i(\hat{L}) = \tilde{s}_j(\hat{L}) = \hat{R}$. In this equilibrium, both players always perceive a signal as supporting their prior belief and thus exhibit confirmatory bias.

**Bayesian Updating Equilibrium (BUE)**

We first characterize the conditions under which a Bayesian Updating Equilibrium can be sustained. Given that Player $j$'s strategy is $\left(\tilde{s}_j(\hat{R}) = \hat{R}, \tilde{s}_j(\hat{L}) = \hat{L}\right)$ (i.e., suppose $j$ acts Bayesian rationally), Player $i$ will correctly interpret the signal when it confirms her current belief (i.e., $\tilde{s}_i(\hat{R}) = \hat{R}$ is Player $i$'s best response) if and only if

$$
\begin{aligned}
&-(\theta(1-q)+(1-\theta)q)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \\
&\quad+\frac{\theta q}{\theta q+(1-\theta)(1-q)}\delta \\
&\geq -(\theta q+(1-\theta)(1-q))\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \\
&\quad+\frac{(1-\theta)(1-q)}{\theta q+(1-\theta)(1-q)}\delta \\
\iff &\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2(2\theta-1)(2q-1) \\
&\geq \frac{1-\theta-q}{\theta q+(1-\theta)(1-q)}\delta
\end{aligned}
$$

Note that under the assumption of $\theta \in (0.5, 1)$ and $q \in (\theta, 1)$, the above inequality always holds.[4] Intuitively, since misinterpreting a confirming signal not only reduces the probability of selecting the optimal policy but also increases the likelihood of misaligned beliefs between the two players, there is no reason for a player to misread signal $\hat{R}$.

[4] $(2\theta-1)(2q-1) > 0 > \frac{1-\theta-q}{\theta q+(1-\theta)(1-q)}$ for every $\theta \in (0.5, 1)$ and $q \in (\theta, 1)$.

Alternatively, when Player $i$'s signal conflicts with her current belief, she will correctly interpret it (i.e., $\tilde{s}_i(\hat{L}) = \hat{L}$ is Player $i$'s best response) if and only if

$$\begin{aligned}
&- (\theta q + (1-\theta)(1-q)) \left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta) q} \right)^2 \\
&\quad + \frac{(1-\theta) q}{\theta(1-q) + (1-\theta) q} \delta \\
\geq &- (\theta(1-q) + (1-\theta) q) \left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta) q} \right)^2 \\
&\quad + \frac{\theta(1-q)}{\theta(1-q) + (1-\theta) q} \delta \\
\iff &\left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta) q} \right)^2 (2\theta - 1)(2q - 1) \\
\leq &\frac{q - \theta}{\theta(1-q) + (1-\theta) q} \delta
\end{aligned} \tag{3.1}$$

The left hand side of inequality (3.1) is the benefit from conforming to the belief that the other player is more likely to hold. The term $(2\theta - 1)(2q - 1)$ comes from the fact that, if a player views conflicting evidence as confirming evidence, (she believes that) the probability that she has a different belief from the other player will decrease by $(\theta q + (1-\theta)(1-q)) - (\theta(1-q) + (1-\theta) q) = (2\theta - 1)(2q - 1)$.

The right hand side of inequality (3.1) is the gain from implementing the policy that is more likely to match the state because of correct interpretation of conflicting evidence. Compared to misreading $\hat{L}$ as $\hat{R}$, taking $\hat{L}$ as $\hat{L}$ and making a policy choice accordingly increases the probability of choosing the optimal policy by $\frac{q-\theta}{\theta(1-q)+(1-\theta)q}$. If the expected gain from choosing a better policy exceeds the potential utility loss from two players making misaligned interpretations of their signals, a player will not have enough motivation to exhibit confirmatory bias when her partner does not distort the meaning of objective signals. Note that inequality (3.1) is also the condition under which an individual, in our model setup, does not exhibit confirmatory bias when she believes that the other decision maker is a non-strategic Bayesian rational agent.

**Confirmatory Bias Equilibrium (CBE)**

We now turn to characterize the conditions under which a Confirmatory Bias Equilibrium can be sustained. Given that Player $j$'s strategy is $\left(\tilde{s}_j(\hat{R}) = \hat{R}, \tilde{s}_j(\hat{L}) = \hat{R}\right)$ (i.e., suppose $j$ exhibits confirmatory bias), $\tilde{s}_i(\hat{R}) = \hat{R}$ is Player $i$'s best response if and only if

$$\frac{\theta q}{\theta q+(1-\theta)(1-q)}\delta \geq -\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 + \frac{(1-\theta)(1-q)}{\theta q+(1-\theta)(1-q)}\delta$$

$$\Longleftrightarrow \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \geq \frac{1-\theta-q}{\theta q+(1-\theta)(1-q)}\delta$$

Again, when the objective signal a player receives is consistent with her prior (i.e., $s_i = \hat{R}$), she does not misinterpret the signal for sure. Otherwise, the player will hold a different belief from her partner and get a lower expected payoff from her policy choice.

Alternatively, when the signal does not support Player $i$'s current belief, she will *misinterpret* it (i.e., $\tilde{s}_i(\hat{L}) = \hat{R}$ is Player $i$'s best response) if and only if

$$\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\delta \geq -\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 + \frac{(1-\theta)q}{\theta(1-q)+(1-\theta)q}\delta$$

$$\Longleftrightarrow \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \geq \frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta \tag{3.2}$$

Notice that the above inequality does not involve a player's belief about her partner's type (i.e., the objective signal her partner gets) since how the other player interprets a signal does not depend on his type—a signal is always interpreted as confirming the prior. As a result, unlike the equilibrium condition for a BUE (i.e., inequality (3.1)), the utility loss term for two players' misaligned beliefs, $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$, is not multiplied by the difference in a player's beliefs about her partner's type, $(2\theta - 1)(2q - 1)$, in the equilibrium condition for a CBE (i.e, inequality (3.2)).

inequality (3.2) shows that, when a player exhibits confirmatory bias, the other player will also exhibit confirmatory bias if the expected gain from choosing a better policy is less than the *certain* utility loss from two players making misaligned interpretations of their signals. Note that inequality (3.2) is also the condition under which an individual, in our model setup, exhibits confirmatory bias when she believes that the other decision maker is a non-strategic agent who has confirmatory bias due to other intrinsic motivations.

Before we proceed to the properties of the BUE and CBE, it is noteworthy that inequality (3.1) and (3.2) cannot be violated at the same time given $\theta \in (0.5, 1)$ and $q \in (\theta, 1)$. Thus, the existence of the equilibria is guaranteed under our model setup, as summarized in Theorem 1.

**Theorem 1.** *For $\theta \in (0.5, 1)$ and $q \in (\theta, 1)$, there always exists either a BUE or a CBE (or both).*

*Proof.* See Appendix B. □

**Some Properties of BUE and CBE**

In this subsection, we derive some properties of the Bayesian Updating Equilibrium and Confirmatory Bias Equilibrium with respect to the strength of prior belief ($\theta$) and the precision of a signal ($q$). To preview the main result, we show that confirmatory bias is more likely to occur as players hold a more extreme prior belief, in the sense that the BUE (CBE, resp.) can only be sustained when $\theta$ is below (above, resp.) some threshold value. Nevertheless, the impact of a signal's precision on the occurrence of confirmatory bias is ambiguous. We will have a more detailed discussion on the effect of $q$ in the next subsection about comparative statics.

Before we start the derivation of properties of equilibria, we define the following two functions:

$$f(\theta, q, \delta) = \left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q} \right)^2 (2\theta - 1)(2q - 1) - \frac{q - \theta}{\theta(1-q) + (1-\theta)q} \delta$$

$$g(\theta, q, \delta) = \left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q} \right)^2 - \frac{q - \theta}{\theta(1-q) + (1-\theta)q} \delta$$

The functions $f(\cdot)$ and $g(\cdot)$ correspond to the equilibrium condition for BUE and CBE, respectively. The BUE is supported if and only if $f(\cdot) \leq 0$, and the CBE is supported if and only if $g(\cdot) \geq 0$.

We first consider the case when $\delta \geq 1$, that is, when the payoff from a policy choice is relatively large. Proposition 1 shows that, when $\delta \geq 1$, a BUE can be sustained if and only if the prior belief is not too extreme.

**Proposition 1.** *If $\delta \geq 1$, then given any $q \in (0.5, 1)$, there exists $\theta_{q,\delta}^{BUE} \in (0.5, q)$ such that a BUE is supported iff $\theta \leq \theta_{q,\delta}^{BUE}$.*

*Proof.* The proof can be completed in three steps:

1. $f(\theta = 0.5, q, \delta) < 0$.
2. $f(\theta = q, q, \delta) > 0$.
3. When $\delta \geq 1$, $\frac{\partial}{\partial\theta} f(\theta, q, \delta) > 0$ for $\theta \in (0.5, 1)$.

Thus, given some $\delta \geq 1$ and $q \in (0.5, 1)$, $f(\cdot; q, \delta)$ only has one root in $(0.5, q)$, and $f(\theta', q, \delta) \leq 0$ iff $\theta'$ is less than or equal to that root given $\theta' \in (0.5, q)$.

For a complete proof, see Appendix B. □

At first glance, the above result seems straightforward. Given $s_i = \hat{L}$, on the one hand, the instrumental value of a signal as captured by $\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$ decreases in $\theta$. On the other hand, a player believes that the likelihood that two players' posterior beliefs are misaligned, characterized by $(2\theta - 1)(2q - 1)$, increases in $\theta$ if both players correctly interpret their signals. This argument seems to suggest that the benefit of deviating from a BUE must be increasing in $\theta$ (i.e., $\frac{\partial}{\partial\theta} f(\theta, q, \delta) > 0$), and therefore a BUE cannot be supported by a large $\theta$.

Nevertheless, notice that the utility loss from a failure in the coordination in posterior beliefs, characterized by $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$, decreases in $\theta$. The reason is that, if two players have different interpretations of a signal, their posteriors will be more dispersed when their common prior is less extreme. As a consequence, it is unclear whether the gain of deviating from a BUE and exhibiting confirmatory bias (i.e., the left hand side of inequality (3.1)) increases or decreases in the strength of prior. The above sketch of the proof of Proposition 1 shows that, when $\delta \geq 1$, the benefit of deviation from a BUE is indeed increasing in $\theta$.

Parallel to Proposition 1, Proposition 2 shows that, when $\delta \geq 1$, a CBE can be sustained if and only if the prior belief is extreme enough.

**Proposition 2.** *If $\delta \geq 1$, then given any $q \in (0.5, 1)$, there exists $\theta^{CBE}_{q,\delta} \in (0.5, q)$ such that a CBE is supported iff $\theta \geq \theta^{CBE}_{q,\delta}$.*

*Proof.* The proof can be completed in three steps.

1. $g(\theta = 0.5, q, \delta) < 0$.
2. $g(\theta = q, q, \delta) > 0$.

3. $\frac{\partial}{\partial\theta}g(\theta, q, \delta) > 0$.

Thus, given some $\delta \geq 1$ and $q \in (0.5, 1)$, $g(\cdot; q, \delta)$ only has one root in $(0.5, q)$, and $g(\theta', q, \delta) \geq 0$ iff $\theta'$ is greater than or equal to that root given $\theta' \in (0.5, q)$.

For a complete proof, see Appendix B. □

From previous discussion, we know that both the instrumental value of a signal and the distance between two misaligned posterior beliefs decrease in $\theta$. The proof of Proposition 2 indicates that, when $\delta \geq 1$, the change in a signal's instrumental value is more dominant.

We then turn to the properties of equilibria with respect to the accuracy of a signal. Proposition 3 shows that, when $\delta \geq 1$, a BUE can be sustained if and only if a signal is accurate enough.

**Proposition 3.** *If $\delta \geq 1$, then given any $\theta \in (0.5, 1)$, there exists $q^{BUE}_{\theta,\delta} \in (\theta, 1)$ such that a BUE is supported iff $q \geq q^{BUE}_{\theta,\delta}$.*

*Proof.* The proof can be completed in three steps:

1. $f(\theta, q = \theta, \delta) > 0$.
2. $f(\theta, q = 1, \delta) < 0$.
3. When $\delta \geq 1$, $\frac{\partial}{\partial q}f(\theta, q, \delta) < 0$ for $q \in (0.5, 1)$.

Thus, given some $\delta \geq 1$ and $\theta \in (0.5, 1)$, $f(\cdot; \theta, \delta)$ only has one root in $(\theta, 1)$, and $f(\theta, q', \delta) \leq 0$ iff $q'$ is greater than or equal to that root given $q' \in (\theta, 1)$.

For a complete proof, see Appendix B. □

Note that the instrumental value of a signal ($\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$), the likelihood that the players' posterior beliefs are misaligned ($(2\theta - 1)(2q - 1)$), and the utility loss from two players having misaligned beliefs ($(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$) all increase in $q$. The above sketch of the proof of Proposition 3 shows that, when $\delta \geq 1$, a change in a signal's instrumental value caused by an increase in $q$ is more dominant than the change in the product of the other two terms (i.e., the left hand side of inequality (3.1)).

Parallel to Proposition 3, Proposition 4 shows that, when $\delta \geq 1$, a CBE can be sustained if and only if a signal is not too accurate.

**Proposition 4.** *If $\delta \geq 1$, then given any $\theta \in (0.5, 1)$, there exists $q^{CBE}_{\theta,\delta} \in (\theta, 1)$ such that a CBE is supported iff $q \leq q^{CBE}_{\theta,\delta}$.*

*Proof.* The proof can be completed in four steps:

1. $g(\theta, q = \theta, \delta) > 0$.
2. $g(\theta, q = 1, \delta) < 0$.
3. When $\delta = 2$, $\frac{\partial}{\partial q} g(\theta, q, \delta) < 0$. Thus, $\frac{\partial}{\partial q} g(\theta, q, \delta) < 0$ for $\delta \geq 2$.
4. We can further show that, given any $\delta \in [1, 2]$ and $\theta \in (0.5, 1)$, $g(\cdot; \theta, \delta)$ only has one (real) root in $(\theta, 1)$.

Thus, given some $\delta \geq 1$ and $\theta \in (0.5, 1)$, $g(\cdot; \theta, \delta)$ only has one root in $(\theta, 1)$, and $g(\theta, q', \delta) \geq 0$ iff $q'$ is less than or equal to that root given $q' \in (\theta, 1)$.

For a complete proof, see Appendix B. □

As we have mentioned above, both a signal's instrumental value and the utility loss from belief misalignment increase in $q$. If the payoff from policy choice is sufficiently large, the change in the loss from belief misalignment will be dominated. However, this cannot be guaranteed under the assumption that $\delta \geq 1$. In other words, $\frac{\partial}{\partial q} g(\theta, q, \delta)$ may be positive when $\delta$ is close to one.[5] The above sketch of the proof of Proposition 4 reveals that, if the value of $q$ is already above the threshold $q^{CBE}_{\theta,\delta}$ such that a CBE cannot be sustained, the increase in the utility loss from belief misalignment caused by an increase in $q$ will not be large enough to offset the difference between this utility loss term and the instrumental value of a signal.

Next we explore the properties of BUE and CBE for the case when $\delta \in (0, 1)$, that is, when the payoff from policy choice is relatively small. Proposition 5 shows that, when $\delta \in (0, 1)$, there still exists a threshold of $\theta$ above which a BUE cannot be sustained. In other words, the result of Proposition 1 holds regardless of the value of $\delta$.

**Proposition 5.** *If $\delta \in (0, 1)$, then given any $q \in (0.5, 1)$, there exists $\theta^{BUE}_{q,\delta} \in (0.5, q)$ such that a BUE is supported iff $\theta \leq \theta^{BUE}_{q,\delta}$.*

*Proof.* The proof can be completed in three steps:

[5] For instance, given $\delta = 1.05$ and $\theta = 0.55$, $\frac{\partial}{\partial q} g(\theta, q, \delta)$ is positive when $q$ is in $(0.7953, 1)$.

1. $f(\theta = 0.5, q, \delta) < 0$.
2. $f(\theta = q, q, \delta) > 0$.
3. We can (numerically) show that, given any $\delta \in (0,1)$ and $q \in (0.5, 1)$, $\frac{\partial}{\partial\theta} f(\theta_0, q, \delta) > 0$ for $\theta_0 \in (0.5, q)$ such that $f(\theta_0, q, \delta) = 0$.

Thus, given some $\delta \in (0,1)$ and $q \in (0.5, 1)$, $f(\cdot; q, \delta)$ only has one root in $(0.5, q)$, and $f(\theta', q, \delta) \leq 0$ iff $\theta'$ is less than or equal to that root given $\theta' \in (0.5, q)$.

For a complete proof, see Appendix B. □

Unlike the proof of Proposition 1, $\frac{\partial}{\partial\theta} f(\theta, q, \delta)$ is not necessarily greater than 0 when $\delta \in (0,1)$.[6] Note that the impact of the payoff from policy choice is limited when $\delta$ is small. Thus, when $\delta \in (0,1)$, a player's utility is mainly determined by (a player's belief about) the likelihood that the players' beliefs are misaligned, and the utility loss from such belief misalignment. The above sketch of the proof of Proposition 5 reveals that, as the value of $\theta$ changes around the threshold, the change in the likelihood term (i.e., $(2\theta - 1)(2q - 1)$) dominates the change in the utility loss term. Moreover, although $\frac{\partial}{\partial\theta} f(\theta, q, \delta)$ may become negative as $\theta$ increases and approaches $q$, the above proposition suggests that the decrease in the utility loss from belief misalignment will not be large enough to offset the benefit of deviating from a BUE as long as $\theta \geq \theta^{BUE}_{q,\delta}$.

The next proposition indicates that, if $\delta \in (0,1)$, then the equilibrium condition for a CBE (i.e., inequality (3.2)) will always hold given that $q$ is sufficiently large.

**Proposition 6.** *Given any $\delta \in (0,1)$, there exists $q^{CBE}_{\delta} \in (0.5, 1)$ such that a CBE is supported by all $\theta \in (0.5, q'')$ when $q'' \geq q^{CBE}_{\delta}$.*

*Proof.* We can show that, when $\delta \in (0,1)$, $g(\theta, q, \delta)$ is greater than 0 for all $q \in [\frac{1+\delta}{2}, 1]$ given any $\theta \in (0.5, 1)$ by showing that

1. $g(\theta, q = \frac{1+\delta}{2}, \delta) > 0$.
2. $\frac{\partial}{\partial q} g(\theta, q, \delta) > 0$ for $q \in [\frac{1+\delta}{2}, 1]$.

Thus, $g(\theta, q, \delta) > 0$ for $q \in [\frac{1+\delta}{2}, 1]$.

For a complete proof, see Appendix B. □

[6]For instance, given $\delta = 0.1$ and $q = 0.9$, $\frac{\partial}{\partial\theta} f(\theta, q, \delta)$ is negative when $\theta$ is in $(0.8282, 0.9)$.

Proposition 6 suggests that, when $\delta$ is relatively small, a highly accurate signal may actually facilitate the occurrence of confirmatory bias. In words, if the payoff from policy choice is insignificant and the objective signal is accurate enough, the benefit from having the same belief as the other player will always dominate the payoff from policy choice, and a player who believes that her partner is exhibiting confirmatory bias will also exhibit confirmatory bias no matter how the prior is distributed. Note that, contrary to the result derived above, Proposition 4 indicates that a CBE is not sustainable under a large $q$ when $\delta \geq 1$. We defer the further discussion about how the impact of $q$ on the existence of confirmatory bias depends on the size of $\delta$ to the next subsection of comparative statics.

The last proposition in this subsection shows that, when $\delta \in (0, 1)$ and given that $q$ is small enough, a CBE can be sustained if and only if the prior belief if extreme enough.

**Proposition 7.** *If $\delta \in (0, 1)$, then given any $q \in (0.5, \frac{1+\delta}{2})$, there exists $\theta^{CBE}_{q,\delta} \in (0.5, q)$ such that a CBE is supported iff $\theta \geq \theta^{CBE}_{q,\delta}$.*

*Proof.* The proof can be completed in three steps.

1. $g(\theta = 0.5, q, \delta) < 0$.
2. $g(\theta = q, q, \delta) > 0$.
3. $\frac{\partial}{\partial\theta} g(\theta, q, \delta) > 0$ when $q \in (0.5, \frac{1+\delta}{2})$.

Thus, given some $\delta \geq 1$ and $q \in (0.5, 1)$, $g(\cdot; q, \delta)$ only has one root in $(0.5, q)$, and $g(\theta', q, \delta) \geq 0$ iff $\theta'$ is greater than or equal to that root given $\theta' \in (0.5, q)$.

For a complete proof, see Appendix B. □

The result and the proof of Proposition 7 are parallel to Proposition 2, with the exception that the statement in Proposition 7 holds only when the accuracy of a signal is below some threshold. The above sketch of proof reveals that, when $\delta \in (0, 1)$, the change in a signal's instrumental value induced by an increase in $\theta$ will still be more dominant than the change in the utility loss from belief misalignment if $q < \frac{1+\delta}{2}$.

Theorem 2 summarizes the main result in this subsection and characterizes the relationship between the strength of prior belief ($\theta$) and the existence of a BUE and a CBE.

**Theorem 2.** *Given some $\delta > 0$. For any $q \in (0.5, \frac{1+\delta}{2})$, there exist $\bar{\theta}_{q,\delta}$ and $\underline{\theta}_{q,\delta}$ where $0.5 < \underline{\theta}_{q,\delta} < \bar{\theta}_{q,\delta} < 1$ such that*

1. *only a BUE is supported when $\theta \in (0.5, \underline{\theta}_{q,\delta})$;*
2. *only a CBE is supported when $\theta \in (\bar{\theta}_{q,\delta}, 1)$;*
3. *both a BUE and a CBE are supported when $\theta \in (\underline{\theta}_{q,\delta}, \bar{\theta}_{q,\delta})$.*

*Proof.* Follows from Theorem 1, Proposition 1, 2, 5, 6, and 7. □

In words, given that the accuracy of a signal is moderate (below $\frac{1+\delta}{2}$), there are two thresholds of $\theta$ (depending on $q$ and $\delta$), $\underline{\theta}_{q,\delta}$ and $\bar{\theta}_{q,\delta}$, that divide the parameter space into three regions. When the prior is moderate (i.e., $\theta$ is less than $\underline{\theta}_{q,\delta}$), a confirmatory bias equilibrium cannot be sustained and a player always correctly interprets an objective signal. As the prior becomes more skewed towards state $R$ but not too extreme (i.e., $\theta$ is between $\underline{\theta}_{q,\delta}$ and $\bar{\theta}_{q,\delta}$), both a confirmatory bias equilibrium and a Bayesian updating equilibrium can be sustained, and whether a player will exhibit confirmatory bias is determined by her belief about the partner's strategy. When the prior is extreme (i.e., $\theta$ is above $\bar{\theta}_{q,\delta}$), a Bayesian updating equilibrium cannot be sustained and a player always interprets an objective signal as in favor of her current belief. Therefore, the above result suggests that the occurrence of confirmatory bias is positively associated with the extremeness of the players' prior belief.

**Comparative Statics**

In this subsection, we consider how the set of parameters that support a BUE or a CBE changes in response to changes in the skewness of the prior ($\theta$) and the accuracy of an objective signal ($q$). In other words, we examine whether an equilibrium is easier to be sustained under a higher or a lower value of $\theta$ and $q$.

Proposition 8 shows that a BUE (CBE, resp.) is easier to be supported by a smaller (larger, resp.) $\theta$.

**Proposition 8.** *Given any $\theta \in (0.5, 1)$, let $BUE_\theta \equiv \{(q,\delta) : \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 (2\theta-1)(2q-1) \leq \frac{q-\theta}{\theta(1-q)+(1-\theta q)}\delta \mid q \in (\theta, 1), \delta \in (0,\infty)\}$, then $BUE_{\theta''} \subseteq BUE_{\theta'}$ for all $\theta' \leq \theta''$. Alternatively, let $CBE_\theta \equiv \{(q,\delta) : \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \geq \frac{q-\theta}{\theta(1-q)+(1-\theta q)}\delta \mid q \in (\theta, 1), \delta \in (0,\infty)\}$, then $(\{(q,\delta) : q > \theta''\} \cap CBE_{\theta'}) \subseteq CBE_{\theta''}$ for all $\theta' \leq \theta''$.*

*Proof.* This proposition follows from Proposition 1, 2, 5, 6, and 7.

When a triple of parameters $(\theta, q, \delta)$ can support a BUE, by Proposition 1 and 5, we have $\theta \leq \theta^{BUE}_{q,\delta}$; furthermore, Proposition 1 and 5 imply that $(\theta_0, q, \delta)$ also supports a BUE given any $\theta_0 \leq \theta$, since $\theta_0 \leq \theta^{BUE}_{q,\delta}$.

When a tuple of parameters $(\theta, q, \delta)$ can support a CBE, there are two possible cases. First, if $\delta \in (0, 1)$ and $q \in [\frac{1+\delta}{2}, 1]$, then by Proposition 6, we know that $(\theta_0, q, \delta)$ also supports a CBE given any $\theta_0 \in (0.5, q)$. Otherwise, by Proposition 2 and 7, we have $\theta \geq \theta^{CBE}_{q,\delta}$; furthermore, Proposition 2 and 7 imply that $(\theta_0, q, \delta)$ also supports a CBE given any $\theta_0 \in (\theta, q)$, since $\theta_0 \geq \theta^{CBE}_{q,\delta}$. □

In words, the above proposition reveals that, with fixed $(q, \delta)$, a BUE being sustainable under a given strength of prior implies that it is also sustainable under a weaker prior. Alternatively, if a CBE can be sustained under a given strength of prior, then it is also sustainable under a stronger prior. Note that this proposition directly follows from the results in the previous subsection. Therefore, it can be viewed as another way to characterize the fact that a larger $\theta$ will facilitate the occurrence of confirmatory bias.

Proposition 9 shows that a BUE (CBE, resp.) is easier to be supported by a smaller $q$ given a small (large, resp.) $\delta$, whereas it is easier to be supported by a larger $q$ given a large (small, resp.) $\delta$.

**Proposition 9.** *Given $q \in (0.5, 1)$, let $BUE_q \equiv \{(\theta, \delta) : \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 (2\theta-1)(2q-1) \leq \frac{q-\theta}{\theta(1-q)+(1-\theta q)}\delta \mid \theta \in (0.5, q), \delta \in (0, \infty)\}$. Given $q'$ and $q''$ such that $q' \leq q''$, there exists $\delta^{BUE}_{q',q''} > 0$ such that*

1. *$(\theta, \delta) \in BUE_{q''} \implies (\theta, \delta) \in BUE_{q'}$ for all $\delta \in (0, \delta^{BUE}_{q',q''})$;*
2. *$(\theta, \delta) \in BUE_{q'} \implies (\theta, \delta) \in BUE_{q''}$ for all $\delta \in (\delta^{BUE}_{q',q''}, \infty)$.*

*Alternatively, given $q \in (0.5, 1)$, let $CBE_q \equiv \{(\theta, \delta) : \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \geq \frac{q-\theta}{\theta(1-q)+(1-\theta q)}\delta \mid \theta \in (0.5, q), \delta \in (0, \infty)\}$. Given $q'$ and $q''$ such that $q' \leq q''$, there exists $\delta^{CBE}_{q',q''} > 0$ such that*

1. *$(\theta, \delta) \in CBE_{q'} \implies (\theta, \delta) \in CBE_{q''}$ for all $\delta \in (0, \delta^{CBE}_{q',q''})$;*
2. *$(\theta, \delta) \in CBE_{q''} \implies (\theta, \delta) \in CBE_{q'}$ for all $\delta \in (\delta^{CBE}_{q',q''}, \infty)$.*

*Proof.* See Appendix B. □

We can formally illustrate the intuition behind the above proposition with the partial derivative of $f(\theta, q, \delta)$ and $g(\theta, q, \delta)$ with respect to $q$. Recall that a BUE can be sustained when $f(\theta, q, \delta) \leq 0$ and a CBE can be sustained when $g(\theta, q, \delta) \geq 0$. Moreover, it can be shown that

$$\begin{aligned}\frac{\partial}{\partial q} f(\theta, q, \delta) = 2(2\theta - 1) \Bigg[ &\left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q} \right)^2 + \\ &(2q-1) \left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q} \right) \cdot \\ &\theta(1-\theta) \left( \frac{1}{(\theta q + (1-\theta)(1-q))^2} + \frac{1}{(\theta(1-q) + (1-\theta)q)^2} \right) \Bigg] \\ &- \frac{2\theta(1-\theta)}{(\theta(1-q) + (1-\theta)q)^2} \delta\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial q} g(\theta, q, \delta) = 2\theta(1-\theta) \Bigg[ &\left( \frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q} \right) \cdot \\ &\left( \frac{1}{(\theta q + (1-\theta)(1-q))^2} + \frac{1}{(\theta(1-q) + (1-\theta)q)^2} \right) \Bigg] \\ &- \frac{2\theta(1-\theta)}{(\theta(1-q) + (1-\theta)q)^2} \delta\end{aligned}$$

Note that, in both $\frac{\partial}{\partial q} f(\theta, q, \delta)$ and $\frac{\partial}{\partial q} g(\theta, q, \delta)$, the first term (i.e., the term followed by a minus sign) is positive while the second term (i.e., $-\frac{2\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\delta$) is negative. When $\delta$ is small, $\frac{\partial}{\partial q} f(\theta, q, \delta)$ and $\frac{\partial}{\partial q} g(\theta, q, \delta)$ are greater than 0; thus, a BUE is easier to be supported by a smaller $q$ and a CBE is easier to be supported by a larger $q$. Alternatively, when $\delta$ is large, $\frac{\partial}{\partial q} f(\theta, q, \delta)$ and $\frac{\partial}{\partial q} g(\theta, q, \delta)$ are less than 0, which implies that a BUE is easier to be supported by a larger $q$ and a CBE is easier to be supported by a smaller $q$ in this case.

The above proposition, again, highlights the trade-off between correctly interpreting an objective signal (and choosing the optimal policy accordingly) and having the same beliefs as the other player. When $q$ is large (i.e., the objective signal is accurate), the instrumental value of a signal is relatively large; however, the distance between the posteriors induced by different signals is also large. If $\delta$ is small, the disutility from having opposite beliefs as the other will overwhelm the (expected) gain from choosing the policy matching an objective signal.

We notice that there is empirical evidence indicating that, on aggregate, the credibility of information source does not have a significant impact on how individuals interpret the information. Nyhan and Reifler (2010) find that, in their experiment, news source manipulation has a null effect on how a subject perceives the information that tries to correct the subject's misperceptions about political events. Our model suggests that such finding may arise from the heterogeneity in people's perceptions about the importance of those events or about whether their choices really matter in determining the final policy outcome.

### 3.4 Extensions

In this section, we consider two extensions of the basic model developed in Section 3. First, we consider the scenario in which there are more than two members in a group. Second, we discuss the case in which a player uses the objective signal she receives to form her belief about her partner's type (i.e., the likelihood of her partner's objective signal) before she interprets the signal.

#### Group Size

In this subsection, we explore how the group size affects the equilibrium conditions. Formally, assume that there are $N$ players (i.e., $i \in \{1, 2, ..., N\}$), where $N > 2$. For simplicity of exposition, we further assume that $N$ is even (so that $N - 1$ is odd). Player $i$ will, in addition to the payoff from policy choice, experience a quadratic utility loss from the distance between her subjective belief (induced by $\tilde{s}_i$) and the *median* of her partners' subjective beliefs. In other words, a player will suffer a utility loss of $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$ if her subjective posterior belief differs from the posterior belief that the majority of the other players hold. Note that the baseline model with $N = 2$ is a special case of this setup.

As in the previous section, we focus on Bayesian Updating Equilibrium (BUE) and Confirmatory Bias Equilibrium (CBE). Notice that the equilibrium condition for a CBE (i.e., inequality (3.2)) does not change with the group size. Suppose that the players except Player $i$ always interpret an objective signal as the supporting one (i.e., $\tilde{s}_j(s_j) = \hat{R}$ for all $j \neq i$ and for all $s_j$), Player $i$ will suffer the utility loss from belief misalignment $-(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})^2$ *for sure* if she chooses to correctly interpret the objective signal that contradicts the prior belief. As a result, when $s_i = \hat{L}$, Player $i$ faces the same trade-off as characterized in Section 3.2 between having the same posterior belief with the majority *for sure* and correctly interpreting the signal, which induces a policy choice that is more likely to match the true state.

However, if players never misread their signals, then the likelihood (derived from a player's belief about others' types) that the majority receives $\hat{R}$ will change in the group size. Thus, the equilibrium condition for a BUE will be different as $N$ changes. For example, when $N = 4$, Player $i$ believes that the probability that the median posterior belief of the other players is induced by $\hat{R}$ is equal to

$$\begin{aligned}
&Pr_i(s_j = \hat{R} \quad \forall j \neq i) + Pr_i(s_{j_1} = \hat{R}, s_{j_2} = \hat{R}, s_{j_3} = \hat{L} \quad \text{for } j_k \neq i \text{ for all } k \text{ and } j_k \neq j_l \text{ for all } k, l) \\
&= (\theta q + (1-\theta)(1-q))^3 + 3(\theta q + (1-\theta)(1-q))^2(\theta(1-q) + (1-\theta)q)
\end{aligned}$$

Let $\lambda \equiv \theta q + (1-\theta)(1-q)$. Then a BUE can be sustained as an equilibrium if and only if

$$\begin{aligned}
&-(\lambda^3 + 3\lambda^2(1-\lambda))\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)^2 \\
&\quad + \frac{(1-\theta)q}{\theta(1-q) + (1-\theta)q}\delta \\
&\geq -((1-\lambda)^3 + 3(1-\lambda)^2\lambda)\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)^2 \\
&\quad + \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\delta \\
\iff &(2\theta - 1)(2q - 1)(1 + 2\lambda(1-\lambda))\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)^2 \\
&\quad \leq \frac{q - \theta}{\theta(1-q) + (1-\theta)q}\delta
\end{aligned}$$

Note that the LHS of the above inequality is greater than the LHS of inequality (3.1), which implies that a BUE is harder to be supported when the group size is four instead of two.

In general, as the group size is $2k$ ($k \in \mathbb{N}$), the equilibrium condition for a BUE becomes

$$
\begin{aligned}
&-\left(\sum_{j=0}^{k-1}\binom{2k-1}{j}\lambda^{2k-1-j}(1-\lambda)^j\right)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \\
&\quad+\frac{(1-\theta)q}{\theta(1-q)+(1-\theta)q}\delta \\
\geq\ &-\left(1-\left(\sum_{j=0}^{k-1}\binom{2k-1}{j}\lambda^{2k-1-j}(1-\lambda)^j\right)\right)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2+ \\
&\quad\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\delta \\
\iff\ &\left(2\left(\sum_{j=0}^{k-1}\binom{2k-1}{j}\lambda^{2k-1-j}(1-\lambda)^j\right)-1\right)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \\
&\quad\leq\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta
\end{aligned}
$$

The following proposition shows that the LHS of the above inequality is increasing in $k$.

**Proposition 10.** *Given* $\lambda \in (0.5, 1)$, *let* $f_\lambda(k) = 2\left(\sum_{j=0}^{k-1}\binom{2k-1}{j}\lambda^{2k-1-j}(1-\lambda)^j\right)-1$. *Then* $f_\lambda(k+1) - f_\lambda(k) > 0$ *for all* $k \in \mathbb{N}$ *and* $\lambda \in (0.5, 1)$.

*Proof.* See Appendix B. □

In words, Proposition 10 reveals that, in general, a BUE becomes harder to be sustained as the group size $N$ increases. Since the equilibrium condition for a CBE does not change in $N$, this result suggests that confirmatory bias becomes more likely to occur as a group expands. This finding is consistent with the intuition that, in a larger group, an individual is more willing to conform to the belief that is shared by the majority of the group, since he/she has more interaction opportunities with other group members and (thus) faces greater peer pressure.

**Belief about the Partner's Type**

In our basic model, we assume that a player's belief about her partner's type, or her partner's (objective) signal $s_j$, is purely induced by the common prior belief and is not dependent on the player's own objective signal $s_i$. This assumption accounts for the idea that an individual does not use the evidence at hand to update her belief about the state before she interprets the evidence. However, a caveat of this assumption is that it leads to inconsistency between a player's belief about the other player's type and the probability of the true state adopted to evaluate the expected policy payoff, which is conditional on $s_i$. Hence, in this subsection, we discuss how the equilibrium conditions will change if we allow Player $i$'s belief about Player $j$'s type to depend on $s_i$.

Formally, when Player $i$'s belief about her partner's type depends on both $\theta$ and $s_i$, given Player $j$'s strategy $\left(\tilde{s}_j(s_j = \hat{R}), \tilde{s}_j(s_j = \hat{L})\right)$, Player $i$'s maximization problem can be rewritten as:

$$\max_{\tilde{s}_i \in \{\hat{L},\hat{R}\}} - \mathbb{P}(s_j = \hat{R}|s_i;\theta)\left(p_i[\tilde{s}_i] - p_j[\tilde{s}_j(s_j = \hat{R})]\right)^2 - \mathbb{P}(s_j = \hat{L}|s_i;\theta)\left(p_i[\tilde{s}_i] - p_j[\tilde{s}_j(s_j = \hat{L})]\right)^2$$
$$+ \mathbb{P}(\omega = R|s_i)\cdot\delta\mathbb{I}_{\{p_i[\tilde{s}_i]\geq\frac{1}{2}\}} + \mathbb{P}(\omega = L|s_i)\cdot\delta\mathbb{I}_{\{p_i[\tilde{s}_i]<\frac{1}{2}\}}$$

where $\mathbb{P}(s_j = \hat{R}|s_i;\theta) = \mathbb{P}(\omega = R|s_i)q + \mathbb{P}(\omega = L|s_i)(1-q)$ and $\mathbb{P}(s_j = \hat{L}|s_i;\theta) = \mathbb{P}(\omega = R|s_i)(1-q) + \mathbb{P}(\omega = L|s_i)q$.

We first consider the equilibrium condition for a BUE. Under the assumption of $q > \theta$, both $\mathbb{P}(\omega = R|s_i = \hat{R})$ and $\mathbb{P}(\omega = L|s_i = \hat{L})$ are greater than one half, which implies that $\mathbb{P}(s_j = \hat{R}|s_i = \hat{R};\theta) > \frac{1}{2}$ and $\mathbb{P}(s_j = \hat{L}|s_i = \hat{L};\theta) > \frac{1}{2}$. Thus, if Player $j$ always correctly interprets his objective signal (i.e., $\tilde{s}_j(s_j) = s_j$ for $s_j = \hat{R}, \hat{L}$), Player $i$ will never have an incentive to misinterpret $s_i$ as well, since she will have a higher probability of both forming the same posterior belief as the other player and choosing the optimal policy when following the literal meaning of $s_i$. This implies that, when a player's belief about the objective signal received by her partner depends on $s_i$, a BUE can be sustained for all $\theta \in (0.5, 1)$ and $q \in (\theta, 1)$.

It is noteworthy, however, that the equilibrium condition for a CBE remains the same under this new assumption. Given that Player $j$ always interprets his signal as supporting evidence (i.e., $\tilde{s}_j(s_j) = \hat{R}$ for $s_j = \hat{R}, \hat{L}$), Player $i$ clearly has no incentive to misinterpret her objective signal when $s_i = \hat{R}$. Moreover, when $s_i = \hat{L}$, Player $i$ will have a different posterior belief from Player $j$ with certainty under correct interpretation of $s_i$, regardless of how Player $i$'s belief about $j$'s type is established. As a result, Player $i$'s best response is to interpret conflicting evidence as supporting one as well if (and only if) the utility loss from belief misalignment is greater than the expected gain from choosing a policy that is more likely to match the state, that is, $\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 \geq \frac{q-\theta}{\theta(1-q)+(1-\theta q)}\delta$, which is the same as inequality (3.2). This result implies that the properties and comparative statics related to a CBE in the basic model also hold under the assumption that a player's belief about the partner's type depends on $s_i$, and thus demonstrates the robustness of our findings in the previous section.

### 3.5 Conclusion

In this paper, we demonstrate how conformity in beliefs can lead to confirmatory bias in a strategic interaction environment. We build our model upon a standard decision-making model under incomplete information in which two agents individually update their belief about the binary state of nature after receiving a (binary) signal and make a policy choice accordingly. We incorporate the elements of strategic interaction by (i) introducing a cost from misaligned posterior beliefs between the agents to their utility, and (ii) allowing an agent to manipulate the interpretation of a binary signal.

Specifically, in our setup, an agent's perception of a signal need not be the same as the signal's literal meaning. When observing a signal that conflicts with the (asymmetric) common prior, an agent has to trade off between correctly perceiving the signal to get higher expected reward from policy choice, and misinterpreting the signal as in consistent with the prior to increase the likelihood that two agents' posteriors are aligned. We study how such trade-off is affected by the strength of the prior and the quality of an information source.

Under this framework, we show that a stronger prior belief about the world will facilitate the emergence of confirmatory bias. The intuition is that the instrumental value of a signal decreases as the prior becomes more biased, and the incentive to correctly interpret the signal will eventually be dominated by the incentive to conform to the majority belief. This result is consistent with previous empirical works studying political polarization which report the backfire effect of information—providing information that is inconsistent with an individual's viewpoint may prompt the individual to exhibit more polarized belief.

In addition, we find that an increase in the accuracy of the information source does not necessarily alleviate confirmatory bias. The reason is that the distance between (and thus the loss associated with) misaligned posterior beliefs induced by opposite signals goes up as a signal becomes more accurate. As a result, when the policy choice is relatively unimportant, the high accuracy of a signal could increase an individual's tendency to misinterpret conflicting evidence. This finding suggests that exposing individuals to opposing views may not be an effective way to reduce political polarization, even if the information source is of high quality.

We finish by noting several model extensions that are worth further investigation.

First, in our current model, the disutility term from belief misalignment is a quadratic loss from the distance between two players' posterior beliefs. A reasonable robustness check is to explore whether our findings still hold under a more general assumption on the form of disutility term (e.g., a concave function). Second, to capture conformity in a group, the players in our model are homogeneous—they share a common prior belief and have the same utility function. It may be interesting to introduce heterogeneity into the environment. For instance, we can consider the interaction between two ideological groups, each of which prefers to form a belief that is distanced from the other group's ideology. Such heterogeneity may provide further insights on the origin and dynamics of political polarization.

**References**

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics*, *94*(4), 749–775.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Banks, J. S., & Calvert, R. L. (1992). A battle-of-the-sexes game with incomplete information. *Games and Economic Behavior*, *4*(3), 347–372.

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, *80*(2), 429–462.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *Quarterly Journal of Economics*, *117*(3), 871–915.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics*, *126*(2), 805–855.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141–64.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations 2* (pp. 69–186). Elsevier.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, *102*(5), 841–877.

Bracha, A., & Brown, D. J. (2012). Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, *75*(1), 67–80.

Caplin, A., & Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, *116*(1), 55–79.

Carlsson, H., & Van Damme, E. (1993). Global games and equilibrium selection. *Econometrica*, *61*(5), 989–1018.

Chauvin, K. P. (2020). Belief updating with dissonance reduction. *Working Paper*.

Fryer Jr, R. G., Harms, P., & Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, *17*(5), 1470–1501.

Gibbons, R. S. (1992). *Game theory for applied economists*. Princeton University Press.

Gottlieb, D. (2014). Imperfect memory and choice under risk. *Games and Economic Behavior*, *85*, 127–158.

Jones, S. R. (1984). *The economics of conformism*. Oxford: Basil Blackwell.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Luzzati, T. (1999). Economic theory and conformism. In M. Gallegati & A. Kirman (Eds.), *Beyond the representative agent* (pp. 111–140). Edward Elgar.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, *114*(1), 37–82.

*Appendix A*

# PROOFS FOR CHAPTER 2

## A.1 Omitted Proofs of General Properties

**Proof of Lemma 1**

By definition 1, for any $(\mu, \sigma) \in \Psi^{\chi}$, any history $h^{t-1}$, any player $i$ and any type profile $\theta = (\theta_i, \theta_{-i})$,

$$\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)[\chi\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i) + (1-\chi)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i})]$$

$$= \chi \underbrace{\left[\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\right]}_{=1} \bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i) + (1-\chi) \underbrace{\left[\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i})\right]}_{=\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i)}$$

$$= \bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i).$$

Therefore, since $(\mu, \sigma) \in \Psi^{\chi}$, with some rearrangement, it follows that

$$\begin{aligned}
\mu_i(\theta_{-i}|h^t, \theta_i) &= \frac{\mu_i(\theta_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i}, \theta_i)}{\sum_{\theta'_{-i}\in\Theta_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma^{\chi}_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i}, \theta_i)} \\
&= \frac{\mu_i(\theta_{-i}|h^{t-1}, \theta_i)[\chi\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i) + (1-\chi)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i})]}{\bar{\sigma}_{-i}(a^t_{-i}|h^{t-1}, \theta_i)} \\
&= \chi\mu_i(\theta_{-i}|h^{t-1}, \theta_i) + (1-\chi)\left[\frac{\mu_i(\theta_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta_{-i})}{\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a^t_{-i}|h^{t-1}, \theta'_{-i})}\right]. \blacksquare
\end{aligned}$$

**Proof of Proposition 1**

The proof is similar to the proof for sequential equilibrium and proceeds in three steps. First, for any finite multi-stage games with observed actions, $\Gamma$, we construct an $\epsilon$-perturbed game $\Gamma^{\epsilon}$ that is identical to $\Gamma$ but every player in every information set has to play any available action with probability at least $\epsilon$. Second, we defined a cursed best-response correspondence for $\Gamma^{\epsilon}$ and prove that the correspondence has a fixed point by Kakutani's fixed point theorem. Finally, in step 3, we use a sequence of fixed points in perturbed games, with $\epsilon$ converging to 0, where the limit of this sequence is a $\chi$-CSE.

**Step 1:** Let $\Gamma^{\epsilon}$ be a game identical to $\Gamma$ but for each player $i \in N$, player $i$ must play any available action in every information set $\mathcal{I}_i = (\theta_i, h^t)$ with probability at least $\epsilon$ where $\epsilon < \frac{1}{\sum_{j=1}^{n} |A_j|}$. Let $\Sigma^\epsilon = \times_{j=1}^{n} \Sigma_j^\epsilon$ be set of feasible behavioral strategy profiles for players in the perturbed game $\Gamma^\epsilon$. For any behavioral strategy profile $\sigma \in \Sigma^\epsilon$, let $\mu^\chi(\cdot) \equiv (\mu_i^\chi(\cdot))_{i=1}^{n}$ be the belief system induced by $\sigma$ via $\chi$-cursed Bayes' rule. That is, for each player $i \in N$, information set $\mathcal{I}_i = (\theta_i, h^t)$ where $h^t = (h^{t-1}, a^t)$ and type profile $\theta_{-i} \in \Theta_{-i}$,

$$\mu_i^\chi(\theta_{-i}|h^t, \theta_i) = \chi \mu_i^\chi(\theta_{-i}|h^{t-1}, \theta_i) + (1-\chi)\left[\frac{\mu_i^\chi(\theta_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a_{-i}^t|h^{t-1}, \theta_{-i})}{\sum_{\theta'_{-i} \in \Theta_{-i}} \mu_i^\chi(\theta'_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a_{-i}^t|h^{t-1}, \theta'_{-i})}\right].$$

Notice that the $\chi$-cursed Bayes' rule is only defined on the family of multi-stage games with observed actions. As $\sigma$ is fully mixed, the belief system is uniquely pinned down.

Finally, let $B^\epsilon : \Sigma^\epsilon \rightrightarrows \Sigma^\epsilon$ be the cursed best response correspondence which maps any behavioral strategy profile $\sigma \in \Sigma^\epsilon$ to the set of $\epsilon$-constrained behavioral strategy profiles $\tilde{\sigma} \in \Sigma^\epsilon$ that are best replies given the belief system $\mu^\chi(\cdot)$.

**Step 2:** Next, fix any $0 < \epsilon < \frac{1}{\sum_{j=1}^{n} |A_j|}$ and show that $B^\epsilon$ has a fixed point by Kakutani's fixed point theorem. We check the conditions of the theorem:

1. It is straightforward that $\Sigma^\epsilon$ is compact and convex.
2. For any $\sigma \in \Sigma^\epsilon$, as $\mu^\chi(\cdot)$ is uniquely pinned down by $\chi$-cursed Bayes' rule, it is straightforward that $B^\epsilon(\sigma)$ is non-empty and convex.
3. To verify that $B^\epsilon$ has a closed graph, take any sequence of $\epsilon$-constrained behavioral strategy profiles $\{\sigma^k\}_{k=1}^{\infty} \subseteq \Sigma^\epsilon$ such that $\sigma^k \to \sigma \in \Sigma^\epsilon$ as $k \to \infty$, and any sequence $\{\tilde{\sigma}^k\}_{k=1}^{\infty}$ such that $\tilde{\sigma}^k \in B^\epsilon(\sigma^k)$ for any $k$ and $\tilde{\sigma}^k \to \tilde{\sigma}$. We want to prove that $\tilde{\sigma} \in B^\epsilon(\sigma)$.

   Fix any player $i \in N$ and information set $\mathcal{I}_i = (\theta_i, h^t)$. For any $\sigma \in \Sigma^\epsilon$, recall that $\sigma_{-i}^\chi(\cdot)$ is player $i$'s $\chi$-cursed perceived behavioral strategies of other players induced by $\sigma$. Specifically, for any type profile $\theta \in \Theta$, non-terminal history $h^{t-1}$ and action profile $a_{-i}^t \in A_{-i}(h^{t-1})$, $\sigma_{-i}^\chi(a_{-i}^t|h^{t-1}, \theta_{-i}, \theta_i) = \chi\bar{\sigma}_{-i}(a_{-i}^t|h^{t-1}, \theta_i) + (1-\chi)\sigma_{-i}(a_{-i}^t|h^{t-1}, \theta_{-i})$.

   Additionally, recall that $\rho_i^\chi(\cdot)$ is player $i$'s belief about the terminal nodes (conditional on the history and type profile), which is also induced by $\sigma$. Since $\mu^\chi(\cdot)$ is continuous in $\sigma$ we have that $\sigma_{-i}^\chi(\cdot)$ and $\rho_i^\chi(\cdot)$ are also continuous in $\sigma$. We further define $\mathcal{S}^k_{\mathcal{I}_i} \equiv \left\{\sigma'_i \in \Sigma^\epsilon_i : \sigma'_i(\,\cdot\,|\mathcal{I}_i) = \tilde{\sigma}^k_i(\,\cdot\,|\mathcal{I}_i)\right\}$ and $\mathcal{S}_{\mathcal{I}_i} \equiv \left\{\sigma'_i \in \Sigma^\epsilon_i : \sigma'_i(\,\cdot\,|\mathcal{I}_i) = \tilde{\sigma}_i(\,\cdot\,|\mathcal{I}_i)\right\}$. Since $\tilde{\sigma}^k \in B^\epsilon(\sigma^k)$, for any $\sigma'_i \in \Sigma^\epsilon_i$, we can obtain that

$$\max_{\sigma''_i \in \mathcal{S}^k_{\mathcal{I}_i}} \left\{ \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu^\chi_i[\sigma^k](\theta_{-i}|h^t,\theta_i)\rho^\chi_i(h^T|h^t,\theta,\sigma^\chi_{-i}[\sigma^k],\sigma''_i)u_i(h^T,\theta_i,\theta_{-i}) \right\}$$
$$\geq \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu^\chi_i[\sigma^k](\theta_{-i}|h^t,\theta_i)\rho^\chi_i(h^T|h^t,\theta,\sigma^\chi_{-i}[\sigma^k],\sigma'_i)u_i(h^T,\theta_i,\theta_{-i}).$$

By continuity, as we take limits on both sides, we can find $\tilde{\sigma} \in B^\epsilon(\sigma)$ because

$$\max_{\sigma''_i \in \mathcal{S}_{\mathcal{I}_i}} \left\{ \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu^\chi_i[\sigma](\theta_{-i}|h^t,\theta_i)\rho^\chi_i(h^T|h^t,\theta,\sigma^\chi_{-i}[\sigma],\sigma''_i)u_i(h^T,\theta_i,\theta_{-i}) \right\}$$
$$\geq \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu^\chi_i[\sigma](\theta_{-i}|h^t,\theta_i)\rho^\chi_i(h^T|h^t,\theta,\sigma^\chi_{-i}[\sigma],\sigma'_i)u_i(h^T,\theta_i,\theta_{-i}).$$

By Kakutani's fixed point theorem, $B^\epsilon$ has a fixed point.

**Step 3:** For any $\epsilon$, let $\sigma^\epsilon$ be a fixed point of $B^\epsilon$ and $\mu^\epsilon$ be the belief system induced by $\sigma^\epsilon$ via $\chi$-cursed Bayes' rule. We combine these two components and let $(\mu^\epsilon,\sigma^\epsilon)$ be the induced assessment. We now consider a sequence of $\epsilon \to 0$, where $\{(\mu^\epsilon,\sigma^\epsilon)\}$ is the corresponding sequence of assessments. By compactness and the finiteness of $\Gamma$, the Bolzano-Weierstrass theorem guarantees the existence of a convergent subsequence of the assessments. As $\epsilon \to 0$, let $(\mu^\epsilon,\sigma^\epsilon) \to (\mu^*,\sigma^*)$. By construction, the limit assessment $(\mu^*,\sigma^*)$ satisfies $\chi$-consistency and sequential rationality. Hence, $(\mu^*,\sigma^*)$ is a $\chi$-CSE. ■

**Proof of Proposition 2**

To prove $\Phi(\chi)$ is upper hemi-continuous in $\chi$, consider any sequence of $\{\chi_k\}_{k=1}^\infty$ such that $\chi_k \to \chi^* \in [0,1]$, and any sequence of CSE, $\{(\mu^k,\sigma^k)\}$, such that $(\mu^k,\sigma^k) \in \Phi(\chi_k)$ for all $k$. Let $(\mu^*,\sigma^*)$ be the limit assessment, i.e., $(\mu^k,\sigma^k) \to (\mu^*,\sigma^*)$. We need to show that $(\mu^*,\sigma^*) \in \Phi(\chi^*)$.

For simplicity, for any player $i \in N$, any information set $\mathcal{I}_i = (h^t,\theta_i)$, any $\sigma'_i \in \Sigma_i$, and any $\sigma \in \Sigma$, the expected payoff under the belief system $\mu^\chi(\cdot)$ induced by $\sigma$ is denoted as:

$$\mathbb{E}_{\mu^\chi[\sigma]}\left[u_i(\sigma'_i,\sigma_{-i}|h^t,\theta_i)\right] \equiv \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{h^T\in\mathcal{H}^T} \mu^\chi_i(\theta_{-i}|h^t,\theta_i)\rho^\chi_i(h^T|h^t,\theta,\sigma^\chi_{-i},\sigma'_i)u_i(h^T,\theta_i,\theta_{-i}).$$

Suppose $(\mu^*, \sigma^*) \notin \Phi(\chi^*)$. Then there exists some player $i \in N$, some information set $\mathcal{I}_i = (h^t, \theta_i)$, some $\sigma_i' \in \Sigma_i$, and some $\epsilon > 0$ such that

$$\mathbb{E}_{\mu^{\chi^*}[\sigma^*]}\left[u_i(\sigma_i', \sigma_{-i}^*|h^t, \theta_i)\right] - \mathbb{E}_{\mu^{\chi^*}[\sigma^*]}\left[u_i(\sigma_i^*, \sigma_{-i}^*|h^t, \theta_i)\right] > \epsilon. \tag{A}$$

Since $\mu^\chi(\cdot)$ is continuous in $\chi$, it follows that for any strategy profile $\sigma$, $\sigma_{-i}^\chi(\cdot)$ and $\rho_i^\chi(\cdot)$ are both continuous in $\chi$. Thus, there exists a sufficiently large $M_1$ such that for every $k \geq M_1$,

$$\left|\mathbb{E}_{\mu^{\chi^k}[\sigma^k]}\left[u_i(\sigma_i^k, \sigma_{-i}^k|h^t, \theta_i)\right] - \mathbb{E}_{\mu^{\chi^*}[\sigma^*]}\left[u_i(\sigma_i^*, \sigma_{-i}^*|h^t, \theta_i)\right]\right| < \frac{\epsilon}{3}. \tag{B}$$

Similarly, there exists a sufficiently large $M_2$ such that for every $k \geq M_2$,

$$\left|\mathbb{E}_{\mu^{\chi^k}[\sigma^k]}\left[u_i(\sigma_i', \sigma_{-i}^k|h^t, \theta_i)\right] - \mathbb{E}_{\mu^{\chi^*}[\sigma^*]}\left[u_i(\sigma_i', \sigma_{-i}^*|h^t, \theta_i)\right]\right| < \frac{\epsilon}{3}. \tag{C}$$

Therefore, for any $k \geq \max\{M_1, M_2\}$, inequalities (A), (B) and (C) imply:

$$\mathbb{E}_{\mu^{\chi^k}[\sigma^k]}\left[u_i(\sigma_i', \sigma_{-i}^k|h^t, \theta_i)\right] - \mathbb{E}_{\mu^{\chi^k}[\sigma^k]}\left[u_i(\sigma_i^k, \sigma_{-i}^k|h^t, \theta_i)\right] > \frac{\epsilon}{3},$$

implying that $\sigma_i'$ is a profitable deviation for player $i$ at information set $\mathcal{I}_i = (h^t, \theta_i)$, which contradicts $(\mu^k, \sigma^k) \in \Phi(\chi_k)$. Therefore, $(\mu^*, \sigma^*) \in \Phi(\chi^*)$, as desired. ■

**Proof of Proposition 3**

Fix any $\chi \in [0,1]$ and let $(\mu, \sigma)$ be a $\chi$-consistent assessment. We prove the result by contradiction. Suppose $(\mu, \sigma)$ does not satisfy $\chi$-dampened updating property. Then there exists $i \in N$, $\tilde{\theta} \in \Theta$ and a non-terminal history $h^t$ such that $\mu_i(\theta_{-i}|h^t, \tilde{\theta}_i) < \chi\mu_i(\theta_{-i}|h^{t-1}, \tilde{\theta}_i)$.

Since $(\mu, \sigma)$ is $\chi$-consistent, there exists a sequence $\{(\mu^k, \sigma^k)\} \subseteq \Psi^\chi$ such that $(\mu^k, \sigma^k) \to (\mu, \sigma)$ as $k \to \infty$. By Lemma 1, we know for this $i, \tilde{\theta}$ and $h^t$, $\mu_i^k(\tilde{\theta}_{-i}|h^t, \tilde{\theta}_i)$ equals to

$$\chi\mu_i^k(\tilde{\theta}_{-i}|h^{t-1}, \tilde{\theta}_i) + (1-\chi)\left[\frac{\mu_i^k(\tilde{\theta}_{-i}|h^{t-1}, \tilde{\theta}_i)\sigma_{-i}^k(a_{-i}^t|h^{t-1}, \tilde{\theta}_{-i})}{\sum_{\theta_{-i}'}\mu_i^k(\theta_{-i}'|h^{t-1}, \tilde{\theta}_i)\sigma_{-i}^k(a_{-i}^t|h^{t-1}, \theta_{-i}')}\right] \geq \chi\mu_i^k(\tilde{\theta}_{-i}|h^{t-1}, \tilde{\theta}_i).$$

There will be a contradiction as we take the limit $k \to \infty$ on both sides:

$$\mu_i(\tilde{\theta}_{-i}|h^t, \tilde{\theta}_i) = \lim_{k\to\infty}\mu_i^k(\tilde{\theta}_{-i}|h^t, \tilde{\theta}_i) \geq \lim_{k\to\infty}\chi\mu_i^k(\tilde{\theta}_{-i}|h^{t-1}, \tilde{\theta}_i) = \chi\mu_i(\tilde{\theta}_{-i}|h^{t-1}, \tilde{\theta}_i). \ ■$$

**Proof of Corollary 2**

We prove the statement by induction on $t$. For $t = 1$, by Proposition 3, $\mu_i(\theta_{-i}|h^1, \theta_i) \geq \chi\mu_i(\theta_{-i}|h_\emptyset, \theta_i) = \chi\mathcal{F}(\theta_{-i}|\theta_i)$. Next, suppose there is $t'$ such that the statement holds for all $1 \leq t \leq t' - 1$. At stage $t'$, by Proposition 3 and the induction hypothesis, we can find that

$$\mu_i(\theta_{-i}|h^{t'}, \theta_i) \geq \chi\mu_i(\theta_{-i}|h^{t'-1}, \theta_i) \geq \chi\left[\chi^{t'-1}\mathcal{F}(\theta_{-i}|\theta_i)\right] = \chi^{t'}\mathcal{F}(\theta_{-i}|\theta_i). \blacksquare$$

**Proof of Proposition 5**

Let the assessment $(\mu, \sigma)$ be a pooling $\chi$-CSE. We want to show that for any $\chi' \leq \chi$, the assessment $(\mu, \sigma)$ is also a $\chi'$-CSE. Consider any non-terminal history $h^{t-1}$, any player $i$, any $a_i^t \in A_i(h^{t-1})$ and any $\theta \in \Theta$. We can first observe that

$$\begin{aligned}
\bar{\sigma}_{-i}(a_{-i}^t|h^{t-1}, \theta_i) &= \sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\sigma_{-i}(a_{-i}^t|h^{t-1}, \theta'_{-i}) \\
&= \sigma_{-i}(a_{-i}^t|h^{t-1}, \theta_{-i})\left[\sum_{\theta'_{-i}} \mu_i(\theta'_{-i}|h^{t-1}, \theta_i)\right] = \sigma_{-i}(a_{-i}^t|h^{t-1}, \theta_{-i})
\end{aligned}$$

where the second equality holds because $\sigma$ is a pooling behavioral strategy profile, so $\sigma_{-i}$ is independent of other players' types. For this pooling $\chi$-CSE, let $G^\sigma$ be the set of on-path histories and $\tilde{G}^\sigma$ be the set of off-path histories. We can first show that for every $h \in G^\sigma$, $i \in N$ and $\theta \in \Theta$, $\mu_i(\theta_{-i}|h, \theta_i) = \mathcal{F}(\theta_{-i}|\theta_i)$.

This can be shown by induction on $t$. For $t = 1$, any $h^1 = (h_\emptyset, a^1)$ and any $\theta \in \Theta$, by Lemma 1, we can obtain that

$$\begin{aligned}
\mu_i(\theta_{-i}|h^1, \theta_i) &= \chi\mu_i(\theta_{-i}|h_\emptyset, \theta_i) + (1-\chi)\left[\frac{\mu_i(\theta_{-i}|h_\emptyset, \theta_i)\sigma_{-i}(a_{-i}^1|h_\emptyset, \theta_{-i})}{\bar{\sigma}_{-i}(a_{-i}^1|h_\emptyset, \theta_i)}\right] \\
&= \chi\mathcal{F}(\theta_{-i}|\theta_i) + (1-\chi)\mathcal{F}(\theta_{-i}|\theta_i)\underbrace{\left[\frac{\sigma_{-i}(a_{-i}^1|h_\emptyset, \theta_{-i})}{\bar{\sigma}_{-i}(a_{-i}^1|h_\emptyset, \theta_i)}\right]}_{=1} = \mathcal{F}(\theta_{-i}|\theta_i).
\end{aligned}$$

Now, suppose there is $t'$ such that the statement holds for $1 \leq t \leq t' - 1$. At stage $t'$ and $h^{t'} = (h^{t'-1}, a^{t'}) \in G^\sigma$, by Lemma 1 and the induction hypothesis, we can again obtain that the posterior belief is the prior belief

$$\mu_i(\theta_{-i}|h^{t'},\theta_i) = \chi\mu_i(\theta_{-i}|h^{t'-1},\theta_i) + (1-\chi)\left[\frac{\mu_i(\theta_{-i}|h^{t'-1},\theta_i)\sigma_{-i}(a^{t'}_{-i}|h^{t'-1},\theta_{-i})}{\bar{\sigma}_{-i}(a^{t'}_{-i}|h^{t'-1},\theta_i)}\right]$$
$$= \chi\mathcal{F}(\theta_{-i}|\theta_i) + (1-\chi)\mathcal{F}(\theta_{-i}|\theta_i)\underbrace{\left[\frac{\sigma_{-i}(a^{t'}_{-i}|h^{t'-1},\theta_{-i})}{\bar{\sigma}_{-i}(a^{t'}_{-i}|h^{t'-1},\theta_i)}\right]}_{=1} = \mathcal{F}(\theta_{-i}|\theta_i).$$

Therefore, we have shown that players will not update their beliefs at every on-path information set, so the belief system is independent of $\chi$. Finally, for any off-path history $h^t \in \tilde{G}^\sigma$, by Proposition 3, we can find that the belief system satisfies for any $\theta \in \Theta$,

$$\mu_i(\theta_{-i}|h^t,\theta_i) \geq \chi\mu_i(\theta_{-i}|h^{t-1},\theta_i) \geq \chi'\mu_i(\theta_{-i}|h^{t-1},\theta_i),$$

implying that when $\chi' \leq \chi$, $\mu$ will still satisfy the dampened updating property. Therefore, $(\mu,\sigma)$ remains a $\chi'$-CSE. ■

### A.2 Omitted Proofs of Section 2.4

**Pooling Equilibria in Signaling Games**

**Proof of Claim 1** First observe that after player 1 chooses $B$, it is strictly optimal for player 2 to choose $R$ for all beliefs $\mu_2(\theta_1|B)$, and after player 1 chooses $A$, it is optimal for player 2 to choose $L$ if and only if

$$2\mu_2(\theta_1|A) + [1-\mu_2(\theta_1|A)] \geq 4\mu_2(\theta_1|A) \iff \mu_2(\theta_1|A) \leq 1/3.$$

**Equilibrium 1.**

If both types of player 1 choose $A$, then $\mu_2(\theta_1|A) = 1/4$, so it is optimal for player 2 to choose $L$. Given $a(A) = L$ and $a(B) = R$, it is optimal for both types of player 1 to choose $A$ as $2 > 1$. Hence $m(\theta_1) = m(\theta_2) = A$, $a(A) = L$ and $a(B) = R$ is a pooling $\chi$-CSE for any $\chi \in [0,1]$.

**Equilibrium 2.**

In order to support $m(\theta_1) = m(\theta_2) = B$ to be an equilibrium, player 2 has to choose $R$ at the off-path information set $A$, which is optimal if and only if $\mu_2(\theta_1|A) \geq 1/3$. In addition, by Proposition 3, we know in a $\chi$-CSE, the belief system satisfies

$$\mu_2(\theta_2|A) \geq \frac{3}{4}\chi \iff \mu_2(\theta_1|A) \leq 1 - \frac{3}{4}\chi.$$

Therefore, the belief system has to satisfy that $\mu_2(\theta_1|A) \in \left[\frac{1}{3}, 1 - \frac{3}{4}\chi\right]$, which requires $\chi \leq 8/9$.

Finally, it is straightforward to verify that for any $\mu \in \left[\frac{1}{3}, 1 - \frac{3}{4}\chi\right]$, $\mu_2(\theta_1|A) = \mu$ satisfies $\chi$-consistency. Suppose type $\theta_1$ player 1 chooses $A$ with probability $p$ and type $\theta_2$ player 1 chooses $A$ with probability $q$ where $p, q \in (0, 1)$. Given this behavioral strategy profile for player 1, by Lemma 1, we have:

$$\mu_2(\theta_1|A) = \frac{1}{4}\chi + (1-\chi)\left[\frac{p}{p+3q}\right].$$

In other words, as long as $(p, q)$ satisfies

$$q = \left[\frac{4 - 4\mu - 3\chi}{12 - 3\chi}\right] p,$$

we can find that $\mu_2(\theta_1|A) = \mu$. Therefore, if $\{(p^k, q^k)\} \to (0, 0)$ such that

$$q^k = \left[\frac{4 - 4\mu - 3\chi}{12 - 3\chi}\right] p^k,$$

then $\mu_2^k(\theta_1|A) = \mu$ for all $k$. Hence, $\lim_{k\to\infty} \mu_2^k(\theta_1|A) = \mu$, suggesting that $\mu_2(\theta_1|A) = \mu$ is indeed $\chi$-consistent. This completes the proof. ■

**Proof of Proposition 6** Here we provide a characterization of $\chi$-CSE of Game 1 and Game 2. For the analysis of both games, we denote $\mu_I \equiv \mu_2(\theta_1|m = I)$ and $\mu_S \equiv \mu_2(\theta_1|m = S)$.

**Analysis of Game BH 3.**

At information set $S$, given $\mu_S$, the expected payoffs of $C$, $D$, $E$ are $90\mu_S$, $30 - 15\mu_S$ and 15, respectively. Therefore, for any $\mu_S$, $E$ is never a best response. Moreover, $C$ is the best response if and only if $90\mu_S \geq 30 - 15\mu_S$ or $\mu_S \geq 2/7$. Similarly, at information set $I$, given $\mu_I$, the expected payoffs of $C$, $D$, $E$ are 30, $45 - 45\mu_I$ and 15, respectively. Therefore, $E$ is strictly dominated, and $C$ is the best response if and only if $30 \geq 45 - 45\mu_I$ or $\mu_I \geq 1/3$. Now we consider four cases.

**Case 1 [$m(\theta_1) = I, m(\theta_2) = S$]:**

By Lemma 1, $\mu_I = 1 - \chi/2$ and $\mu_S = \chi/2$. Moreover, since $\mu_I = 1 - \chi/2 \geq 1/2$ for any $\chi$, player 2 will choose $C$ at information set $I$. To support this equilibrium, player 2 has to choose $C$ at information set $S$. In other words, $[(I, S); (C, C)]$ is separating $\chi$-CSE if and only if $\mu_S \geq 2/7$ or $\chi \geq 4/7$.

**Case 2** [$m(\theta_1) = S, m(\theta_2) = I$]**:**

By Lemma 1, $\mu_I = \chi/2$ and $\mu_S = 1 - \chi/2$. Because $\mu_S \geq 1 - \chi/2 \geq 1/2$, it is optimal for player 2 to choose $C$ at information set $S$. To support this as an equilibrium, player 2 has to choose $D$ at information set $I$. Yet, in this case, type $\theta_2$ player 1 will deviate to $S$. Therefore, this profile cannot be supported as an equilibrium.

**Case 3** [$m(\theta_1) = I, m(\theta_2) = I$]**:**

Since player 1 follows a pooling strategy, player 2 will not update his belief at information set $I$, i.e., $\mu_I = 1/2$. $\chi$-dampened updating property implies $\chi/2 \leq \mu_S \leq 1 - \chi/2$. Since $\mu_I > 1/3$, player 2 will choose $C$ at information set $I$. To support this profile to be an equilibrium, player 2 has to choose $D$ at information set $S$, and hence, it must be the case that $\mu_S \leq 2/7$. Coupled with the requirement from $\chi$-dampened updating, the off-path belief has to satisfy $\chi/2 \leq \mu_S \leq 2/7$. That is, $[(I, I); (C, D)]$ is pooling $\chi$-CSE if and only if $\chi/2 \leq 2/7$ or $\chi \leq 4/7$.

**Case 4** [$m(\theta_1) = S, m(\theta_2) = S$]**:**

Similar to the previous case, since player 1 follows a pooling strategy, player 2 will not update his belief at information set $S$, i.e., $\mu_S = 1/2$. Also, the $\chi$-dampened updating property suggests $\chi/2 \leq \mu_I \leq 1 - \chi/2$. Because $\mu_S > 2/7$, it is optimal for player 2 to choose $C$ at information set $S$. To support this as an equilibrium, player 2 has to choose $D$ at information set $I$. Therefore, it must be that $\mu_I \leq 1/3$. Combined with the requirement of $\chi$-dampened updating, the off-path belief has to satisfy $\chi/2 \leq \mu_I \leq 1/3$. As a result, $[(S, S); (D, C)]$ is a pooling $\chi$-CSE if and only if $\chi \leq 2/3$.

**Analysis of Game BH 4.**

At information set $I$, given $\mu_I$, the expected payoffs of $C$, $D$, $E$ are 30, $45 - 45\mu_I$ and $35\mu_I$. Hence, $D$ is the best response if and only if $\mu_I \leq 1/3$ while $E$ is the best response if $\mu_I \geq 6/7$. For $1/3 \leq \mu_I \leq 6/7$, $C$ is the best response. On the other hand, since player 2's payoffs at information set $S$ are the same as in Game 1, player 2 will adopt the same decision rule—player 2 will choose $C$ if and only if $\mu_S \geq 2/7$, and choose $D$ if and only if $\mu_S \leq 2/7$. Now, we consider the following four cases.

**Case 1** [$m(\theta_1) = I, m(\theta_2) = S$]**:**

In this case, by Lemma 1, $\mu_I = 1 - \chi/2$ and $\mu_S = \chi/2$. To support this profile to be an equilibrium, player 2 has to choose $E$ and $C$ at information set $I$ and $S$, respectively. To make it profitable for player 2 to choose $E$ at information set $I$, it must be that:

$$\mu_I = 1 - \chi/2 \geq 6/7 \iff \chi \leq 2/7.$$

On the other hand, player 2 will choose $C$ at information set $S$ if and only if $\chi/2 \geq 2/7$ or $\chi \geq 4/7$, which is not compatible with the previous inequality. Therefore, this profile cannot be supported as an equilibrium.

**Case 2 [$m(\theta_1) = S, m(\theta_2) = I$]:**

In this case, by Lemma 1, $\mu_I = \chi/2$ and $\mu_S = 1 - \chi/2$. To support this as an equilibrium, player 2 has to choose $D$ at both information sets. Yet, $\mu_S = 1 - \chi/2 > 2/7$, implying that it is not a best reply for player 2 to choose $D$ at information set $S$. Hence this profile also cannot be supported as an equilibrium.

**Case 3 [$m(\theta_1) = I, m(\theta_2) = I$]:**

Since player 1 follows a pooling strategy, player 2 will not update his belief at information set $I$, i.e., $\mu_I = 1/2$. The $\chi$-dampened updating property implies $\chi/2 \leq \mu_S \leq 1 - \chi/2$. Because $1/3 < \mu_I = 1/2 < 6/7$, player 2 will choose $C$ at information set $I$. To support this profile as an equilibrium, player 2 has to choose $D$ at information set $S$, and hence, it must be the case that $\mu_S \leq 2/7$. Coupled with the requirement of $\chi$-dampened updating, the off-path belief has to satisfy $\chi/2 \leq \mu_S \leq 2/7$. That is, $[(I, I); (C, D)]$ is pooling $\chi$-CSE if and only if $\chi/2 \leq 2/7$ or $\chi \leq 4/7$.

**Case 4 [$m(\theta_1) = S, m(\theta_2) = S$]:**

Similar to the previous case, since player 1 follows a pooling strategy, player 2 will not update his belief at information set $S$, i.e., $\mu_S = 1/2$. Also, the $\chi$-dampened updating property implies $\chi/2 \leq \mu_I \leq 1 - \chi/2$. Because $\mu_S > 2/7$, it is optimal for player 2 to choose $C$ at information set $S$. To support this as an equilibrium, player 2 can choose either $C$ or $D$ at information set $I$.

**Case 4.1:** To make it a best reply for player 2 to choose $D$ at information set $I$, it must be that $\mu_I \leq 1/3$. Combined with the requirement from $\chi$-dampened updating, the off-path belief has to satisfy $\chi/2 \leq \mu_I \leq 1/3$. As a result, $[(S, S); (D, C)]$ is a pooling $\chi$-CSE if and only if $\chi \leq 2/3$.

**Case 4.2:** To make it a best reply for player 2 to choose $C$ at information set $I$, it must be that $1/3 \leq \mu_I \leq 6/7$. Combined with the requirement from $\chi$-dampened updating, the off-path belief has to satisfy

$$\max\left\{\frac{1}{2}\chi,\ \frac{1}{3}\right\} \leq \mu_I \leq \min\left\{\frac{6}{7}, 1-\frac{1}{2}\chi\right\}.$$

For any $\chi \in [0,1]$, one can find $\mu_I$ that satisfies both inequalities. Hence $[(S,S);(C,C)]$ is a pooling $\chi$-CSE for any $\chi$.

This completes the analysis of Game BH 3 and Game BH 4. ■

**A Public Goods Game with Communication**

**Proof of Proposition 7** To prove this set of cost cutoffs form a $\chi$-CSE, we need to show that there is no profitable deviation for any type at any subgame. First, at the second stage where there are exactly $0 \leq k \leq N-1$ players sending 1 in the first stage, since no players will contribute, setting $C_k^\chi = 0$ is indeed a best response. At the subgame where all $N$ players send 1 in the first stage, we use $\mu_i^\chi(c_{-i}|N)$ to denote player $i$'s cursed belief density. By Lemma 1, the cursed belief about all other players having a cost lower than $c$ is simply:

$$\begin{aligned}
F^\chi(c) &\equiv \int_{\{c_j \leq c,\, \forall j \neq i\}} \mu_i^\chi(c'_{-i}|N)dc'_{-i} \\
&= \begin{cases} \chi\,(c/K)^{N-1} + (1-\chi)\left(c/C_c^\chi\right)^{N-1} & \text{if } c \leq C_c^\chi \\ 1-\chi+\chi\left(c/C_c^\chi\right)^{N-1} & \text{if } c > C_c^\chi, \end{cases}
\end{aligned}$$

and $C_N^\chi$ is the solution of the fixed point problem of $C_N^\chi = F^\chi(C_N^\chi)$.

Moreover, in equilibrium, $C_c^\chi$ type of players would be indifferent between sending 1 and 0 in the communication stage. Thus, given $C_N^\chi$, $C_c^\chi$ is the solution of the following equation

$$0 = \left(\frac{C_c^\chi}{K}\right)^{N-1}\left[-C_c^\chi + F^\chi(C_N^\chi)\right].$$

As a result, we obtain that in equilibrium, $C_c^\chi = C_N^\chi = F^\chi(C_N^\chi) \leq 1$ and denote this cost cutoff by $C^*(N,K,\chi)$. Substituting it into $F^\chi(c)$, gives:

$$C^*(N,K,\chi) - \chi\left[\frac{C^*(N,K,\chi)}{K}\right]^{N-1} = 1-\chi.$$

In the following, we show that for any $N \geq 2$ and $\chi$, the cutoff $C^*(N,K,\chi)$ is unique.

**Case 1:** When $N = 2$, the cutoff $C^*(2, K, \chi)$ is the unique solution of the linear equation

$$C^*(2, K, \chi) - \chi \left[ \frac{C^*(2, K, \chi)}{K} \right] = 1 - \chi \iff C^*(2, K, \chi) = \frac{K - K\chi}{K - \chi}.$$

**Case 2:** For $N \geq 3$, we define the function $h(y) : [0, 1] \to \mathbb{R}$ where

$$h(y) = y - \chi \left( \frac{y}{K} \right)^{N-1} - (1 - \chi).$$

It suffices to show that $h(y)$ has a unique root in $[0, 1]$. When $\chi = 0$, $h(y) = y - 1$ which has a unique root at $y = 1$. In the following, we will focus on the case where $\chi > 0$. Since $h(y)$ is continuous, $h(0) = -(1 - \chi) < 0$ and $h(1) = \chi \left[ 1 - (1/K)^{N-1} \right] > 0$, there exists a root $y^* \in (0, 1)$ by the intermediate value theorem. Moreover, as we take the second derivative, we can find that for any $y \in (0, 1)$,

$$h''(y) = - \left( \frac{\chi}{K^{N-1}} \right) (N - 1)(N - 2) y^{N-3} < 0,$$

implying that $h(y)$ is strictly concave in $[0, 1]$. Furthermore, $h(0) < 0$ and $h(1) > 0$, so the root is unique, as illustrated in the left panel of Figure 3. This completes the proof. ■

**Proof of Corollary 3** By Proposition 7, we know the cutoff $C^*(N, K, \chi) \leq 1$ and it satisfies

$$C^*(N, K, \chi) - \chi \left[ \frac{C^*(N, K, \chi)}{K} \right]^{N-1} = 1 - \chi.$$

Therefore, when $\chi = 0$, the condition becomes $C^*(N, K, 0) = 1$. In addition, when $\chi = 1$, the condition becomes

$$C^*(N, K, 1) - \left[ \frac{C^*(N, K, 1)}{K} \right]^{N-1} = 0,$$

implying $C^*(N, K, 1) = 0$.

For $\chi \in (0, 1)$, to prove $C^*(N, K, \chi)$ is strictly decreasing in $N$, $K$ and $\chi$, we consider a function $g(y; N, K, \chi) : (0, 1) \to \mathbb{R}$ where $g(y; N, K, \chi) = y - \chi [y/K]^{N-1}$. For any $y \in (0, 1)$ and fix any $K$ and $\chi$, we can observe that when $N \geq 2$,

$$g(y; N + 1, K) - g(y; N, K) = -\chi \left[ \frac{y}{K} \right]^N + \chi \left[ \frac{y}{K} \right]^{N-1} > 0,$$

so $g(\cdot; N, K, \chi)$ is strictly increasing in $N$. Therefore, the cutoff $C^*(N, K, \chi)$ is strictly decreasing in $N$. Similarly, for any $y \in (0,1)$ and fix any $N$ and $\chi$, observe that when $K > 1$,

$$\frac{\partial g}{\partial K} = \chi(N-1)\left(\frac{y^{N-1}}{K^N}\right) > 0,$$

which implies that cutoff $C^*(N, K, \chi)$ is also strictly decreasing in $K$. For the comparative statics of $\chi$, we can rearrange the equilibrium condition where

$$\frac{1 - C^*(N, K, \chi)}{\chi} = 1 - \left[\frac{C^*(N, K, \chi)}{K}\right]^{N-1}.$$

Since LHS is strictly decreasing in $\chi$, the equilibrium cutoff is also strictly decreasing in $\chi$. Finally, taking the limit on both sides of the equilibrium condition, we obtain:

$$\lim_{N\to\infty} C^*(N, K, \chi) = \lim_{K\to\infty} C^*(N, K, \chi) = 1 - \chi.$$

This completes the proof. ■

**The Centipede Game with Altruistic Types**

**Proof of Claim 2** By backward induction, we know selfish player two will choose $T4$ for sure. Given that player two will choose $T4$ at stage four, it is optimal for selfish player one to choose $T3$. Now, suppose selfish player one will choose $P1$ with probability $q_1$ and player two will choose $P2$ with probability $q_2$. Given this behavioral strategy profile, player two's belief about the other player being altruistic at stage two is:

$$\mu = \frac{\alpha}{\alpha + (1-\alpha)q_1}.$$

In this case, it is optimal for selfish player two to pass if and only if

$$32\mu + 4(1-\mu) \geq 8 \iff \mu \geq \frac{1}{7}.$$

At the equilibrium, selfish player two is indifferent between $T2$ and $P2$. If not, say $32\mu + 4(1-\mu) > 8$, player two will choose $P2$. Given that player two will choose $P2$, it is optimal for selfish player one to choose $P1$, which makes $\mu = \alpha$ and $\alpha > 1/7$. However, we know $\alpha \leq 1/7$ which yields a contradiction. On the other hand, if $32\mu + 4(1-\mu) < 8$, then it is optimal for player two to choose $T2$ at stage two. As a result, selfish player one would choose $T1$ at stage one, causing $\mu = 1$. In this case, player two would deviate to choose $P2$, which again yields a contradiction. To summarize, in equilibrium, player two has to be indifferent between $T2$ and $P2$, i.e., $\mu = 1/7$. As we rearrange the equality, we can obtain that

$$\frac{\alpha}{\alpha + (1-\alpha)q_1^*} = \frac{1}{7} \iff q_1^* = \frac{6\alpha}{1-\alpha}.$$

Finally, since the equilibrium requires selfish player one to mix at stage one, selfish player one has to be indifferent between $P1$ and $T1$. Therefore,

$$4 = 16q_2^* + 2(1 - q_2^*) \iff q_2^* = \frac{1}{7}.$$

This completes the proof. ■

**Proof of Proposition 8** By backward induction, we know selfish player two will choose $T4$ for sure. Given this, it is optimal for selfish player one to choose $T3$. Now, suppose selfish player one will choose $P1$ with probability $q_1$ and player two will choose $P2$ with probability $q_2$. Given this behavioral strategy profile, by Lemma 1, player two's *cursed* belief about the other player being altruistic at stage 2 is:

$$\mu^\chi = \chi\alpha + (1-\chi)\left[\frac{\alpha}{\alpha + (1-\alpha)q_1}\right].$$

In this case, it is optimal for player two to pass if and only if

$$32\mu^\chi + 4(1-\mu^\chi) \geq 8 \iff \mu^\chi \geq \frac{1}{7}.$$

We can first show that in equilibrium, it must be that $\mu^\chi \leq 1/7$. If not, then it is strictly optimal for player two to choose $P2$. Therefore, it is optimal for selfish player one to choose $P1$ and hence $\mu^\chi = \alpha \leq 1/7$, which yields a contradiction. In the following, we separate the discussion into two cases.

**Case 1:** $\chi \leq \frac{6}{7(1-\alpha)}$

In this case, we argue that player two is indifferent between $P2$ and $T2$. If not, then $32\mu^\chi + 4(1-\mu^\chi) < 8$ and it is strictly optimal for player two to choose $T2$. This would cause selfish player one to choose $T1$ and hence $\mu^\chi = 1 - (1-\alpha)\chi$. This yields a contradiction because

$$\mu^\chi = 1 - (1-\alpha)\chi < \frac{1}{7} \iff \chi > \frac{6}{7(1-\alpha)}.$$

Therefore, in this case, player two is indifferent between $T2$ and $P2$ and thus,

$$\begin{aligned}
\mu^\chi = \frac{1}{7} &\iff \chi\alpha + (1-\chi)\left[\frac{\alpha}{\alpha + (1-\alpha)q_1^\chi}\right] = \frac{1}{7} \\
&\iff \chi + \frac{1-\chi}{\alpha + (1-\alpha)q_1^\chi} = \frac{1}{7\alpha} \\
&\iff \alpha + (1-\alpha)q_1^\chi = (1-\chi)\Big/\left[\frac{1}{7\alpha} - \chi\right] \\
&\iff q_1^\chi = \left[\frac{7\alpha - 7\alpha\chi}{1 - 7\alpha\chi} - \alpha\right]\Big/(1-\alpha).
\end{aligned}$$

Since the equilibrium requires selfish player one to mix at stage 1, selfish player one has to be indifferent between $P1$ and $T1$. Therefore,

$$4 = 16q_2^\chi + 2(1 - q_2^\chi) \iff q_2^\chi = \frac{1}{7}.$$

**Case 2:** $\chi > \frac{6}{7(1-\alpha)}$

In this case, we know for any $q_1^\chi \in [0, 1]$,

$$\mu^\chi = \chi\alpha + (1 - \chi)\left[\frac{\alpha}{\alpha + (1 - \alpha)q_1^\chi}\right] \leq 1 - (1 - \alpha)\chi < \frac{1}{7},$$

implying that it is strictly optimal for player two to choose $T2$, and hence it is strictly optimal for selfish player one to choose $T1$ at stage 1. This completes the proof. ■

**Sequential Voting over Binary Agendas**

**Proof of Proposition 9** If $a^1(\theta_1) = b$ and all other types of voters as well as type $\theta_1$ at stage 2 vote sincerely, voter $i$'s $\chi$-cursed belief in the second stage upon observing $a_{-i}^1 = (a, b)$ is

$$\mu_i^\chi(\theta_{-i}|a_{-i}^1 = (a, b)) = \begin{cases} p_1p_3\chi + \frac{p_1}{p_1+p_2}(1 - \chi) & \text{if } \theta_{-i} = (\theta_3, \theta_1) \\ p_2p_3\chi + \frac{p_2}{p_1+p_2}(1 - \chi) & \text{if } \theta_{-i} = (\theta_3, \theta_2) \\ p_kp_l\chi & \text{otherwise.} \end{cases}$$

As mentioned in Section 4.4, a voter would act as if he perceives the other voters' (behavioral) strategies correctly in the last stage. However, misunderstanding the link between the other voters' types and actions would distort a voter's belief updating process. In other words, a voter would perceive the strategies correctly but form beliefs incorrectly. As a result, the continuation value of the $a$ vs $c$ subgame to a type $\theta_1$ voter is simply the voter's $\chi$-cursed belief, conditional on being pivotal, about there being at least one type $\theta_1$ voter among his opponents. Similarly, the continuation value of the $b$ vs $c$ subgame is equal to the voter's conditional $\chi$-cursed belief about there being at least one type $\theta_1$ or $\theta_2$ voter among his opponents multiplied by $v$. Therefore, the continuation values to a type $\theta_1$ voter in the two possible subgames of the second stage are (let $\tilde{p}_2 \equiv \frac{p_1}{p_1+p_2}$):

$$\begin{aligned} a \text{ vs } c: &\quad \chi\left(1 - (1 - p_1)^2\right) + (1 - \chi)\tilde{p}_2 \\ b \text{ vs } c: &\quad \left(1 - p_3^2\chi\right)v \end{aligned}$$

It is thus optimal for a type $\theta_1$ voter to vote for $b$ in the first stage if

$$\chi\left(1-(1-p_1)^2\right)+(1-\chi)\tilde{p}_2 \leq \left(1-p_3^2\chi\right)v$$
$$\Longleftrightarrow [2p_1-p_1^2-\tilde{p}_2+p_3^2v]\chi \leq v-\tilde{p}_2 \tag{A.1}$$

Notice that the statement would automatically hold when $\chi=0$. In the following, we want to show that given $v$ and $p$, if condition (A.1) holds for some $\chi \in (0,1]$, then it will hold for all $\chi' \leq \chi$. As $\chi > 0$, we can rewrite condition (A.1) as

$$2p_1-p_1^2-\tilde{p}_2+p_3^2v \leq \frac{v-\tilde{p}_2}{\chi}. \tag{2'}$$

**Case 1:** $v-\tilde{p}_2 < 0$.

In this case, we want to show that voting $b$ in the first stage is never optimal for type $\theta_1$ voter. That is, we want to show condition (2') never holds for $v < \tilde{p}_2$. To see this, we can first observe that the RHS is strictly increasing in $\chi$. Therefore, it suffices to show

$$2p_1-p_1^2-\tilde{p}_2+p_3^2v > v-\tilde{p}_2.$$

This is true because

$$\begin{aligned}2p_1-p_1^2-\tilde{p}_2+p_3^2v-(v-\tilde{p}_2) &= 2p_1-p_1^2-(1-p_3^2)v\\ &> 2p_1-p_1^2-(1+p_3)p_1 = p_1p_2 \geq 0\end{aligned}$$

where the second inequality holds as $v < \frac{p_1}{p_1+p_2}$.

**Case 2:** $v-\tilde{p}_2 \geq 0$.

Since the RHS of condition (2') is greater or equal to 0, it will weakly increase as $\chi$ decreases. Thus, if condition (2') holds for some $\chi \in (0,1]$, it will also hold for all $\chi' \leq \chi$. This completes the proof. ■

**Proof of Proposition 10** Assuming that all voters vote sincerely in both stages, voter $i$'s $\chi$-cursed belief in the second stage upon observing $a_{-i}^1=(a,b)$ is

$$\mu_i^\chi(\theta_{-i}|a_{-i}^1=(a,b)) = \begin{cases} p_1p_2\chi+\frac{p_1}{p_1+p_3}(1-\chi) & \text{if } \theta_{-i}=(\theta_1,\theta_2)\\ p_2p_3\chi+\frac{p_3}{p_1+p_3}(1-\chi) & \text{if } \theta_{-i}=(\theta_3,\theta_2)\\ p_kp_l\chi & \text{otherwise.}\end{cases}$$

Similar to the proof of Proposition 9, the continuation values to a type $\theta_1$ voter in the two possible subgames of the second stage are (let $\tilde{p}_3 \equiv \frac{p_1}{p_1+p_3}$):

$$\begin{aligned} a \text{ vs } c: &\quad \chi\left(1-(1-p_1)^2\right)+(1-\chi)\tilde{p}_3 \\ b \text{ vs } c: &\quad \left(1-p_3^2\chi\right)v \end{aligned}$$

Thus, it is optimal for a type $\theta_1$ voter to vote for $a$ in the first stage if

$$\begin{aligned} &\chi\left(1-(1-p_1)^2\right)+(1-\chi)\tilde{p}_3 \geq \left(1-p_3^2\chi\right)v \\ \Longleftrightarrow\ &\chi\left(2p_1-p_1^2-\tilde{p}_3+p_3^2v\right) \geq v-\tilde{p}_3. \end{aligned} \tag{A.2}$$

**Case 1:** $v-\tilde{p}_3>0$.

In this case, we want to show that given $p$ and $v$, there exists $\tilde{\chi}$ such that condition (A.2) holds if and only if $\chi \geq \tilde{\chi}$. Let $\tau \equiv 2p_1-p_1^2-\tilde{p}_3+p_3^2v$. If $\tau>0$, then condition (A.2) holds if and only if $\chi \geq \tilde{\chi} \equiv \frac{v-\tilde{p}_3}{\tau}$. On the other hand, if $\tau \leq 0$, condition (A.2) will not hold for all $\chi \in [0,1]$ and hence we can set $\tilde{\chi}=2$.

**Case 2:** $v-\tilde{p}_3 \leq 0$.

In this case, we want to show that given $p$ and $v$, there exists $\tilde{\chi}$ such that condition (A.2) holds if and only if $\chi \leq \tilde{\chi}$. If $\tau<0$, then condition (A.2) holds if and only if $\chi \leq \frac{v-\tilde{p}_3}{\tau}$ where the RHS is greater or equal to 0. On the other hand, if $\tau \geq 0$, then condition (A.2) will hold for any $\chi \in [0,1]$ and hence we can again set $\tilde{\chi}=2$. This completes the proof. ■

**The Dirty Faces Game**

**Proof of Proposition 11** When observing a clean face, a player will know that he has a dirty face immediately. Therefore, choosing 1 (i.e., choosing $D$ at stage 1) when observing a clean face is a strictly dominant strategy. In other words, for any $\chi \in [0,1]$, $\hat{\sigma}^{\chi}(O)=1$.

The analysis of the case where the player observes a dirty face is separated into two cases.

**Case 1:** $\chi>\bar{\alpha}$

In this case, we show that $\hat{\sigma}^{\chi}(X)=T+1$ is the only $\chi$-CE. If not, suppose $\hat{\sigma}^{\chi}(X)=t$ where $t \leq T$ can be supported as a $\chi$-CE. We can first notice that $\hat{\sigma}^{\chi}(X)=1$ cannot be supported as a $\chi$-CE because it is strictly dominated to choose 1 when observing a dirty face. For $2 \le t \le T$, given the other player $-i$ chooses $\hat{\sigma}^{\chi}(X) = t$, we can find player $-i$'s *average strategy* is

$$\bar{\sigma}_{-i}(j) = \begin{cases} 1-p & \text{if} \quad j = 1 \\ p & \text{if} \quad j = t \\ 0 & \text{if} \quad j \neq 1, t. \end{cases}$$

Therefore, the other player $-i$'s $\chi$-cursed strategy is:

$$\sigma^{\chi}_{-i}(j|x_i = O) = \begin{cases} \chi(1-p) + (1-\chi) & \text{if} \quad j = 1 \\ \chi p & \text{if} \quad j = t \\ 0 & \text{if} \quad j \neq 1, t, \end{cases} \qquad \text{and}$$

$$\sigma^{\chi}_{-i}(j|x_i = X) = \begin{cases} \chi(1-p) & \text{if} \quad j = 1 \\ \chi p + (1-\chi) & \text{if} \quad j = t \\ 0 & \text{if} \quad j \neq 1, t. \end{cases}$$

In this case, given (player $i$ perceives that) player $-i$ chooses the $\chi$-cursed strategy, player $i$'s expected payoff to choose $2 \le j \le t$ when observing a dirty face is:

$$(1-p)\left[-\delta^{j-1}\chi p\right] + p\left\{\delta^{j-1}\alpha\left[\chi p + (1-\chi)\right]\right\} = p\delta^{j-1}\underbrace{\left[\alpha - \chi(1+\alpha)(1-p)\right]}_{<0 \iff \chi > \bar{\alpha}} < 0.$$

Hence, given the other player chooses $t$ when observing a dirty face, it is strictly dominated to choose any $j \le t$. Therefore, the only $\chi$-CE is $\hat{\sigma}^{\chi}(X) = T + 1$.

**Case 2:** $\chi < \bar{\alpha}$

In this case, we want to show that $\hat{\sigma}^{\chi}(X) = 2$ is the only $\chi$-CE. If not, suppose $\hat{\sigma}(X) = t$ for some $t \ge 3$ can be supported as a $\chi$-CE. We can again notice that since when observing a dirty face, it is strictly dominated to choose 1, 1 is never a best response. Given player $-i$ chooses $\hat{\sigma}^{\chi}(X) = t$, by the same calculation as in **Case 1**, the expected payoff to choose $2 \le j \le t$ is:

$$p\delta^{j-1}\underbrace{\left[\alpha - \chi(1+\alpha)(1-p)\right]}_{>0 \iff \chi < \bar{\alpha}} > 0,$$

which is decreasing in $j$. Therefore, the best response to $\hat{\sigma}^{\chi}(X) = t$ is to choose 2 when observing a dirty face. As a result, the only $\chi$-CE in this case is $\hat{\sigma}^{\chi}(X) = 2$. This completes the proof. ■

**Proof of Proposition 12** When observing a clean face, the player would know that his face is dirty. Thus, choosing $D$ at stage 1 is a strictly dominant strategy, and $\tilde{\sigma}^{\chi}(O) = 1$ for all $\chi \in [0, 1]$. On the other hand, the analysis for the case where the player observes a dirty face consists of several steps.

**Step 1:** Assume that both players choosing $D$ at some stage $\bar{t}$. We claim that at stage $t \leq \bar{t}$, the cursed belief $\mu^{\chi}(X|t, X) = 1 - (1-p)\chi^{t-1}$. We can prove this by induction on $t$. At stage $t = 1$, the belief about having a dirty face is simply the prior belief $p$. Hence this establishes the base case. Now suppose the statement holds for any stage $1 \leq t \leq t'$ (and $t' < \bar{t}$). At stage $t' + 1$, by Lemma 1,

$$\begin{aligned}\mu^{\chi}(X|t'+1, X) &= \chi\mu^{\chi}(X|t', X) + (1-\chi)\\ &= \chi\left[1 - (1-p)\chi^{t'-1}\right] + (1-\chi)\\ &= 1 - (1-p)\chi^{t'}\end{aligned}$$

where the second equality holds by the induction hypothesis. This proves the claim.

**Step 2:** Given the cursed belief computed in the previous step, the expected payoff to choose $D$ at stage $t$ is:

$$\begin{aligned}\mu^{\chi}(X|t, X)\alpha - [1 - \mu^{\chi}(X|t, X)] &= \left[1 - (1-p)\chi^{t-1}\right]\alpha - \left[(1-p)\chi^{t-1}\right]\\ &= \alpha - (1-p)(1+\alpha)\chi^{t-1},\end{aligned}$$

which is increasing in $t$. Notice that at the first stage, the expected payoff is $\alpha - (1-p)(1+\alpha) < 0$ by Assumption (1), so choosing $U$ at stage 1 is strictly dominated. Furthermore, the player would choose $U$ at every stage when observing a dirty face if and only if

$$\begin{aligned}\mu^{\chi}(X|T, X)\alpha - [1 - \mu^{\chi}(X|T, X)] \leq 0 &\iff \alpha - (1-p)(1+\alpha)\chi^{T-1} \leq 0\\ &\iff \chi \geq \bar{\alpha}^{\frac{1}{T+1}}.\end{aligned}$$

As a result, both players choosing $\tilde{\sigma}^{\chi}(X) = T+1$ is a $\chi$-CSE if and only if $\chi \geq \bar{\alpha}^{\frac{1}{T+1}}$.

**Step 3:** In this step, we show both players choosing $\tilde{\sigma}^{\chi}(X) = 2$ is a $\chi$-CSE if and only if $\chi \leq \bar{\alpha}$. We can notice that given the other player chooses $D$ at stage 2, the player would know stage 2 would be the last stage regardless of his face type. Therefore, it is optimal to choose $D$ at stage 2 as long as the expected payoff of $D$

at stage 2 is positive. Consequently, both players choosing $\tilde{\sigma}^{\chi}(X) = 2$ is a $\chi$-CSE if and only if

$$\begin{aligned}\mu^{\chi}(X|2,X)\alpha - [1-\mu^{\chi}(X|2,X)] \geq 0 &\iff \alpha - (1-p)(1+\alpha)\chi \\ &\iff \chi \leq \bar{\alpha}.\end{aligned}$$

**Step 4:** Given the other player chooses $\tilde{\sigma}^{\chi}(X) > t$, as the game reaches stage $t$, the belief about the other player choosing $U$ at stage $t$ is:

$$\begin{aligned}&\underbrace{\mu^{\chi}(X|t,X)}_{\text{prob. of dirty}} \; [\chi\mu^{\chi}(X|t,X) + (1-\chi)] \\ &\qquad + \underbrace{[1-\mu^{\chi}(X|t,X)]}_{\text{prob. of clean}} [\chi\mu^{\chi}(X|t,X)] = \mu^{\chi}(X|t,X).\end{aligned}$$

Furthermore, we denote the expected payoff of choosing $D$ at stage $t$ as

$$\mathbb{E}\left[u^{\chi}(D|t,X)\right] \equiv \mu^{\chi}(X|t,X)\alpha - (1-\mu^{\chi}(X|t,X)).$$

In the following, we claim that for any stage $2 \leq t \leq T-2$, given the other player will stop at some stage later than stage $t+2$ or never stop, if it is optimal to choose $U$ at stage $t+1$, then it is also optimal for you to choose $U$ at stage $t$. That is,

$$\begin{aligned}&\mathbb{E}\left[u^{\chi}(D|t+1,X)\right] < \delta\mu^{\chi}(X|t+1,X)\mathbb{E}\left[u^{\chi}(D|t+2,X)\right] \\ \Longrightarrow\ &\mathbb{E}\left[u^{\chi}(D|t,X)\right] < \delta\mu^{\chi}(X|t,X)\mathbb{E}\left[u^{\chi}(D|t+1,X)\right].\end{aligned}$$

To prove this claim, first observe that

$$\begin{aligned}&\mathbb{E}\left[u^{\chi}(D|t+1,X)\right] < \delta\mu^{\chi}(X|t+1,X)\mathbb{E}\left[u^{\chi}(D|t+2,X)\right] \\ \iff\ &(1+\alpha)\mu^{\chi}(X|t+1,X) - 1 < \delta\mu^{\chi}(X|t+1,X)\left[(1+\alpha)\mu^{\chi}(X|t+2,X) - 1\right].\end{aligned}$$

After rearrangement, the inequality is equivalent to

$$\delta\chi\left[\mu^{\chi}(X|t+1,X)\right]^2 + \left[\delta(1-\chi) - \frac{\delta}{1+\alpha} - 1\right]\mu^{\chi}(X|t+1,X) + \frac{1}{1+\alpha} > 0.$$

Consider a function $F : [0,1] \to \mathbb{R}$ where

$$F(y) = \delta\chi y^2 + \left[\delta(1-\chi) - \frac{\delta}{1+\alpha} - 1\right]y + \frac{1}{1+\alpha}.$$

Since $\mu^{\chi}(X|j,X) = 1-(1-p)\chi^{j-1}$ is increasing in $j$, it suffices to complete the proof of the claim by showing there exists a unique $y^* \in (0,1)$ such that $F$ is single-crossing on $[0,1]$ where $F(y^*) = 0$, $F(y) < 0$ for all $y > y^*$, and $F(y) > 0$ for all $y < y^*$. Because $F$ is continuous and

- $F(0) = \frac{1}{1+\alpha} > 0$,
- $F(1) = \delta\chi + \left[\delta(1-\chi) - \frac{\delta}{1+\alpha} - 1\right] + \frac{1}{1+\alpha} = -\frac{\alpha(1-\delta)}{1+\alpha} < 0$.

By intermediate value theorem, there exists a $y^* \in (0,1)$ such that $F(y^*) = 0$. Moreover, $y^*$ is the unique root of $F$ on $[0,1]$ because $F$ is a strictly convex parabola and $F(1) < 0$. This establishes the claim.

**Step 5:** For any $3 \leq t \leq T$, in this step, we find the conditions to support both players choosing $\tilde{\sigma}^\chi(X) = t$ as a $\chi$-CSE. We can first notice that both players choosing $\tilde{\sigma}^\chi(X) = t$ is a $\chi$-CSE if and only if

1. $\mathbb{E}\left[u^\chi(D|t,X)\right] \geq 0$
2. $\mathbb{E}\left[u^\chi(D|t-1,X)\right] \leq \delta\mu^\chi(X|t-1,X)\mathbb{E}\left[u^\chi(D|t,X)\right]$.

Condition 1 is necessary because if it fails, then it is better for the player to choose $U$ at stage $t$ and get at least 0. Condition 2 is also necessary because if the condition doesn't hold, it would be profitable for the player to choose $D$ before stage $t$. Furthermore, these two conditions are jointly sufficient to support $\tilde{\sigma}^\chi(X) = t$ as a $\chi$-CSE by the same argument as **step 3**. From condition 1, we can obtain that

$$\begin{aligned}\mathbb{E}\left[u^\chi(D|t,X)\right] \geq 0 &\iff (1+\alpha)\mu^\chi(X|t,X) - 1 \geq 0 \\ &\iff 1-(1-p)\chi^{t-1} \geq \frac{1}{1+\alpha} \iff \chi \leq \bar{\alpha}^{\frac{1}{t-1}}.\end{aligned}$$

In addition, by the calculation of **step 4**, we know

$$\mathbb{E}\left[u^\chi(D|t-1,X)\right] \leq \delta\mu^\chi(X|t-1,X)\mathbb{E}\left[u^\chi(D|t,X)\right] \iff F\left(\mu^\chi(X|t-1,X)\right) \geq 0,$$

which is equivalent to

$$\begin{aligned}\mu^\chi(X|t-1,X) &\leq \frac{\left[1+\frac{\delta}{1+\alpha} - \delta(1-\chi)\right] - \sqrt{\left[1+\frac{\delta}{1+\alpha} - \delta(1-\chi)\right]^2 - 4\delta\chi\left(\frac{1}{1+\alpha}\right)}}{2\delta\chi} \\ &= \frac{[(1+\alpha)(1+\delta\chi) - \alpha\delta] - \sqrt{[(1+\alpha)(1+\delta\chi) - \alpha\delta]^2 - 4\delta\chi(1+\alpha)}}{2\delta\chi(1+\alpha)} \equiv \kappa(\chi).\end{aligned}$$

Therefore, condition 2 holds if and only if

$$1-(1-p)\chi^{t-2} \leq \kappa(\chi) \iff \chi \geq \left(\frac{1-\kappa(\chi)}{1-p}\right)^{\frac{1}{t-2}}.$$

In summary, both players choosing $\tilde{\sigma}^{\chi}(X) = t$ is a $\chi$-CSE if and only if

$$\left(\frac{1-\kappa(\chi)}{1-p}\right)^{\frac{1}{t-2}} \leq \chi \leq \bar{\alpha}^{\frac{1}{t-1}}.$$

This completes the proof. ■

*Appendix B*

# PROOFS FOR CHAPTER 3

## Proof of Theorem 1

*Proof.* Suppose (towards contradiction) that both BUE and CBE do not exist. This implies both inequality (3.1) and inequality (3.2) do not hold, i.e.,

$$\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2(2\theta-1)(2q-1)>\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$$
$$\text{and }\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2<\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$$
$$\Longrightarrow (2\theta-1)(2q-1)>1$$

However, we have $2\theta-1\in(0,1)$ and $2q-1\in(2\theta-1,1)\subset(0,1)$, which implies $(2\theta-1)(2q-1)<1$, a contradiction. □

## Proofs of Propositions

Before we start the proofs, recall that $f(\theta,q,\delta)=\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2(2\theta-1)(2q-1)-\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$ and $g(\theta,q,\delta)=\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2-\frac{q-\theta}{\theta(1-q)+(1-\theta)q}\delta$. A Bayesian Updating Equilibrium (BUE) can be sustained if and only if $f(\theta,q,\delta)\leq 0$, and a Confirmatory Bias Equilibrium (CBE) can be sustained if and only if $g(\theta,q,\delta)\geq 0$.

## Proof of Proposition 1

<u>Claim:</u> When $\delta\geq 1$ (and $q\in(0.5,1)$), $\frac{\partial}{\partial\theta}f(\theta,q,\delta)>0$ for $\theta\in(0.5,q)$.

*Proof.* To prove the claim, it is sufficient to show that, given $\delta\geq 1$, $\frac{1}{2}\frac{\partial}{\partial\theta}f(\theta,q,\delta)>0$

for $\theta \in (0.5, 1)$.

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial\theta}f(\theta,q,\delta) &= (2q-1)\left(\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2+q(1-q)(2\theta-1)\right.\\
&\qquad\times\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\
&\qquad\left.\times\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2}-\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\right)\\
&\quad+\frac{q(1-q)}{(\theta(1-q)+(1-\theta)q)^2}\delta\\
&\geq (2q-1)\left(\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2+q(1-q)(2\theta-1)\right.\\
&\qquad\times\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\
&\qquad\left.\times\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2}-\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\right)\\
&\quad+\frac{q(1-q)}{(\theta(1-q)+(1-\theta)q)^2}\\
&= \left.\begin{aligned}&(2q-1)\left(\tfrac{\theta q}{\theta q+(1-\theta)(1-q)}-\tfrac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\cdot\\
&\left(\left(\tfrac{\theta q}{\theta q+(1-\theta)(1-q)}-\tfrac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)+q(1-q)(2\theta-1)\tfrac{1}{(\theta q+(1-\theta)(1-q))^2}\right)\end{aligned}\right\} \quad (1)\\
&\quad\left.-\left(\tfrac{\theta q}{\theta q+(1-\theta)(1-q)}-\tfrac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)(2\theta-1)(2q-1)q(1-q)\tfrac{1}{(\theta(1-q)+(1-\theta)q)^2}\right\} \quad (2)\\
&\quad\left.+q(1-q)\tfrac{1}{(\theta(1-q)+(1-\theta)q)^2}\right\} \quad (3)
\end{aligned}$$

Note that the first and third term above after the last equality are positive while the second term is negative. However, we know $(2q-1) \in (0,1)$, $(2\theta-1) \in (0,1)$, and $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}) \in (0,1)$, which implies $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q})(2\theta-1)(2q-1) \in (0,1)$. Thus, the sum of the second and the third term is still positive, implying that $\frac{1}{2}\frac{\partial}{\partial\theta}f(\theta,q,\delta) > 0$. □

Notice that $f(0.5,q,\delta) = -(2q-1)\delta < 0$ and $f(q,q,\delta) = (\frac{q^2}{q^2+(1-q)^2} - \frac{1}{2})^2(2q-1)^2 > 0$. Therefore, given any $\delta \geq 1$ and $q_0 \in (0.5,1)$, there exists a unique $\theta_0 \in (0.5, q_0)$ such that $f(\theta_0, q_0, \delta) = 0$; moreover, since $f(\cdot)$ increases in $\theta$, we have $f(\theta, q_0, \delta) < 0$ for $\theta \in (0.5, \theta_0)$, which completes the proof of Proposition 1.

**Proof of Proposition 2**

Claim: When $\delta \geq 1$ (and $q \in (0.5,1)$), $\frac{\partial}{\partial\theta}g(\theta,q,\delta) > 0$ for $\theta \in (0.5, q)$.

*Proof.* To prove the claim, it is sufficient to show that, given $\delta \geq 1$, $\frac{1}{2}\frac{\partial}{\partial\theta}g(\theta, q, \delta) > 0$ for $\theta \in (0.5, 1)$.

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial\theta}g(\theta, q, \delta) = q(1-q)&\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) \\
&\times\left(\frac{1}{(\theta q + (1-\theta)(1-q))^2} - \frac{1}{(\theta(1-q) + (1-\theta)q)^2}\right) \\
&+ \frac{q(1-q)}{(\theta(1-q) + (1-\theta)q)^2}\delta \\
\geq q(1-q)&\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) \\
&\times\left(\frac{1}{(\theta q + (1-\theta)(1-q))^2} - \frac{1}{(\theta(1-q) + (1-\theta)q)^2}\right) \\
&+ \frac{q(1-q)}{(\theta(1-q) + (1-\theta)q)^2} \\
= &\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)q(1-q)\frac{1}{(\theta q + (1-\theta)(1-q))^2} \\
&- \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)q(1-q)\frac{1}{(\theta(1-q) + (1-\theta)q)^2} \\
&+ q(1-q)\frac{1}{(\theta(1-q) + (1-\theta)q)^2}
\end{aligned}$$

Note that the first and third term above after the last equality are positive while the second term is negative. However, since $(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}) \in (0, 1)$, the sum of the second and third term is positive, implying that $\frac{1}{2}\frac{\partial}{\partial\theta}g(\theta, q, \delta) > 0$. □

Notice that $g(0.5, \delta, q) = (2q-1)(2q-1-\delta) < 0$ and $g(q, q, \delta) = (\frac{q^2}{q^2+(1-q)^2} - \frac{1}{2})^2 > 0$. Therefore, given any $\delta \geq 1$ and $q_0 \in (0.5, 1)$, there exists a unique $\theta_0 \in (0.5, q_0)$ such that $g(\theta_0, q_0, \delta) = 0$; moreover, since $g(\cdot)$ increases in $\theta$, we have $g(\theta, q_0, \delta) > 0$ for $\theta \in (\theta_0, q_0)$, which completes the proof of Proposition 2.

**Proof of Proposition 3**

<u>Claim:</u> When $\delta \geq 1$ (and $\theta \in (0.5, 1)$), $\frac{\partial}{\partial q}f(\theta, q, \delta) < 0$ for $q \in (\theta, 1)$.

*Proof.* Let $h(\theta, q) \equiv \frac{1}{2\theta(1-\theta)}(\theta q + (1-\theta)(1-q))^2(\theta(1-q) + (1-\theta)q)^2$. To prove the claim, it is sufficient to show that, given $\delta \geq 1$, $h(\theta, q) \cdot \frac{\partial}{\partial q}f(\theta, q, \delta) < 0$ for

$q \in (0.5, 1)$.

$$\begin{aligned}
h(\theta,q)\cdot\frac{\partial}{\partial q}f(\theta,q,\delta) = h(\theta,q)\cdot\Bigg[&2(2\theta-1)\Bigg(\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2\\
&+\theta(1-\theta)(2q-1)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\
&\times\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2}+\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\Bigg)\\
&-\frac{2\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\delta\Bigg]\\
<h(\theta,q)\cdot\Bigg[&2(2\theta-1)\Bigg(\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\
&+2\theta(1-\theta)(2q-1)\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2}+\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\Bigg)\\
&-\frac{2\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\Bigg]\\
=\ &2(2\theta-1)^3q^3+2(2\theta-1)^2(1-3\theta)q^2\\
&+(2\theta-1)(6\theta^2-4\theta+1)q+\theta(-2\theta^2+2\theta-1)\\
\equiv\ &\gamma(q;\theta)
\end{aligned}$$

We then complete the proof of claim by showing that $\gamma(q;\theta) < 0$ for any $\theta \in (0.5, 1)$ and $q \in (0.5, 1)$.

From the discriminant of a cubic polynomial, we know that, if $\kappa(x) = ax^3+bx^2+cx+d$ (where $(a, b, c, d) \in \mathbb{R}^4$), then $\left(\frac{bc}{6a^2}-\frac{b^3}{27a^3}-\frac{d}{2a}\right)^2+\left(\frac{c}{3a}-\frac{b^2}{9a^2}\right) > 0$ implies that $\kappa(x)$ only has one real root. Since $\frac{(2\theta-1)(6\theta^2-4\theta+1)}{6(2\theta-1)^3} - (\frac{2(2\theta-1)^2(1-3\theta)}{6(2\theta-1)^3})^2 = \frac{1}{18(2\theta-1)^2} > 0$, we can conclude that $\gamma(q;\theta)$ only has one real root (with respect to $q$).

Furthermore, we have $\gamma(1;\theta) = (\theta-1)(2\theta^2-2\theta+1) < 0$ for $\theta \in (0.5, 1)$. Notice that the coefficient of the cubic term of $\gamma(q;\theta)$ is $2(2\theta-1)^3 > 0$ for $\theta \in (0.5, 1)$. Thus, the (unique) real root of $\gamma(q;\theta)$ is greater than 0, which implies $\gamma(q;\theta) < 0$ for all $q \in (0.5, 1)$ (given $\theta \in (0.5, 1)$), as desired. □

Notice that $f(\theta,\theta,\delta) = (\frac{\theta^2}{\theta^2+(1-\theta)^2}-\frac{1}{2})^2(2\theta-1)^2 > 0$ and $f(\theta,1,\delta) = -\delta < 0$. Therefore, given any $\delta \geq 1$ and $\theta_0 \in (0.5, 1)$, there exists a unique $q_0 \in (\theta_0, 1)$ such that $f(\theta_0, q_0, \delta) = 0$; moreover, since $f(\cdot)$ decreases in $q$, we have $f(\theta_0, q, \delta) < 0$ for $q \in (q_0, 1)$, which completes the proof of Proposition 3.

**Proof of Proposition 4**

First, note that $g(\theta,\theta,\delta) = (\frac{\theta^2}{\theta^2+(1-\theta)^2} - \frac{1}{2})^2 > 0$, $g(\theta,1,\delta) = 1-\delta < 0$ when $\delta > 1$, $g(\theta,1,\delta=1) = 0$, and $\frac{\partial}{\partial q} g(\theta,q,\delta)|_{q=1,\delta=1} = \frac{1-\theta}{\theta} > 0$. Thus, $g(\cdot;\theta,\delta)$ has at least one root in $(\theta,1)$ given any $\theta \in (0.5,1)$ and $\delta \geq 1$. We then complete our proof of Proposition 4 by showing that the following two claims are true.

Claim 1: When $\delta \geq 2$ (and $\theta \in (0.5,1)$), $\frac{\partial}{\partial q} g(\theta,q,\delta) < 0$ for $q \in (\theta,1)$.

*Proof.*

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial q} g(\theta,q,\delta) = \theta(1-\theta)&\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\ &\times\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2} + \frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\\ &- \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\delta\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial q} g(\theta,q,\delta=2) = \theta(1-\theta)&\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\cdot\frac{1}{(\theta q+(1-\theta)(1-q))^2}\\ &- \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\\ &+\theta(1-\theta)\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\cdot\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\\ &- \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}\\ <0&\end{aligned}$$

Last, notice that $\frac{1}{2}\frac{\partial}{\partial q} g(\theta,q,\delta=2) < 0$ implies that $\frac{\partial}{\partial q} g(\theta,q,\delta=2) < 0$ for all $\delta \geq 2$. □

Claim 2: Given any $\delta \in [1,2]$ and $\theta \in (0.5,1)$, $g(\cdot;\theta,\delta)$ only has one (real) root in $(\theta,1)$.

*Proof.* Let $\kappa(q;\theta,\delta) = (\theta q+(1-\theta)(1-q))^2(\theta(1-q)+(1-\theta)q)^2 g(q;\theta,\delta)$. To prove Claim 2, it is sufficient to show that $\kappa(\cdot;\theta,\delta)$ only has one root in $(\theta,1)$ since $(\theta q+(1-\theta)(1-q))^2(\theta(1-q)+(1-\theta)q)^2 > 0$ for $\theta \in (0.5,1)$ and $q \in (\theta,1)$.

Recall that $g(\cdot;\theta,\delta)$ (and thus $\kappa(\cdot;\theta,\delta)$) has at least one root in $(\theta,1)$. We then prove that it has at most one root in $(\theta,1)$ by showing that $\frac{d^2}{dq^2}\kappa(q;\theta,\delta) > 0$ for $\delta \in [1,2]$ and $q \in (\theta,1)$.

After some tedious algebra, it can be shown that

$$\begin{aligned}\frac{d^2}{dq^2}\kappa(q;\theta,\delta) &= 12q\delta(2\theta-1)^2((2\theta-1)q+(1-\theta-\theta^2)) \\ &\quad + 2(4\theta^2(1-\theta^2)+\delta(2\theta-1)(6\theta^3-4\theta^2-2\theta+1)) \\ &> 12\theta\delta(2\theta-1)^2(\theta-1)^2 \\ &\quad + 2(4\theta^2(1-\theta^2)+\delta(2\theta-1)(6\theta^3-4\theta^2-2\theta+1)) \qquad [\because q>\theta] \\ &= 8\theta^2(1-\theta)^2+\delta(2\theta-1)(24\theta^4-48\theta^3+40\theta^2-16\theta+2) \equiv \Gamma(\theta;\delta)\end{aligned}$$

When $\delta = 2$, we have $\Gamma(\theta;2) = 96\theta^5 - 232\theta^4 + 240\theta^3 - 136\theta^2 + 40\theta - 4$. It can be numerically shown that $\Gamma(\theta;2)$ only has one real root ($\approx 0.183453$). Since $\Gamma(1;2) > 0$ and $\Gamma(0;2) < 0$, we can conclude that $\Gamma(\theta;2) > 0$ for $\theta \in (0.5,1)$. Furthermore, this implies that, given any $\theta \in (0.5,1)$ , $\Gamma(\theta;\delta) \geq 0$ for $\delta \in [1,2]$. If $(2\theta-1)(24\theta^4-48\theta^3+40\theta^2-16\theta+2) < 0$, then $\Gamma(\theta;\delta) \geq \Gamma(\theta;2) > 0$ for $\delta \in [1,2]$; if $(2\theta-1)(24\theta^4-48\theta^3+40\theta^2-16\theta+2) \geq 0$, then $\Gamma(\theta;\delta) > 0$ for $\delta \in [1,2]$ since $8\theta^2(1-\theta)^2 > 0$. Therefore, we get $\frac{d^2}{dq^2}\kappa(q;\theta,\delta) > \Gamma(\theta;\delta) \geq 0$ for any $\theta \in (0.5,1)$ and $\delta \in (1,2)$. □

Recall that $g(\cdot;\theta,\delta)$ has at least one root in $(\theta,1)$ given any $\theta \in (0.5,1)$ and $\delta \geq 1$; Claim 1 and Claim 2 imply that $g(\cdot;\theta,\delta)$ only has one root in $(\theta,1)$. That is, given any $\delta \geq 1$ and $\theta_0 \in (0.5,1)$, there exists a unique $q_0 \in (\theta_0,1)$ such that $g(\theta_0,q_0,\delta) = 0$. Moreover, since $g(\theta_0,\theta_0,\delta) > 0$ and $g(\theta_0,1,\delta) \leq 0$, we have $g(\theta_0,q,\delta) > 0$ for $q \in (\theta_0,q_0)$, which completes the proof of Proposition 4.

**Proof of Proposition 5**

Recall that $f(0.5,q,\delta) = -(2q-1)\delta < 0$ and $f(q,q,\delta) = (\frac{q^2}{q^2+(1-q)^2}-\frac{1}{2})^2(2q-1)^2 > 0$ given any $\delta > 0$. We then complete the proof of Proposition 5 by showing that, given any $\delta \in (0,1)$ and $q \in (0.5,1)$, $\frac{\partial}{\partial\theta}f(\theta_0,q,\delta)$ must be strictly positive for any $\theta_0 \in (0.5,q)$ such that $f(\theta_0,q,\delta) = 0$ (which implies that $f(\cdot;q,\delta)$ only intersects with the $x$-axis once for $\theta \in (0.5,q)$).

<u>Claim:</u> Given any $\delta \in (0,1)$ and $q \in (0.5,1)$, $\frac{\partial}{\partial\theta}f(\theta_0,q,\delta) > 0$ for $\theta_0 \in (0.5,q)$ such that $f(\theta_0,q,\delta) = 0$.

*Proof.* When $f(\theta,q,\delta) = 0$, we have $\delta = \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2(2\theta-1)(2q-1)\cdot\frac{\theta(1-q)+(1-\theta)q}{q-\theta} \equiv \delta_0$. To complete the proof, we need to show that $\frac{\partial}{\partial\theta}f(\theta,q,\delta)|_{\delta=\delta_0} > 0$.

Recall that

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial\theta}f(\theta,q,\delta) = (2q-1)\Bigg(&\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2\\ &+q(1-q)(2\theta-1)\cdot\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)\\ &\times\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2}-\frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right)\Bigg)\\ &+\frac{q(1-q)}{(\theta(1-q)+(1-\theta)q)^2}\delta\end{aligned}$$

From the above equation, we can derive $\frac{1}{2}(q-\theta)(\theta q+(1-\theta)(1-q))^2(\theta(1-q)+(1-\theta)q)^2(2q-1)^{-1}\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)}-\frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^{-1}\frac{\partial}{\partial\theta}f(\theta,q,\delta)\equiv f'(\theta,q,\delta)$. Note that given any $q\in(0.5,1)$, the terms preceding $\frac{\partial}{\partial\theta}f(\theta,q,\delta)$ are greater than 0 for $\theta\in(0.5,q)$. Thus, to complete the proof of the claim, it is sufficient to show that $f'(\theta,q,\delta_0)>0$ given any $q\in(0.5,1)$ and $\theta\in(0.5,q)$.

After some tedious algebra, we get

$$\begin{aligned}f'(\theta,q,\delta_0) =& -(2q-1)^3\theta^5+(16q^4-12q^3-8q^2+9q-2)\theta^4+(-32q^4+40q^3-8q^2-4q+1)\theta^3\\ &+(28q^4-40q^3+13q^2)\theta^2+(-12q^4+18q^3-6q^2)\theta+(2q^4-3q^3+q^2)\\ \equiv& F(\theta;q)\end{aligned}$$

It can be numerically shown that, given any $q\in(0.5,1)$, the largest real root of $F(\cdot;q)$ is greater than one, whereas the second largest real root (if exists) of $F(\cdot;q)$ is less than zero.[1] This implies that $F(\theta;q)>0$ for $\theta\in(0.5,1)$ since the coefficient of $\theta^5$ in $F(\cdot;q)$ is negative.[2]

□

**Proof of Proposition 6**

We complete the proof by showing that, given any $\theta\in(0.5,1)$ and $\delta\in(0,1)$, $g(\theta,q=\frac{1+\delta}{2},\delta)>0$ and $\frac{\partial}{\partial q}g(\theta,q,\delta)>0$ for $q\in[\frac{1+\delta}{2},1]$.

Claim 1: $g(\theta,q=\frac{1+\delta}{2},\delta)>0$ for any $\theta\in(0.5,1)$ and $\delta\in(0,1)$.

[1] We use Mathematica to obtain numerical solutions.

[2] In fact, our numerical result shows that $\min_{\theta\in[0.5,1],q\in[0.5,1]}F(\theta;q)=F(1,1)=0$, which also indicates that $F(\theta;q)>0$ for $q\in(0.5,1)$ and $\theta\in(0.5,q)$.

*Proof.* First, after some tedious algebra, we can get

$$g(\theta, q = \frac{1+\delta}{2}, \delta) = \frac{(4\theta\delta(1-\theta))^2 + \left(1-\delta^2\right)(1+(2\theta-1)\delta)^2(1-(2\theta-1)\delta)}{(1+(2\theta-1)\delta)^2(1-(2\theta-1)\delta)^2} - 1$$

Since both the numerator and the denominator of the first term are positive, $g(\theta, q = \frac{1+\delta}{2}, \delta) > 0$ if and only if $\eta(\delta;\theta) > 0$, where

$$\begin{aligned}
\eta(\delta;\theta) &= (4\theta\delta(1-\theta))^2 + \left(1-\delta^2\right)(1+(2\theta-1)\delta)^2(1-(2\theta-1)\delta) \\
&\quad - (1+(2\theta-1)\delta)^2(1-(2\theta-1)\delta)^2 \\
&= \big[(2\theta-1)^3\delta^4 + (2\theta-1)^2(1-(2\theta-1)^2)\delta^3 \\
&\quad - (2\theta-1)(1+(2\theta-1)^2)\delta^2 - (4\theta(1-\theta) - (4\theta(1-\theta))^2)\delta + (2\theta-1)\big]\,\delta \\
&\equiv \tilde{\eta}(\delta;\theta)\cdot\delta
\end{aligned}$$

To complete the proof of Claim 1, it is sufficient to show that, given any $\theta \in (0.5, 1)$, $\tilde{\eta}(\delta;\theta) > 0$ for $\delta \in (0, 1)$.

Since the coefficients of $\delta^4$, $\delta^3$, and the constant term in $\tilde{\eta}(\delta;\theta)$ are positive while the coefficients of $\delta^2$ and $\delta$ are negative, $\tilde{\eta}(\delta;\theta)$ has no or two positive roots (by Descartes' rule of signs). Note that $\tilde{\eta}(\delta = 1;\theta) = 0$; thus, $\tilde{\eta}(\delta;\theta)$ has two positive roots. Moreover, it can be shown that $\frac{d}{d\delta}\tilde{\eta}(\delta;\theta) < 0$ and $\lim_{\delta\to\infty}\tilde{\eta}(\delta;\theta) = +\infty$ (when $\theta \in (0.5, 1)$). Hence, one of the positive roots of $\tilde{\eta}(\delta;\theta)$ is larger than one. This (along with the fact that $\frac{d}{d\delta}\tilde{\eta}(\delta;\theta) < 0$) implies that $\tilde{\eta}(\delta;\theta) > 0$ for $\delta \in (0, 1)$. □

<u>Claim 2:</u> Given any $\theta \in (0.5, 1)$ and $\delta \in (0, 1)$, $\frac{\partial}{\partial q}g(\theta, q, \delta) > 0$ for $q \in [\frac{1+\delta}{2}, 1]$.

*Proof.* Recall that, in the proof of Proposition 4, we have shown that

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial q}g(\theta, q, \delta) = \theta(1-\theta)&\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) \\
&\times\left(\frac{1}{(\theta q + (1-\theta)(1-q))^2} + \frac{1}{(\theta(1-q) + (1-\theta)q)^2}\right) \\
&- \frac{\theta(1-\theta)}{(\theta(1-q) + (1-\theta)q)^2}\delta
\end{aligned}$$

When $q = \frac{1+\delta}{2}$, we have $\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right) = \frac{4\theta\delta(1-\theta)}{1-\delta^2(2\theta-1)^2} > 4\theta\delta(1-\theta)$. In addition, given $\theta \in (0.5, 1)$, we have $4\theta\delta(1-\theta) - \delta = \delta(4\theta(1-\delta) - 1) > 0$. Thus, we get $\frac{4\theta\delta(1-\theta)}{1-\delta^2(2\theta-1)^2}\cdot\frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2} > \delta\cdot\frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}$.

Furthermore, we know $\frac{\partial}{\partial q}\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right) = \left(\frac{1}{(\theta q+(1-\theta)(1-q))^2} + \frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right) > 0$. Therefore, we get $\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right) \cdot \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2} > \frac{4\theta\delta(1-\theta)}{1-\delta^2(2\theta-1)^2} \cdot \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2} > \delta \cdot \frac{\theta(1-\theta)}{(\theta(1-q)+(1-\theta)q)^2}$ for $q \in [\frac{1+\delta}{2}, 1]$, which implies $\frac{\partial}{\partial q} g(\theta, q, \delta) > 0$ for $q \in [\frac{1+\delta}{2}, 1]$. □

**Proof of Proposition 7**

Claim: When $\delta \in (0, 1)$ and $q \in (0.5, \frac{1+\delta}{2})$, $\frac{\partial}{\partial \theta} g(\theta, q, \delta) > 0$ for $\theta \in (0.5, q)$.

*Proof.* Recall that

$$\begin{aligned}
\frac{1}{2}\frac{\partial}{\partial\theta} g(\theta, q, \delta) &= q(1-q)\left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) \\
&\quad \times \left(\frac{1}{(\theta q + (1-\theta)(1-q))^2} - \frac{1}{(\theta(1-q) + (1-\theta)q)^2}\right) \\
&\quad + \frac{q(1-q)}{(\theta(1-q) + (1-\theta)q)^2}\delta \\
&= \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) q(1-q) \frac{1}{(\theta q + (1-\theta)(1-q))^2} \\
&\quad - \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) q(1-q) \frac{1}{(\theta(1-q) + (1-\theta)q)^2} \\
&\quad + \frac{q(1-q)}{(\theta(1-q) + (1-\theta)q)^2}\delta \\
&> \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) q(1-q) \frac{1}{(\theta q + (1-\theta)(1-q))^2} \\
&\quad - \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right) q(1-q) \frac{1}{(\theta(1-q) + (1-\theta)q)^2} \\
&\quad + (2q-1)\frac{q(1-q)}{(\theta(1-q) + (1-\theta)q)^2}
\end{aligned}$$

The last inequality holds since $q < \frac{1+\delta}{2}$ implies $\delta > 2q - 1$.

Moreover, since $\frac{\partial}{\partial\theta}\left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right)^2 = q(1-q)\left(\frac{1}{(\theta q+(1-\theta)(1-q))^2} - \frac{1}{(\theta(1-q)+(1-\theta)q)^2}\right) \leq 0$ when $\theta \in [0.5, 1]$, we have $2q - 1 - \left(\frac{\theta q}{\theta q+(1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q)+(1-\theta)q}\right) \geq 2q - 1 - (q - (1-q)) = 0$, which implies $\frac{1}{2}\frac{\partial}{\partial\theta} g(\theta, q, \delta) > 0$. □

Notice that $g(0.5, \delta, q) = (2q-1)(2q-1-\delta) < 0$ when $q < \frac{1+\delta}{2}$, and $g(q, q, \delta) = (\frac{q^2}{q^2+(1-q)^2} - \frac{1}{2})^2 > 0$. Therefore, given any $\delta \in (0, 1)$ and $q_0 \in (0.5, \frac{1+\delta}{2})$, there exists a unique $\theta_0 \in (0.5, q_0)$ such that $g(\theta_0, q_0, \delta) = 0$; moreover, since $g(\cdot)$ increases in

$\theta$, we have $g(\theta, q_0, \delta) > 0$ for $\theta \in (\theta_0, q_0)$, which completes the proof of Proposition 7.

**Proof of Proposition 9**

In the following, we prove the two statements in Proposition 9 separately.

Proof of the first statement (with respect to the BUE):

We first denote the boundary of the equilibrium condition for a BUE as

$$\delta_f(\theta; q) \equiv \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)^2 (2\theta-1)(2q-1)\cdot\frac{\theta(1-q)+(1-\theta)q}{q-\theta}$$

where $q \in (0.5, 1)$. Note that, given $\bar{q} \in (0.5, 1)$, $(\theta, \delta)$ can support a BUE if and only if $\delta \geq \delta_f(\theta; \bar{q})$.

We then complete the proof with the following four statements:

1. $\delta_f(0.5; q) = 0$ for any $q \in (0.5, 1)$.
2. Given $q'$ and $q''$ such that $q'' > q'$, we have $\lim_{\theta \to q'} \delta_f(\theta; q') = \infty$ and $\delta_f(\theta = q'; q'') < \infty$.
3. It can be numerically shown that $\frac{d}{d\theta}\delta_f(\theta; q'')|_{\theta=0.5} > \frac{d}{d\theta}\delta_f(\theta; q')|_{\theta=0.5} > 0$ when $1 > q'' > q' > 0.5$.
4. It can be numerically shown that $\frac{d^2}{d\theta^2}\delta_f(\theta; q') > \frac{d^2}{d\theta^2}\delta_f(\theta; q'')$ for $\theta \in (0.5, q')$ when $1 > q'' > q' > 0.5$.

Thus, $\delta_f(\cdot; q')$ only intersects with $\delta_f(\cdot; q'')$ once for $\theta \in (0.5, q'')$. Moreover, when $\delta$ is below that intersection point, we have $(\theta, \delta) \in \text{BUE}_{q''} \implies (\theta, \delta) \in \text{BUE}_{q'}$, and when $\delta$ is above that intersection point, we have $(\theta, \delta) \in \text{BUE}_{q'} \implies (\theta, \delta) \in \text{BUE}_{q''}$, as desired.

Proof of the second statement (with respect to the CBE):

We first denote the boundary of the equilibrium condition for a BUE as

$$\delta_g(\theta; q) \equiv \left(\frac{\theta q}{\theta q + (1-\theta)(1-q)} - \frac{\theta(1-q)}{\theta(1-q) + (1-\theta)q}\right)^2 \cdot \frac{\theta(1-q)+(1-\theta)q}{q-\theta}$$

where $q \in (0.5, 1)$. Note that, given $\bar{q} \in (0.5, 1)$, $(\theta, \delta)$ can support a CBE if and only if $\delta \leq \delta_g(\theta; \bar{q})$.

We then complete the proof with the following three statements:

1. $\delta_g(0.5, q) = 2q - 1$; thus, $\delta_g(0.5, q') < \delta_g(0.5, q'')$ when $1 > q'' > q' > 0.5$.

2. $\lim_{\theta \to q'} \delta_g(\theta; q') = \infty$ and $\delta_g(\theta = q'; q'') < \infty$ when $1 > q'' > q' > 0.5$.

3. It can be numerically shown that $\frac{d^2}{d\theta^2}\delta_g(\theta; q') > \frac{d^2}{d\theta^2}\delta_g(\theta; q'')$ for $\theta \in (0.5, q')$ when $1 > q'' > q' > 0.5$.

Thus, $\delta_g(\cdot; q')$ only intersects with $\delta_g(\cdot; q'')$ once for $\theta \in (0.5, q'')$. Moreover, when $\delta$ is below that intersection point, we have $(\theta, \delta) \in \text{CBE}_{q'} \implies (\theta, \delta) \in \text{CBE}_{q''}$, and when $\delta$ is above that intersection point, we have $(\theta, \delta) \in \text{CBE}_{q''} \implies (\theta, \delta) \in \text{CBE}_{q'}$, as desired.

**Proof of Proposition 10**

Let $\tilde{f}_\lambda(k) = \sum_{j=0}^{k-1} \binom{2k-1}{j} \lambda^{2k-1-j}(1-\lambda)^j$. To complete the proof of Proposition 10, it is sufficient to show that $\tilde{f}_\lambda(k+1) - \tilde{f}_\lambda(k) > 0$ for all $k \in \mathbb{N}$ and $\lambda \in (0.5, 1)$.

$$
\begin{aligned}
&\tilde{f}_\lambda(k+1) - \tilde{f}_\lambda(k) \\
&= \lambda\left((\lambda^{2k} - \lambda^{2k-2}) + \sum_{j=1}^{k}\binom{2k+1}{j}\lambda^{2k-j}(1-\lambda)^j - \sum_{j=1}^{k-1}\binom{2k-1}{j}\lambda^{2k-2-j}(1-\lambda)^j\right) \\
&= \lambda(1-\lambda)\left(\sum_{j=1}^{k}\binom{2k+1}{j}\lambda^{2k-j}(1-\lambda)^{j-1} - \sum_{j=1}^{k-1}\binom{2k-1}{j}\lambda^{2k-2-j}(1-\lambda)^{j-1} - \lambda^{2k-1} - \lambda^{2k-2}\right) \\
&= \lambda^k(1-\lambda)\left(\sum_{j=1}^{k}\binom{2k+1}{j}\lambda^{k+1-j}(1-\lambda)^{j-1} - \sum_{j=1}^{k-1}\binom{2k-1}{j}\lambda^{k-1-j}(1-\lambda)^{j-1} - \lambda^{k} - \lambda^{k-1}\right) \\
&= \lambda^k(1-\lambda)^2\left(\sum_{j=2}^{k}\binom{2k+1}{j}\lambda^{k+1-j}(1-\lambda)^{j-2} - \sum_{j=2}^{k-1}\binom{2k-1}{j}\lambda^{k-1-j}(1-\lambda)^{j-2}\right. \\
&\qquad \left. - \left(\binom{2k-1}{1} + \binom{2k-1}{0}\right)\lambda^{k-1} - \binom{2k-1}{1}\lambda^{k-2}\right) \\
&= \lambda^k(1-\lambda)^3\left(\sum_{j=3}^{k}\binom{2k+1}{j}\lambda^{k+1-j}(1-\lambda)^{j-3} - \sum_{j=3}^{k-1}\binom{2k-1}{j}\lambda^{k-1-j}(1-\lambda)^{j-3}\right. \\
&\qquad \left. - \left(\binom{2k-1}{2} + \binom{2k-1}{1}\right)\lambda^{k-2} - \binom{2k-1}{2}\lambda^{k-3}\right) \\
&= \cdots \\
&= \lambda^k(1-\lambda)^{k-1}\left(\binom{2k+1}{k-1}\lambda^2 + \binom{2k+1}{k}\lambda(1-\lambda) - \binom{2k-1}{k-1} - \left(\binom{2k-1}{k-2} + \binom{2k-1}{k-3}\right)\lambda^2 - \binom{2k-1}{k-2}\lambda\right) \\
&= \lambda^k(1-\lambda)^{k-1}\left(\binom{2k-1}{k-1}(1-\lambda)(2\lambda-1)\right) \\
&= \binom{2k-1}{k-1}\lambda^k(1-\lambda)^k(2\lambda-1) \\
&> 0 \qquad \text{for any } \lambda \in (0.5, 1)
\end{aligned}
$$