Title
Essays on the Role of Leadership on Cooperation, Competition and the Consolidation of Power

Author
Hernandez, Pablo Ignacio

Publication Date
2013

Peer reviewed|Thesis/dissertation

Essays on the Role of Leadership on Cooperation, Competition and the Consolidation of Power

By

Pablo Ignacio Hernández

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Ernesto Dal Bó, co-Chair

Professor John Morgan, co-Chair

Professor Rui de Figueiredo

Professor Shachar Kariv

Spring 2013

Abstract

Essays on the Role of Leadership on Cooperation, Competition and the Consolidation of Power

by

Pablo Ignacio Hernández

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Ernesto Dal Bó, co-Chair

Professor John Morgan, co-Chair

We explore the association between leadership and economic success in three different settings. First, we offer a new approach to study how groups lacking formal leaders coordinate actions (e.g. innovation, corporate transformation, civil uprising, etc.). Using controlled laboratory experiments, we find that leadership is a critical catalyst for cooperation. By varying the payoffs for not cooperating when others do so, we identify an interaction effect between the characteristics of leaders and the underlying context in which leadership emerges. In particular, when these payoffs are low, leadership is ubiquitous: no special features distinguish leaders. When change requires overcoming high monetary incentives favoring the status quo, leaders tend to be exceptional. These types of leaders exhibit a distinct non-monetary taste for mutual cooperation. This result implies that recruiting communally oriented individuals may facilitate innovation or change within firms.

Next, we study the extent to which non-pecuniary motives explain collusive behavior and consequently low productivity in firms. We find that classical collusion, obtained by explicit or implicit agreements between individuals, is fueled by other-regarding concerns. Each member showing other-regarding concerns decreases group effort by 15% of the average effort in a group of three individuals. Moreover, the emergence of a selfish leader from an originally leaderless group exacerbates this collusive behavior. Thus, the propensity of collusion is highest in a group of one selfish individual and two other-regarding ones. Contrary to our first essay, leadership in this case is detrimental for productivity because of conflicting interests between the workers and the firm.

Finally, we study how incumbents consolidate power over time. Using a formal theory model, we identify a mechanism by which leaders break away from the following trap: the more growth is fostered, the stronger the incentives for challengers to contest control. The mechanism hinges on two forms of investments by the incumbent: coercive capacity and productive capacity. In an application to state consolidation, we derive lessons for state-building by showing how investments in coercive capacity might have to precede investments in productive capabilities to ensure prosperity and peace. We also show how economic shocks and arms innovations may trigger state consolidation and economic take-off or keep polities in a trap of conflict and economic stagnation.

To my wife, my parents, my brother and sister

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to Ernesto Dal Bó and John Morgan for their invaluable help, support and encouragement throughout this journey. I also wish to thank Rui de Figueiredo, Shachar Kariv, Steve Tadelis and Noam Yuchtman for their insightful suggestions and valuable advice in each stage of this process. I would also like to thank Cameron Anderson, Pnina Feldman, José Guajardo, Dylan Minor, Don Moore, Mariano Moszoro, Dave Mowery, Denis Nekipelov, Santiago Oliveros, Dana Sisak, Pablo Spiller, Felix Várdy, Reed Walker and my fellow PhD students. In particular, I wish to thank Juan Pablo Atal, Ron Berman, Bo Cowgill, Isaac Hacamo, Lucy Hu, Tomás Reyes, Orie Shelef and Santiago Truffa for all the conversations.

I am also grateful to Kim Guilfoyle, Rowie del Castillo, Bradley Jong and Miho Tanaka for their invaluable assistance and support.

Chapter 1

On the Origins of Leadership Through Communication: The Role of Context and Social Preferences

1.1 Introduction

The emergence of a leader—someone who takes the initiative to advocate for a promising idea—may be the necessary catalyst to turn that idea into a successful innovation. Steve Jobs' initiative to promote and sell his teammate Steve Wozniak's Apple I computer in the seventies, for example, gave rise to what is now the largest technology firm in the world.[1] Great ideas may be quickly forgotten if no such leader emerges. Nokia experienced the unfortunate consequences of this seven years ago, when researchers were unable to persuade the company to commercialize their smartphone—a precursor to Apple's iPhone. The reason, *The Wall Street Journal* claims, was "internal rivalries" (see "Nokia's Bad Call on Smartphones," *The Wall Street Journal*, July 18 2012). As of today, Nokia has conceded its place as the biggest cell-phone maker to Samsung Electronics Co. Leadership emergence, as opposed to formal authority, is increasingly important, as organizational structures flatten and communication technologies evolve. Today, more than ever, change is likely to rely on the initiative of informal, emergent leaders. This study focuses on the interplay between the characteristics of those leaders and the context in which they emerge.

Of course, the importance of leadership emergence reaches beyond industry. The Arab Spring, the Occupy movement, and the protests in Moscow following Putin's re-election all relied on leaderless groups coalescing and demonstrating out on the streets. While these groups lacked formal leaders, informal leaders quickly emerged. These were the individuals using communication tools like Twitter to organize the protests. Also in these situations, emergent leadership was critical for the success or failure of these social movements.

There is little understanding, however, of the characteristics of emergent leaders as well as the determinants of their success in fostering change. In order to make progress on this, we study leadership emergence in a laboratory setting.

The challenge of transformational leaders is to convince loyalists—those inclined to favor the status quo. Our main treatment is to vary the rewards for loyalty in the face of dissent. In the example of social movements, these rewards concern to what happens to those who stay home in the face of demonstrations. In business settings, we can think of this as the rewards for those who remain loyal to the company line in, say, heated discussions or debates.

Change, be it regime overthrow or corporate transformation, can only come when everyone chooses the "protest" option. This corresponds to demonstrating in a political

[1]Popular stories from Steve Jobs' teammates at the germinating Apple Computers, Inc. back in the seventies, reflect his skills at persuasion and salesmanship. Andy Hertzfeld (former Steve Jobs' teammate on the Mac project), for instance, wrote in his website (http://folklore.org/StoryView.py?project=Macintosh&story=Reality_Distortion_Field.txt): "He [Jobs] can convince anyone of practically anything," explaining a metaphore extracted from Start Trek (the Reality Distortion Field) about Jobs' working style. As history has shown, this remarkable skill contributed to change the way the computer industry evolved. Officially, Steve Jobs' biographer Walter Isaacson, acknowledges Jobs was aware of this talent.

movement or to pushing to change procedures, policies or even strategies in a corporate context.

In settings where the rewards for loyalty are modest, change happens often. The coordination outcome occurs 70% of the time in our study. Moreover, emergent leadership is universal. In 94% of the cases, a leader emerges. These leaders have no special characteristics that may differentiate them from the whole populace.

When there is a premium for loyalty, the situation is dramatically different. Change comes more rarely, only 23% of the time in our study. A leader emerges 75% of the time, less often than under the low rewards scheme. Moreover, these leaders are different than the populace as a whole. They are more reciprocal altruistic and lying averse. Reciprocal altruism is the idea that a leader gets an intrinsic value when others join him or her embracing change. Lying aversion is the cost of breaking one's word. While the first characteristic is always associated with emergent leadership, the second is not. Among reciprocators, lying aversion is positively associated with leadership, among non reciprocators, it is negatively associated.

Leaders' initiative is not merely cheap talk—words often lead to action. Leaders are more likely to choose the protest option than others, regardless of rewards for loyalty. Moreover, leadership is critical for change to happen. Change occurs 73% of the time when a leader emerges (relative to 24% when no leader emerges) in the low rewards scheme and in 31% of the time when a leader emerges (relative to 1% when no leader emerges) in the high rewards scheme. Moreover, persuasion of a leader is effective only in the high rewards to loyalists scheme. Individuals who do not observe others leading cooperate 35% of the time in this regime, while they do it 45% of the time when they observe someone else leading. In the low rewards scheme, those who do not observe someone else leading cooperate 82% of the time, and when they do, they opt for protesting 81% of the time—not a significant difference.

Rewards to loyalists determine the likelihood of change and the characteristics of those who lead. The negative relationship we find between rewards to loyalists and change is direct. The link between these rewards and the characteristics of those who lead, however, is more subtle. To investigate this link, we rely on simple experimental games. Our setting consists of two individuals who must organize to achieve change—to push for an innovation or to overthrow a regime. If both choose the "protest" option, change obtains, and this is the best outcome from a group perspective. There is, however, a risk to pushing for change. If only one individual adheres, then the incumbent, be it the government or the manager, punishes that individual and he or she suffers the worst possible outcome. An individual can avoid this risk by sticking to the status quo. When everyone is loyal, status quo prevails and both individuals obtain lower payoffs than had they both decided to change. As we mention above, the key variation in context concerns to what happens to "loyalists" in the face of dissent from others. One possibility is that they receive no additional reward. Individuals are inclined to join change because loyalty is not attractive when others are pushing for change. Alternatively, the regime might choose to reward loyalists in this situation. Here, we imagine that the reward is so large that loyalists are

materially better off when no change occurs than when change prevails.

These two situations may be readily recognized as a stag hunt game (a coordination game in which defection is risk dominant) and a prisoner's dilemma game. Leadership consists of initiating messages to organize change. In the stag hunt, it is an equilibrium for everyone to demonstrate though there is obviously some risk involved. Here, the role of a leader is to create the trust necessary for everyone to take to the streets or to support the new idea. In contrast, in the prisoner's dilemma adhering to change is a dominated strategy. Here, a leader must convince others to override their self-interest. There is, however, a temptation present for a successful leader. By rallying the masses to the streets or making others to push for change and then staying loyal, the leader can benefit from the reward for loyalty.

Convincing others through pre-play communication may not be an easy task. In effect, under standard preferences, extant theory on the effects of pre-play communication predicts that it is impossible to reach mutually beneficial outcomes in contexts such as our prisoner's dilemma game. Leadership through speech, therefore, should be of no use. We show that this theoretical prediction is not borne out in the data. Leadership through communication fosters change in the stag hunt game, and perhaps surprisingly, in the prisoner's dilemma game as well. How is this possible? The explanation is that individuals have non-pecuniary concerns. Leadership in the stag hunt does not need to count on non-pecuniary motivations to be effective—it is just a matter of making the best outcome focal. If the context presents a dilemma, however, then reciprocal altruism and lying aversion play a role in leadership emergence.

So how does leadership in these two contexts relate to lying aversion and reciprocal altruism? Lying aversion is the cost of breaking one's word. Since honoring one's word is optimal in the stag hunt, there is no tension here. Leadership decisions do not hinge on this trait. For the prisoner's dilemma, the situation is quite different. The temptation to receive a reward for loyalty (and to avoid the possibility of severe punishment) provides an incentive for a leader to break his or her word. Thus, potential leaders who anticipate that they will succumb to this temptation may choose to eschew leadership altogether; or once committed to leadership, lying aversion provides commitment to follow through and this may make a leader more persuasive. In short, the effect of lying aversion on leadership is context dependent.

Reciprocal altruism is the intrinsic value a leader gets when others join him or her in demonstrating or in embracing new ideas. In the stag hunt, this effect merely reinforces the extrinsic incentives already present and so again it is of no consequence. In contrast, reciprocal altruism undermines the temptation to remain loyal when others are out demonstrating or advocating for change in the prisoner's dilemma. Thus, it undercuts the incentive effects of the rewards to loyalists. These additional gains drive individuals with high intrinsic benefits from cooperation to leadership roles.

Moreover, there is potentially an interaction effect. An individual who is both lying averse and reciprocally altruistic is likely to take a leadership role since the two effects may reinforce one another. A lying averse but not reciprocally altruistic individual will

shy away from leadership since there is no particular benefit to rallying others and a substantial cost should the individual succumb to temptation and remain loyal. An individual who is neither lying averse nor reciprocally altruistic is more likely to take a leadership role. Here, the individual is, in effect, acting as an agent—of the state or the company—seeking to "out" the state's enemies or those who are not aligned with the company line, in exchange for the rewards from loyalty. The point is simply that context is crucial as to whether these characteristics are activated in determining the identity of leaders.

In sum, the heart of the study is to place subjects in leaderless groups where they have an opportunity to communicate. We perform a battery of tests to measure these characteristics—reciprocal altruism and lying aversion—as well as a host of others that have been found to be indicative of leadership. We observe the identity of the person initiating a plan for change, observe whether the plan is agreed to, and then study subsequent actions. The key treatment variation is the context in which leadership takes place, i.e., the reward scheme for loyalists. In our baseline treatment, the game is a stag hunt—the regime offers no additional rewards to loyalists. We compare this to a treatment where the regime lavishly rewards loyalists in the face of demonstrations, i.e. a prisoner's dilemma situation.

The rest of the paper is divided as follows. In Section 2 we provide a literature review, in Section 3 we describe the experimental design, in Section 4 the theory benchmark and in Section 5 the hypotheses. In Section 6 we show the results and in Section 7 we conclude.

1.2 Literature

The literature on leadership is vast, with great many different definitions and approaches to the phenomenon. It seems however, there is consensus in one idea: Leadership is the process of using influence to attain mutual goals [9, 57, 13]. The behavioral theories of leadership seek for sound classifications of such "process of using influence." These theories can be grouped in two broad categories [28]: transformational-transactional leadership, and leadership as consideration and initiating structure. On one hand, the transactional-transformational approach, introduced by [16], analyzes leaders' actions to manage rewards and punishments (the transactional approach), and actions to produce a significant change in people's lives (the transformational approach). These approaches focus on behavior of established leaders. On the other hand, the consideration and initiating structure approach, born from the early studies on leadership emergence from leaderless groups [45, 8, 70], sees leaders as planners who guide groups (initiate structure) taking into account individuals' needs (consideration).

This behavioral approach to leadership was introduced as an alternative to the traditional trait theories, which started back in the 19th century (for instance see, [34]). The rich literature on leaders' traits has documented several distinctive features of leaders. In a nutshell, emotional traits such as the feeling of power, ambition, extroversion

[9, 51], or abilities and skills such as energy, intelligence, verbal fluency, confidence and independence [9, 5] have been related to leadership emergence and effectiveness.

Perhaps surprisingly, the behavioral perspective has developed mostly in parallel with the traits perspective, even though scholars now agree that both are essentially related [28]. In effect, researchers often lament the lack of integration between the behavioral and trait based approaches [4]. The economic view of leadership integrates these two because it posits "traits" as the primitives driving behavior, given contextual incentives. This more integrative view is the one we embrace in this study.

Initiative may be driven by the interplay of traits and context. There are a myriad of possible situations a group of individuals can be immersed in. Perhaps, two of the most important can be represented as coordination games and social dilemmas. In coordination games, leadership has been stressed in several fields. In political science, for instance, [17] stresses the role of leaders as the solvers of the coordination problems groups face. Similarly, in social psychology, [73] and others have pointed out that leadership evolved as the dominant channel through which groups cope with coordination problems. In economics, the role of leaders as coordinators was first addressed theoretically by David Kreps [53]. Kreps' main idea posits leaders as the ones who can coordinate in the presence of multiple equilibria. Although interesting, little subsequent work has been done in economics analyzing leaders as coordinators (for a clear discussion see [47]) in a way consistent with political science and social psychology findings [39].

In social dilemmas, the most common form of leadership found in the literature is "leading-by-example." Benjamin Hermalin [46] shows that individuals endowed with private information about the value of a public good, could lead-by-example or lead-by-sacrifice in order to signal this information. [60] experimentally tests Hermalin's theory. The authors find that, although signaling plays a role, reciprocity (mimicking leaders' contribution) provides a better rationale for the data, in both leading-by-example and leading-by-sacrifice. Along the same lines, [44, 62, 41] study sequential contributions to a public good. All of them find that letting one member to contribute first raises contributions, mainly because of the large contributions of these leaders. These studies focus on exogenously imposed leaders.

There has been research on endogenous leading-by-example as well, particularly on charitable giving. [64], for instance, also assumes that leaders have private information, but analyzes endogenous sequence of contributions. They find that endogenous sequential contribution also improves outcomes. This study, however, do not put emphasis on leaders (and non-leaders) characteristics.

To my knowledge, there are three studies in which traits and endogenous initiative (by example) are put together in a public goods game. The first one is [15]. From a novel experimental design, the authors identify costly endogenous leadership behavior and elicit individuals' traits. Generosity, strong preferences for efficiency, above-average cognitive skills, internal locus of control and patience are related to individuals who take the initiative. The second one is by [66]. In a concise study, they find voluntary leadership is more efficient than exogenously imposed leadership. The third study is [3]. It consists

of a public good experiment similar to [66], that also measures generosity, gender and personality traits. The main treatment is to reveal group members' attributes (gender and generosity) when contribution is voluntary. The authors elicited generosity by asking individuals to give a portion of their show-up fee to a charity, and personality traits through the Big 5 personality test[50]. They also find that leading voluntary contributions yield to more efficient outcomes, especially in groups with a high number of generous individuals. Males are more likely to be leaders, and females contribute more on average.

Overall, these papers find that initiative taking (by example) improves outcomes and that other-regarding concerns influence the decision to lead. They however, do not exogenously vary the level of strategic conflict (in our case, rewards to loyalists in the face of dissent) and they do not allow for communication. We believe that strategic conflict is fundamental in leadership emergence and that communication is perhaps the most pervasive mechanism of exhortation. We believe ours is the first study addressing the former issue. Regarding the latter, however, research is extensive, although no particular emphasis has been placed on leadership [56, 69].

Pre-play communication can be considered as a "device" by which individuals can coordinate in some [correlated] equilibrium ([38, 7]) of a given normal form game. This literature provides a rationale for how communication can extend the set of equilibrium outcomes; and it does well in explaining coordination when it is an equilibrium to do so. It fails, however, to explain the common finding of cooperation in social dilemmas. In effect, when communication is allowed, a robust finding in experimental studies of social dilemmas (for surveys see [56, 69] and [23]) and of coordination games[21] is that pre-play communication increases the frequency of the efficient outcome. Alignment of incentives makes coordination through communication a reasonable explanation in coordination games. In social dilemmas, scholars have proposed behavioral explanations in line with reciprocity, fairness and a cost of letting others down to explain the success of pre-play communication.[2] [32], for instance, develops a model that aims at explaining, among other things, why communication enhances outcomes by assuming preferences for equality and a fixed cost for being caught lying. The nature of these non-monetary drivers of cooperation has been further explored. [61] measures guilt (through self-reported emotional reaction) when individuals do not honor their word in social dilemma games with pre-play communication. The measure of guilt is positively correlated with cooperation in a prisoner's dilemma game. Along the same lines, [18] provides a mechanism through which guilt aversion leads to individuals to reciprocate if other trusted in a trust game:

[2]When non-pecuniary motives are introduced, [54], for example, provides a theoretical explanation for cooperation in a finitely repeated prisoner's dilemma based on the existence of individuals who enjoy a non-monetary gain when mutual cooperation occurs. [2] shows that [54] non-monetary concern or "reciprocal-altruism" explains the data from an experimental design. More recently, [14] provides a model based on other regarding preferences (that imbeds a taste for reciprocity and for fairness) to explain the positive correlation between wage offers and subsequent effort found in the literature (among other non-standard economic behavior). In sum, cooperation in social dilemmas have been found experimentally even without communication [2, 22].

individuals face a cost if they believe are letting others down. [42] argues that individuals may also posses a fixed cost of lying. In his experimental design Gneezy asks individuals to send a truthful or a deceitful message to another person about two possible splits of a pie; he then makes them play a dictator game with the same options in the message. He finds that a considerable proportion of his sample told the truth, but chose the selfish allocation in the dictator game, suggesting an intrinsic cost of lying. If these non-monetary motives rationalize apparently irrational behavior, it is sensible to ask whether these motives have a bearing on who emerges as a leader.

In a nutshell, our contribution is twofold. First, we assess the role of leadership through communication on successful change. Second, we show that leaders are different when leading is "hard" and when leading is "easy." That is, we identify an interaction effect between the characteristics of leaders and the underlying context in which leadership emerges. Thus, leadership, in our case, is context dependent. As a result, we intend to advance in the integration of the trait and the behavioral approaches to leadership.

1.3 Experimental Procedure

The experiment was constructed to study the key drivers of leadership and measure the effectiveness of leadership activity. The underlying idea is that leadership is context specific. When a would-be leader merely needs to achieve coordination on an outcome consistent with selfish behavior, there is little personal risk to the leader and leadership is ubiquitous. In our framework, this is the case were rewards to loyalists are low. In contrast, when the situation is one where the leader must persuade others to override self-interest in choosing an action, only highly motivated individuals will assume the leadership role. In our particular context, motivation comes in the form of a desire for reciprocity and indifference towards lying. Neither of these traits has received much attention in the extant literature on leadership.

The design is structured to ascertain key personal characteristics such as reciprocity, lying aversion, altruism and risk aversion. It also measures standard leadership traits such as extroversion, agreeableness, internal locus of control and intelligence. Finally, we measure demographics such as gender, race and ethnicity. Age and experience might also correspond to leadership; however, our subject population of mainly undergraduates lacks much variation along these dimensions. The heart of the design consists of varying the context in which potential leadership emerges. We now turn to the details.

The experiment consisted of two main treatments of two sessions each (96 subjects in total, 48 in each treatment) and two secondary treatments (96 subjects in total, 48 in each treatment). The only difference between the main and the secondary treatments is that in the former chat box pre-play communication is allowed and in the latter it is not. The instructions for the experiment were passed out to the participants and read aloud before the session began. A copy of the instructions is in the Appendix A-A. Participants did not interact with the experimenter, except to ask questions immediately after the instructions

were read and before the experimental tasks began. The experimental currency was the Berkeley Buck ($) and the exchange rate was $12 per US$1. There are three parts to each experimental session for both main and secondary treatments. We describe them in detail in the same order they were presented to participants in each session.

1.3.1 Part 1: Social Preferences and Lie Aversion

It was intended to elicit unconditional social preferences and a proxy for lying-aversion. The procedure in Part 1 was the same for all the treatments. It consisted of three blocks. In the first block, we elicited unconditional social preferences by asking the subjects to divide 10 tokens. The exact text presented in the computer screen was:

> Divide 10 tokens. Allocate a number of tokens to yourself (hold) and a number of tokens to the other participant (pass).
>
> A token is worth $X to you and $Y to the other participant. Please choose a division (total 10 tokens).
>
> Hold (1 token = $X):_____
>
> Pass (1 token = $Y):_____

Four different situations (values of X and Y) were presented to the subjects in the following order: (X=$1, Y=$1.25), (X=$1, Y=$1), (X=$1, Y=$0.67) and (X=$1, Y=$2). By varying the value of keeping versus passing each token, we can characterize the subject choices as either selfish or non-selfish. A subject was categorized as selfish if she kept all the tokens regardless of relative "prices;" otherwise a subject was labeled as non-selfish. This revealed preference elicitation procedure was first devised by [1] and subsequently extended by [37].

Payouts for this part were calculated by randomly selecting one of the four allocation decisions. Subjects were randomly matched in anonymous pairs to execute the payouts dictated by that selected allocation. As a result, each participant received his/her value held and the value passed by his/her matched partner from the corresponding selected allocation. The matching was performed at the end of the experiment.

In the second block of Part 1 we elicited lying-aversion using a procedure similar to [42]. Subjects faced two options featuring different divisions of $20. The options were: Option 1) keep $15 to him/herself and give $5 to other participant or Option 2) keep $5 to him/herself and give $15 to other participant. Subjects were asked to send a pre-codified message to another subject, who did not know which option corresponded to which set of payoffs. Participants could send a deceitful message reading "Option 1) will earn you [the subject the message was intended for] more money than Option 2);" or they could send a truthful message reading "Option 2) will earn you [the subject the message was intended for] more money than Option 1)." We randomized the order of Option 1) and Option 2) and used colors (Blue and Red) instead of numbers (Option Blue instead of Option 1), etc.) to avoid decisions based on mechanical ordering. Participants were anonymously

matched in pairs at the end of the experiment. Subjects were told the message would be delivered to another randomly matched participant at that time, and the amount of money they both would get depended on this other subject's decision. Each subject received the payout from his/her own decision after observing the matched participant's message and the payout from the matched participant's decision after reading his/her own message.

We also asked subjects the probability their "advice" would be followed by the person with whom they would be matched. This permits us to distinguish between circumstances where the deceitful message was sent with an intent to deceive versus when it was sent to counteract the partner's skepticism about the veracity of the message. For instance, a subject wishing to offer helpful advice, but suspecting her partner will do the opposite of whatever message was sent, could only achieve her objective by sending the deceitful message rather than the truthful one. Thus, it seems important to distinguish white lies, deceitful messages intended to lead the partner to choose the higher payoff option, from black lies, deceitful messages sent with the intent to trick the partner into choosing the lower payoff option. In coding for lying aversion, we treat white lies as equivalent to truthful reports.

The procedure above potentially confounds lying aversion with altruism. A sufficiently altruistic subject who is not lying averse may still send a truthful message purely out of desire to be generous towards her partner. To untangle the two effects, we again follow [42] and implement a non-strategic version of the message game above. Here, every subject gives "advice" to a computer, which follows it with the same probability indicated by the subject in the game before. Subjects did not know the probability was going to be the same in both versions, they were only told in the instructions (see the instructions in the Appendix A-A) the computer may execute their decision with "some" probability. Since lying to a computer does not carry the same moral stigma than lying to another person, choices in this round should purely reflect other-regarding preferences, to the extent they are present. We code a subject as a lying averse if they sent a truthful message or a white lie to a human and a selfish "message" to the computer. A subject is not lying averse otherwise.

Payoffs in this latter procedure were determined exactly as in the previous one, except for the fact that the computer made all the decisions (reversal and matching). Appendix A-B shows a screen-shot of the interface for this portion of the experiment.

1.3.2 Part 2: Belief elicitation and treatment conditions

In part 2 of the experiment, subjects played a one shot prisoner's dilemma followed by a stag hunt (SH treatment) or a prisoner's dilemma (PD treatment), depending on the treatment. This process was repeated 12 times, each with a different partner. The treatment was fixed over each session.

In the one-shot prisoners' dilemma (OSPD for future reference), subjects chose between

Cooperate or Defect.[3] In the same screen in which subjects picked their action, they were asked to forecast how many other subjects would choose to cooperate. If a subject exactly predicted how many of the other 23 subjects in their session cooperated they would get $8. Subjects lost $1 times the (absolute) difference between their guess and the true figure.[4]

Half of the rounds of the OSPD were randomly selected for payment. In a selected round, each subject was randomly matched with a partner to compute payoffs. Similarly, forecasts for half of the rounds were compensated. These determinations were made at the end of the experiment. Thus, after each round, subjects received no feedback about their payoffs nor their forecasts.

This portion of the experiment provides a measure of reciprocal altruism. Subjects who cooperate when they forecast some cooperation from others are coded as reciprocal altruists.

In the stag hunt (SH) or prisoner's dilemma (PD) treatments, in contrast with the other parts of the experiment, subjects were matched with another subject to participate in an interaction that would have immediate, rather than deferred, payoff consequences and feedback. In each round a subject would be matched with one other subject using a rotation matching protocol ([22, 26]). This procedure divided participants in each session into two groups and then matched each subject in one group with one subject in the other group, without repetition. This ensured that any pair of subjects were matched at most once and that one subject was not matched with a participant in his/her own group. The goal of this procedure is to minimize strategic effects across rounds.

Each main treatment consisted of two screens. In the first screen, participants observed a payoff matrix corresponding to the Coordination Game or Prisoner's Dilemma respectively, as in Table 1.1. On the left of each matrix, there was a chat box in which subjects could communicate for 30 seconds prior to making choices. Once the 30 seconds

[3]We labeled options as A or B, and randomized the link to Defect or to Cooperate. That is, in some rounds, A was the "Defect" action and in others it was the "Cooperate" action. The same was true for B. We also randomize the entries in the payoff matrix. We used different payoffs in each of the twelve rounds of the OSPD. We kept constant the net benefit of defection (equal to $5) if the other cooperated and the net benefit of defection if the other defected (equal to $4) to make it comparable to the PD game in Table (1.1). We also varied the order in which options (Cooperate or Defect) were presented. In some of the 12 rounds, the option equivalent to Defection was presented as the first row, and in others the option equivalent to Cooperation was presented in the first row, to avoid mechanical behavior. This ordering was random. These changes across rounds were intended to encourage participants to pay attention to the payoffs and the options in each round to avoid automatic responses.

[4]With this procedure, we elicit an statistic of the distribution (the median) of the number of participants who would cooperate. In theory, a risk neutral agent i chooses y to maximize

$$8 - \sum_{n=1}^{23} |y - n| p_i(n)$$

where $p_i(n)$ is the probability (belief) agent i assigns to n participants cooperating. The optimal choice $y^* = median_i(n)$. Therefore, y^* gives us information about the individual $i's$ beliefs, $p_i(n)$, about overall cooperation.

1\2	**D**efect	**C**ooperate
Defect	$4, 4$	$\mathbf{8}, 0$
Cooperate	$0, \mathbf{8}$	$9, 9$

1. SH

1\2	**D**efect	**C**ooperate
Defect	$4, 4$	$\mathbf{14}, 0$
Cooperate	$0, \mathbf{14}$	$9, 9$

2. PD

Table 1.1: Games in the two Treatment conditions

had elapsed, both subjects were directed to a second screen, again displaying the corresponding payoff matrix in Table 1.1, but now they had to choose whether to Defect or Cooperate simultaneously and without the opportunity to chat.[5]

The point of the chat portion of the design is to measure leadership, which we define to be the initiation of a plan to cooperate. Subsequent play allows us to assess the effectiveness (or lack thereof) of such initiatives. The two main treatments vary the context in which leadership activities occur. In the SH treatment, cooperation is consistent with money maximizing play, though it does require a degree of trust between the two parties. In the PD cooperation is, of course, inconsistent with money maximizing play. Thus, a leader might persuade her partner to cooperate for social motives—mutual cooperation benefits both parties—or selfish motives—the leader stands to gain more from defection if her partner can be persuaded to cooperate.

The belief elicitation phase of the OSPD is designed to track how beliefs about overall cooperativeness might change as a result of these interactions.

In the secondary treatments, the procedure for this part is exactly the same, except that individuals do not have the opportunity to chat in none of the 12 interactions. Everything else before or after the interaction is exactly the same for all the treatments.

1.3.3 Part 3: Questionnaires

Part 3 of the experiment consisted of a Cognitive Reflection Test (CRT, [40]), a Risk Aversion test ([48]), a Big 5 personality test ([50]), an Internal Locus of Control test ([68]) and a questionnaire about basic demographics (gender, major and ethnicity). From these tests, only that for Risk Aversion was incentivized. The goal of these questionnaires was to elicit traits that have been found relevant to leadership in the social psychology literature and to link them to initiative and cooperation in our setting.

Figure 1.1 shows the time line of the experiment.

We conducted 8 sessions in total from April to September of 2012 at the UC Berkeley Xlab (4 sessions with and 4 sessions without communication). 192 UC Berkeley students from the Xlab subject pool participated in the experiment. Sessions lasted approximately

[5]As before we labeled A the Defect option and B the Cooperate option. Different from before, we did not randomize the order of these options as presented to subjects to make sure they are familiarized with the meaning of each option in case they want to communicate intentions to play to the other individual. Figure A.2 in Appendix A-C shows the screens corresponding to this section.

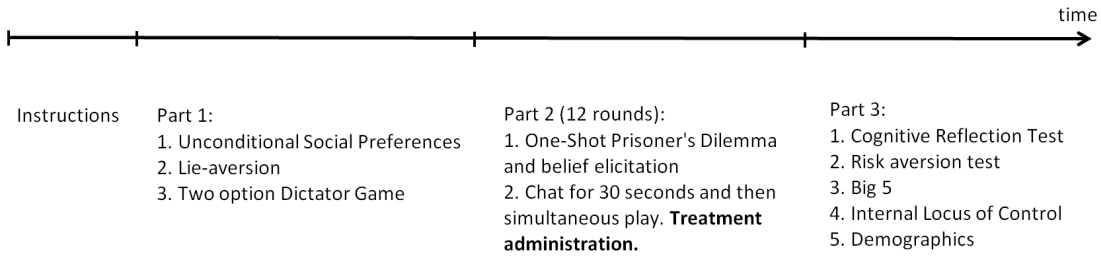

Figure 1.1: Time line experimental procedure (main treatments)

1 hour and payoffs averaged US$16. Each participant took part in only one session. All treatments were programmed and conducted using z-Tree ([36]). Throughout the experiment we sought to ensure anonymity. Participants were separated in workstations and no communication was allowed other than that feasible through the chat-box in the main treatments for the 30 seconds in each round.

1.4 Theory benchmark

It is well known that pre-play communication can affect the equilibrium behavior in normal form games [38]. Pre-play communication offers the opportunity to players jointly condition their actions on the messages exchanged rather than choosing their actions independently. In other words, communication may make the set of correlated equilibria accessible to players. In a correlated equilibrium an external mediator or "correlation device" selects a profile of actions according to an equilibrium distribution, and informs each player only its corresponding equilibrium action. The theoretical results on costless pre-play communication have sought for communication procedures or protocols by which players can replace the mediator.

In two person games the scope for communication is more limited [71]. One of the most important results is [7]. It shows that only a subset of correlated equilibrium outcomes of the normal form game coincides with the Nash equilibrium of an extended game with costless pre-play communication. In our games, these coincide with the convex combination of the pure strategy Nash equilibria in the one-shot game under purely pecuniary payoffs. That is, in the stag hunt, both mutual defection and mutual cooperation can occur; in the prisoner's dilemma, only mutual defection is predicted to happen. Proposition 1 characterizes these equilibrium actions.[6] The proof is in the Appendix A-H.

Proposition 1.

[6]The complete characterization of the equilibria with pre-play communication is showed in Lemma 1 and Lemma 2 in the Appendix A-H.

- *In the SH game with pre-play communication:*
  - *The probability of mutual cooperation* (C, C) *is* p *and of mutual defection* (D, D) *is* $1 - p$*, with* $p \in [0, 1]$*. Outcomes* (D, C) *and* (C, D) *occur with* 0 *probability.*
  - *Every time player* $i = 1, 2$ *is prescribed to play* C *she plays* C *and when player* i *is prescribed to play* D *she plays* D*.*
- *In the PD game with pre-play communication:*
  - *The probability of mutual defection* (D, D) *is* 1*. Outcomes* (C, C)*,* (D, C) *and* (C, D) *occur with* 0 *probability.*
  - *Player* $i = 1, 2$ *plays defection* D *regardless of what she is prescribed to do.*

Proposition 1 describes the outcomes that can be attained in equilibrium with pre-play communication in the games we study. Notice that in both games, communication does not serve as persuasion in this framework—only convex combination of the one-shot pure strategy outcomes can be obtained. The reason is that players cannot be convinced to do something against his or her best interests. Moreover, the identity of the communicator or the particular arguments used are of no particular consequence either. Communication merely serves to partially replace a public randomizing device. Indeed, if a sufficiently rich set of such devices were available, communication would be irrelevant.

If eventually our subjects use the opportunity to communicate to agree on any number of randomizing devices, like the clock in the computer, then mutual cooperation can be attained in the stag hunt game. Of course, none of this is of any use in the prisoner's dilemma game. Indeed, the theory (under purely pecuniary preferences) is unambiguous—there is no useful role of leadership through communication.

As predicted in the low rewards to loyalists scheme (SH treatment), communication is effective in securing cooperation. Pairs that can communicate cooperate 70% of the time, while those who cannot communicate do it only 5% of the time. In the high rewards regime (PD treatment), where communication is predicted to be ineffective, theory does poorly: 23% of the pairs make change to happen (relative to 1% without communication). In this case leadership clearly matters. Groups with emergent leaders reach mutual cooperation 31% of the time, compared to only 1% when no leadership emerges.

As a consequence, in the next section, we offer several hypotheses about the characteristics of emergent leaders. We focus our analysis on social preferences in the form of reciprocal altruism and lying aversion, although we explore some other, more traditional, constructs as well.

1.5 Hypotheses

Our results suggest that choices are virtually at odds with the theory predictions under purely pecuniary preferences in the PD game. This is not that surprising since behavior

1\2	Defect	Cooperate
Defect	$4, 4$	$\mathbf{8}, 0$
Cooperate	$0, \mathbf{8}$	$9+\rho_1, 9+\rho_2$

1. SH

1\2	Defect	Cooperate
Defect	$4, 4$	$\mathbf{14}, 0$
Cooperate	$0, \mathbf{14}$	$9+\rho_1, 9+\rho_1$

2. PD

Table 1.2: Games in the two Treatment conditions

consistent with social preferences is commonly observed in these games. In this section, we examine two social preferences formulations: lying aversion and reciprocal altruism. Under lying aversion, an individual suffers disutility from breaking his word in the event that his partner acted in good faith. Under reciprocal altruism, an individual derives intrinsic utility from repaying a good deed done by his partner.

As we will describe, adding these two aspects of preferences changes our predictions about choices in the games we study. Perhaps more importantly, differences in these characteristics define the qualities of emergent leaders.

Consider panel 1 in Table 1.1. This is our stag hunt game in which cooperation is a best response to cooperation. In this case, having a private taste for collective action does not change this. With communication, it may be a good strategy for anyone to state his or her intention to cooperate to incite collective action. Such statements may help make this equilibrium focal. Consider now panel 2 in Table 1.1, the prisoner's dilemma treatment. When the game offers incentives to remain loyal in the face of dissent, that is, defection is the best response to cooperation, high private motivation for collective action may render cooperation a best response to cooperation. If this motivation is extreme, we should expect it to induce individuals to lead through truthful speech.

In the high rewards to loyalists regime, however, not every leader will be willing to adhere. If the gains for staying loyal in the face of dissent are high, individuals with low intrinsic motivation for collective action may try to exhort others deceitfully. They work for the incumbent, in the sense they may want to expose the "bad cells" out. In the face of deceitful leaders, is there any chance speech can induce others to follow? The difference between a truthful and a deceitful leader is that lying for the former is never a problem, because he or she plans to adhere to his or her word anyway; but for the latter, lying might be a problem when he or she feels a cost of lying. We should expect that individuals who are not specially motivated by collective action do not lead if their cost of lying is high enough.

Let us be a bit more precise. Consider the monetary payoffs in Table 1.1, corresponding to the low and high reward to loyalists (in the face of dissent) schemes—stag hunt and prisoner's dilemma games, respectively. Reciprocal altruism is the private intrinsic reward for collective action. In Table 1.2, the private parameter $\rho_i \geq 0$, $i = 1, 2$ represents this extra gain. For the low reward to loyalists regime (SH), $\rho_i \geq 0$ only reinforces individual incentives for mutual cooperation. In the high reward regime (PD) however, a high enough ρ_i flip incentives from staying at home to adhering when the other does so.

If communication is costless, everyone may try to exhort others to demonstrate. Exhortation to cooperate may not be credible in this regime, unless individuals suffer a cost of lying. Assume that individuals face a private cost of lying $\lambda_i \geq 0$ if and only if: 1) individual i exhorts the partner to cooperate and 2) the partner follows suit, but individual i defects. Precisely, we could extend the games in Table 1.1 by including this non-monetary concern, as in Table 1.3. Two important considerations emerge from this thought experiment. First, a private cost of lying does not affect the incentives to cooperate if the partner cooperates in the SH treatment, but it does change incentives to cooperate in the PD if the cost of lying λ_i is high enough and i suggested cooperation. Second, communication is no longer costless, because the message (if individual suggested to cooperate) enters directly into the payoff function.

Our main hypothesis summarizes this thought experiment.

Hypothesis 1. *Emergent leaders have high reciprocal altruism and low lying aversion in the PD treatment. They have no such special features in the SH treatment.*

In the SH treatment anyone should try to lead by making mutual cooperation focal. In the PD treatment, however, leading may signal a private benefit for mutual cooperation and a private cost for lying—individuals with high lying aversion who do not care about collective action should abstain from leading. In this sense, groups in which leadership emerges should cooperate more often than those in which leadership does not emerge.

Hypothesis 2. *Collective action occurs more often when an individual leads than when no one leads, in both treatments.*

Along the same lines, we should observe that a great majority should end up demonstrating when rewards for loyalty are low. When they are high, however, only special individuals, with high reciprocal altruism and high lying aversion, should do so. This is our third hypothesis.

Hypothesis 3. *Greater cooperation occurs in the SH treatment than in the PD treatment. Moreover, in the PD treatment, individuals with high lying aversion and reciprocal altruism cooperate more often than individuals with low lying aversion and reciprocal altruism.*

Finally, suppose that the incumbent regime suppresses communication altogether. This removes the channel for coordinating in the SH and for signaling good intentions in the PD. As a consequence, cooperation should decline.

Hypothesis 4. *There should be less cooperation in both games absent communication than with communication.*

1.6 Experimental results

Having identified key non-pecuniary aspects of preferences hypothesized to affect leadership and cooperation, it remains to identify individuals having these characteristics. In this section, we first describe in detail how these characteristics may be recovered from choice behavior in the experiment and then how these characteristics relate to leadership

1\2	Defect	Cooperate
Defect	$4, 4$	$\mathbf{8} - \lambda_1 1[said\, C], 0$
Cooperate	$0, \mathbf{8} - \lambda_2 1[said\, C]$	$9 + \rho_1, 9 + \rho_2$

1. SH

1\2	Defect	Cooperate
Defect	$4, 4$	$\mathbf{14} - \lambda_1 1[said\, C], 0$
Cooperate	$0, \mathbf{14} - \lambda_2 1[said\, C]$	$9 + \rho_1, 9 + \rho_1$

2. PD

Table 1.3: Payoffs per treatment

and cooperation.

1.6.1 Constructs

1.6.1.1 Initiative

Our proxy for initiative comes from the first message sent suggesting mutual cooperation. For instance, messages such (A is defection, B is cooperation) "We both should choose B," "B and B," or "Shall we both go B" are all coded as taking the initiative. When the first message is irrelevant such as "Hi" or "Hey cutie" we do not code it as taking the initiative. Some subjects suggested defection. This occurred 0.7% of the time in the SH treatment and 5% of the time in the PD treatment. We do not code these messages as taking the initiative either. Finally, when both players suggest cooperation and their messages occur within 3 seconds of one another, we code both individuals as taking the initiative. Roughly, 14% of the games played exhibit simultaneous initiative.

1.6.1.2 Reciprocal altruism (ρ)

We estimate reciprocal altruism from the decisions in the one-shot prisoner's dilemma game (OSPD) preceding each interaction of the treatment game. Using the preference formulation for reciprocal altruism shown in Table 1.4, choices in this part of the experiment allow us to place bounds on ρ.

To be precise, let us consider the prisoner's dilemma game (OSPD) in round $t = 1, ..., 12$ with utilities as in Table 1.4.

Let us denote the two individuals in a given group by i and j. Suppose $\rho_i \geq 0$ has full support. Let us denote σ_{jt} the belief about individual j cooperating in round t from i's perspective. In the OSPD game, individual i cooperates in round t if the benefit of defection,$(1-\sigma_{jt})d_t + \sigma_{jt}a_t$, is less than the benefit of cooperation, $(1-\sigma_{jt})b_t + \sigma_{jt}(c_t + \rho_i)$. This is equivalent to say that individual i cooperates only if $\rho_i > \frac{(1-\sigma_{jt})}{\sigma_{jt}}(d_t - b_t) + (a_t - c_t) \equiv \rho_t^*(\sigma_{jt})$. Similarly individual i defects in round t if $\rho_i < \rho_t^*(\sigma_{jt})$.

$i \backslash j$	Defect	Cooperate
Defect	d_t, d_t	a_t, b_t
Cooperate	b_t, a_t	$c_t + \rho_i, c_t + \rho_j$

Table 1.4: OSPD game

From individual $i's$ decisions, we observe whether he or she cooperates or defects in the OSPD and we elicit σ_{jt}, the forecasted proportion of individuals who cooperate in round t in the OSPD game. The parameters a_t, b_t, c_t and d_t are known. Using σ_{jt} we compute $\rho_t^*(\sigma_{jt})$ for each round t. Let us denote $c_{it} = C$ if individual i cooperates in round t and $c_{it} = D$ if he or she defects, then we may deduce that, for $t = 1, ...12$,

$$\rho_i \in [\max_t\{\rho_t^* | c_{it} = C\}, \min_t\{\rho_t^* | c_{it} = D\}]. \tag{1.1}$$

Expression (1.1) indicates that a rational individual must have a reciprocal altruism parameter at least as high as the one which leaves him indifferent about cooperating under the most pessimistic beliefs, and at most as high as the one which makes him defect under the most optimistic beliefs about other's cooperating.

This assumes the model is correct, with all that entails. One aspect is that individuals are internally consistent in their choices. To examine this, we compare all the decisions made in pairs of rounds (12 rounds, 66 pairs in total) to see which pairs are consistent. By consistent we mean that if an individual chooses to Cooperate (Defect) given beliefs σ_{jt} in round t then she must choose to cooperate (defect) in round $t' \neq t$ if the beliefs $\sigma_{jt'}$ are more optimistic (pessimistic), $\sigma_{jt'} \geq \sigma_{jt}$ ($\sigma_{jt'} \leq \sigma_{jt}$).

In the SH treatment 23 of 48 (48%) individuals are rational under this model (that is all the 66 pairs of decisions are consistent). In the PD treatment, 18 of 48 (38%) individuals are fully rational assuming this primitives.

When individuals display inconsistent decisions we remove the most inconsistent (this is, the one that triggers the largest number of inconsistent pairs). We continue this elimination procedure until all the remaining choices are consistent.

We follow this procedure for every participant who has at least one inconsistent pair. After this procedure, 42 of 48 participants (88%) in the SH treatment have 10 (out of 12) or more consistent decisions, and 1 of 48 (2%) has the minimum of 7 consistent decisions. In the PD treatment 38 of 48 (79%) participants have 10 (out of 12) or more consistent decisions, and 4 of 48 (8%) have the minimum of 8 consistent decisions. We use these consistent decisions for each subject to create our measure of reciprocal altruism.

For subjects exhibiting variation in choices, this procedure yields an interval $[\underline{\rho}, \bar{\rho}]$. We take the midpoint of this interval for our estimate of ρ. Some subjects, however, exhibited no variation in choices. For these subjects, we only obtain an upper bound or lower bound on ρ depending whether they defected or cooperated in every round, respectively. For these subjects we close the interval by using the highest or lowest possible value of ρ observed for any subject in the corresponding treatment. Using this, we again choose the

midpoint of the interval. The mean ρ in the SH treatment is 16.2 (in B$) and in the PD treatment is 16.5 (in B$).

1.6.1.3 Lying-aversion (λ)

Our construct for lying aversion comes from the second block of Part 1, in which participants can send a truthful or a deceitful message to a randomly matched partner. As in [42] we classify a given individual as lying-averse if he or she sends a truthful message and if he or she declares the other would follow with at least 50% chance, but chooses selfish option in the two-option dictator game that follows. We extend this definition to include subjects who believe the partner will follow with less than 50% chance. In that case a deceitful message is intended to help ("white lie"). Those who send a "white lie" but choose the selfish option afterward are also classified as lying averse.

In sum, we code a given subject as "lying-averse" if he or she sends a truthful message or a "white lie," but chooses the selfish expected payoff in the subsequent equivalent game with the computer; otherwise he or she is coded as not lying averse. In the SH treatment 19 of 48 (40%) participants are coded as lying-averse and 29 of 48 (60%) are coded as not lying averse. In the PD treatment 15 of 48 (31%) are coded as lying-averse and 33 of 48 (69%) are coded as not lying averse by this criterion.[7]

1.6.1.4 Unconditional social preferences, personality traits and demographics

We also elicit unconditional social preferences and perform a battery of tests to measure personality traits and demographics. To classify subjects as selfish or not, we use the four choices in block one of part 1 of the experiment. A subject is coded as selfish if he or she chooses the allocation maximizing his or her monetary payoffs all four times. Otherwise, they are classified as non-selfish. Roughly 28% of the subjects are classified as perfectly selfish.[8]

The first test is the Cognitive Reflection Test developed by [40]. It consists of three questions and aims at measuring cognitive ability. Individuals were given 5 minutes to answer the test. The score is one point for each correct answer, so 0 is the minimum and 3 is the maximum. Immediately after we administered a short version of the risk-aversion test in [48]. It consisted of 4 alternatives, presented in order, in which each individual had to choose among two lotteries, one riskier than the other. (Appendix A-D shows a screen shot.) A risk averse individual should start by choosing the safe alternative and switch to the riskier alternative when it becomes attractive enough. 4 participants presented inconsistent choices (switched more than once), 3 of the 92 (4%) remaining were classified

[7]Notice there are subjects who send the truthful message or a "white lie" who choose the altruistic outcome afterward. We code these subjects as not lying averse.

[8][1] found 23% of their subjects can be classified as perfectly selfish and [37] found that was the case for 26% of their sample.

as risk loving (they never chose the safe lottery) and 16 of 92 (17%) as extremely risk averse (they never switched to the risky lottery).

We conducted the Big 5 personality test ([50]). It consists of 44 questions to characterize individuals based on 5 personality traits: Extroversion (or Extraversion), Agreeableness, Conscientiousness, Neuroticism and Openness. Each question asks for own perception of personality attributes. These attributes have been found to be strongly correlated with leadership [51].[9]

Our last test is a version of the Internal Locus of Control test developed in [68]. It consists of 13 questions, with two statements each, one indicating that people have no control over a certain hypothetical event, and the other suggesting the opposite. Individuals with a higher score on this simple test (more responses associated with the "control over events" alternatives) have been found to be high-achievers (ambitious and task oriented) and directed by own beliefs, rather than inclined to follow advice from others . Finally, we ask for demographic characteristics such as gender and ethnicity.

1.6.1.5 Constructs for agreement and cooperation

We also keep track of the instances in which an individual agrees to the leader's suggestion. After an individual takes the initiative, the matched partner can either reply by agreeing to the suggestion to cooperate, or not (say nothing, say something unrelated to cooperation, or suggest defection). We code as 1 if the former happens and zero otherwise.

Finally, cooperation is equal to 1 if the participant chooses to demonstrate (the cooperative action) and zero otherwise, in each treatment.

Table 1.5 provides a summary of the main variables just described. As the table shows, there are no statistical differences for the preference, personality and demographic variables across treatments. Initiative and cooperation do, however, differ. We explore this in detail next.

1.6.2 Initiative and types

In both games, individuals frequently take the initiative to exhort others towards collective action. Initiative, however, is significantly more pervasive when there is no

[9]Extroversion has been associated to leadership emergence [9, 43] mainly because leaders emerging from leaderless groups are more active, energetic, not silent and assertive [43]. Agreeable individuals tend to be cooperative, sensitive, altruists and to have "tact." The evidence on the direction of the relationship between leadership and agreeableness is not clear, however. Cooperativeness tends to be positively related to leadership, but sensitivity and tact are more likely to be related to modest individuals, who do not usually emerge as leaders [9]. Conscientiousness is related to task oriented behavior which is in turn associated with initiating structure. Two factors, high self-esteem and high self-confidence, are found in most of the trait based studies on leadership. These two characteristics are related to low Neuroticism [9]. Openness main components are creativity and divergent thinking, both of which have been positively related with effective leadership as well.

	SH N	SH mean	PD N	PD mean	diff	se	p
Reciprocal Altruism (ρ)	48	16.18	48	16.53	-0.35	3.95	0.93
Lying Aversion ($\lambda \in \{0,1\}$)	48	0.40	48	0.31	0.08	0.10	0.40
Selfish	48	0.63	48	0.63	0.00	0.10	1.00
InternalLocusofControl	48	6.67	48	6.25	0.42	0.48	0.39
Extraversion	48	3.11	48	3.22	-0.11	0.17	0.50
Agreeableness	48	3.65	48	3.69	-0.05	0.11	0.66
Conscientiousness	48	3.43	48	3.49	-0.06	0.14	0.67
Neuroticism	48	2.82	48	2.79	0.03	0.14	0.83
Openness	48	3.48	48	3.49	-0.01	0.12	0.93
ScoreCRT	48	1.48	48	1.35	0.13	0.24	0.60
RiskAversion	45	3.38	47	3.49	-0.11	0.21	0.59
female	48	0.73	46	0.67	0.06	0.10	0.56
Asian	48	0.71	48	0.65	0.06	0.10	0.52
White	48	0.19	48	0.23	-0.04	0.08	0.62
OtherEthnicity	48	0.10	48	0.13	-0.02	0.07	0.75
Initiate	576	0.58	576	0.41	0.17	0.03	0.00
Cooperate	576	0.82	576	0.39	0.42	0.03	0.00

Table 1.5: Summary statistics

extra benefit for staying at home while others adhere. When the rewards to loyalists are low, subjects initiate cooperation 58% of the time. Raising the rewards dampens this impulse considerably. Subjects initiate cooperation only 41% of the time. Treating each decision as an independent observation, a chi-square test strongly rejects the null hypothesis of no treatment effect (p-value=0.000). The result is unchanged if we instead treat the group as the unit of observation.

The key insight from this finding is that the rewards for loyalists, what we refer to as context, make an enormous difference in the exercise of leadership. One possible rationale for this difference is that leadership is, in a sense, less risky when rewards to loyalists are low. Cooperation is in the interest of both parties, so "persuading" one's partner to undertake this action is not terribly difficult.

The situation is more fraught when incentives to loyalists are high. It is unclear whether an overture to cooperate will be greeted warmly. Even if it is, it is unclear whether the warmth will be genuine or merely a means of trapping the leader into cooperation when the partner has every intention of defecting.

The would be leader's own motives are under scrutiny for the same reasons. Cooperation might be proposed by an individual with pure motives, out of sincere desire to cooperate—or not. Even the would-be leader may be unsure of her motives, whether, at the moment of decision she will succumb to temptation. Rather than risking this, she may prefer to refrain from leadership altogether. Indeed, this type of avoidance behavior is often seen in social dilemmas, as the pronounced absence of leaders in the prisoner's dilemma treatment may be yet another manifestation of this effect.

Apart from the question of how many leaders appear as context changes, it is equally important to ask who these leaders are; the age-old question of what aspects of a man's or a woman's character impel him or her to lead. More subtly, how do these characteristics vary with the situation? Is leadership universal or is it more a matter of the match between one's character and the problem to be confronted?

Our main result shows that it is the interplay between character and context that determines leadership.

Result 1.

Leadership is the interplay of character and context. Specifically, under high rewards for loyalty, leaders tend to be reciprocal altruists. Among non reciprocal altruists, leaders tend to have low lying aversion. In contrast, when there are low rewards for loyalty, neither reciprocal altruism nor lying aversion have any bearing on leadership.

Support for this result comes from Figure 1.2 and Table 1.6. Figure 1.2 shows the rate of initiative by treatment, lying aversion and reciprocal altruism. For exposition purposes, we label high reciprocal altruism if it is above the median, and low otherwise. In the SH treatment, initiative ranges from 53% to 61%. In the PD treatment, initiative ranges from 15% to 49%. Thus, reciprocal altruism is related to more initiative than low reciprocal altruism (48% and 32% respectively). Individuals who are predicted to care little about collective action, who also care little about lying—low reciprocal altruists and low lying averse types—take the initiative 36% of the time, whereas individuals who care

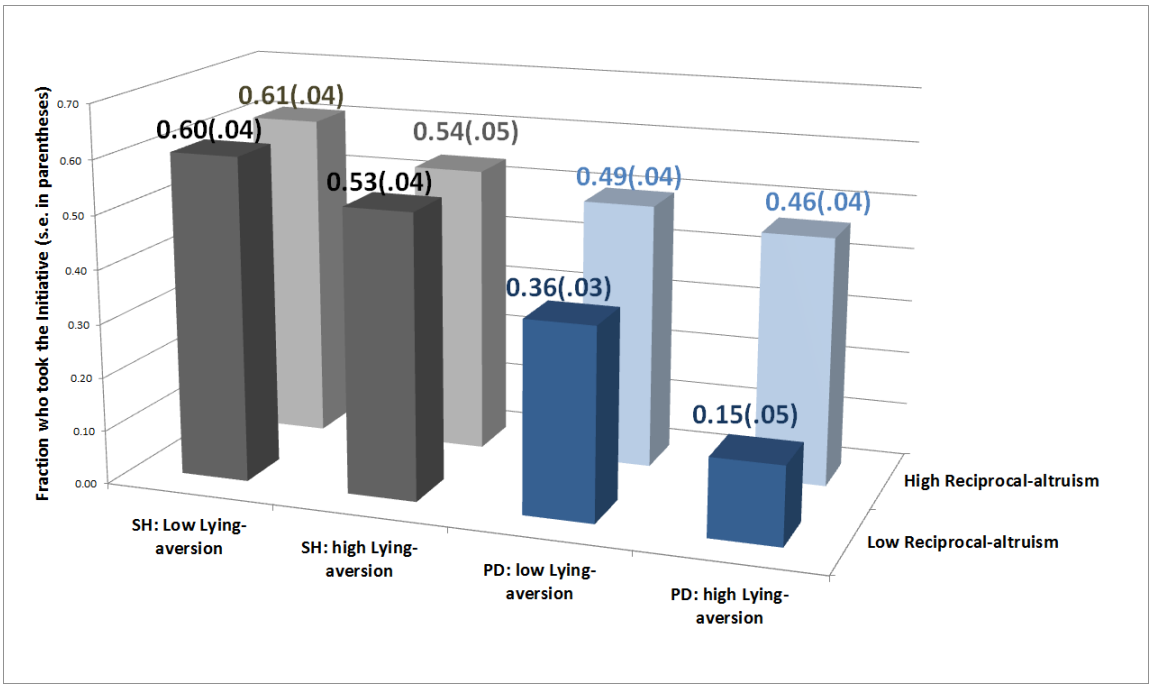

Figure 1.2: Rates of initiative by Treatment for each type (ρ, λ)

little about collective action but have high lying aversion take the initiative only 15% of the time.

Table 1.6 shows the same pattern in a regression of initiative on reciprocal altruism and lying aversion, and the interaction between them. Including the interaction is important because it may capture an effect of lying aversion for different values of reciprocal altruism on initiative.[10] Now we estimate a normal probability model and cluster standard errors at the individual level. The first two columns in Table 1.6 exhibit the results for the SH treatment. The first column shows the estimations for the parsimonious specification that includes only our constructs for reciprocal altruism and lying aversion; the second includes all the controls. Adding the controls does not make any of the point estimates of the coefficients for reciprocal altruism and lying aversion significant. Among the controls, Internal Locus of Control and risk aversion are negatively related to initiative. Ethnicity (Asian) and Agreeableness are positively related in the SH treatment.

The last two columns show the results for the PD treatment. The coefficient for lying aversion reciprocal altruism and the interaction are significant in the specification and the fourth specification. In particular, and consistent with Figure 1.2, high lying aversion individuals initiate less often on average. Reciprocal altruism, on the contrary, is positively related to initiative. To be precise, we compute the marginal effects of

[10]For instance, it might be the case that lying aversion matters for initiative only when the individual's reciprocal altruism is low, because when it is high, exhortation is likely to be truthful.

	(1) SH Pr{Initiate}	(2) SH Pr{Initiate}	(3) PD Pr{Initiate}	(4) PD Pr{Initiate}
High LA	-0.19	-0.16	-0.71***	-1.18***
	(0.28)	(0.24)	(0.15)	(0.29)
High RA	0.02	-0.23	0.33	0.44**
	(0.25)	(0.24)	(0.22)	(0.20)
RA x LA	0.00	-0.11	0.63**	1.37***
	(0.35)	(0.37)	(0.30)	(0.40)
_cons	0.26	0.75	-0.35***	-2.22**
	(0.19)	(1.72)	(0.13)	(0.91)
N	576	540	576	540
CONTROLS	NO	YES	NO	YES
pseudo R^2	0.004	0.078	0.031	0.078

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 1.6: Reduced form of initiative on reciprocal altruism and lying-aversion

reciprocal altruism and of lying aversion on the probability of initiating (at the mean of all the other covariates) from this last specification. The marginal effect of lying aversion is -15% (p-value=0.024). That is, an individual coded as lying averse is 14% less likely to take the initiative than an individual who is not, at the mean values of the covariates. The marginal effect of an individual who is above the median of reciprocal altruism is 33% (p-value=0.000).[11]

These results are consistent with our first hypothesis. Honor and non-pecuniary benefit for collective action do not characterize leaders in the low reward scheme (SH treatment); but they do matter for leadership when the incumbent highly rewards loyalists in the face of dissent (PD treatment). In one hand, individuals for whom collective action is not especially attractive, but care about honoring their word seldom emerge as leaders. In the other hand, individuals who care about positive collective action often take the initiative to exhort others to join them in demonstrating.

[11]Standard errors of the marginal effect function were estimated using the Delta method.

1.6.3 Initiative, agreement and cooperation

As expected, individual participation in revolts is more frequent in the SH than in the PD treatment. In 470 of 576 (82%) cases individuals cooperated in the SH treatment. In the PD treatment, this was the case in 227 of 576 (39%) decision instances. Looking at group outcomes, in the SH treatment in 202 of 288 (70%) game-rounds the group was able to overthrow the regime while in 20 of 288 (7%) it ended up without any manifestation (in the 23% remaining only one individual cooperated). In the PD treatment, the equivalent figures are 67 of 288 (23%) and 128 of 288 (44%), and the 33% remaining, only one individual cooperated.

1.6.3.1 Initiative and cooperation

Initiative by itself does not guarantee effective leadership; follow-through is critical. The leader's partner may remain unpersuaded or wary of the leader's suggestion. The leader herself may renege. The broader point is that initiative may not reflect leadership if it consists merely of empty words without subsequent follow through. In this section, we investigate the connection between words and deeds in our experiment. Our main finding is easily summarized—initiative increases cooperation regardless of the rewards for loyalty.

In the low reward scheme (SH treatment), initiative occurred in 94% (271 of 288) of the games, and 73% of those (198 of 271) ended up in collective action. When no initiative occurred (only 17 of 288 games), collective action occurred in 24% (4 of 17) of the games. In the high reward scheme (PD treatment), initiative occurred in 75% (216 of 288) of the games, and 31% (66 of 216) of those ended up in collective action. When no initiative occurred (in 72 of 288 games), collective action obtained in only one case (1 of 72 games). In short, leadership in the form of initiative is almost a necessary condition for collective action.

For leadership to be effective in overthrowing the regime then, individuals willing to mutually cooperate ought to believe that other's exhortation to cooperate must carry some truth; some leaders have to be willing to honor their word. Consistent with this we find that, on average, cooperation is more frequent among individuals who take the initiative than among those who do not, in both contexts.

In the SH treatment, among all the decisions in which an individual does not take the initiative, in 70% (170 of 244) of them he or she cooperates; whereas among all the decisions a given individual does take the initiative, in 90% (300 of 332) of them he or she ends up cooperating (chi-squared p-value=0.000). The pattern is similar in the PD treatment: 29% (98 of 342) of those who do not take the initiative cooperate, while 55% (129 of 234) does among those who take the initiative (chi-squared p-value=0.000). In Appendix A-E we estimate the corresponding reduced form probability model, clustering standard errors at the individual level and controlling for the individual characteristics. We summarize these two findings as Result 2.

Result 2.

A. Those taking the initiative are more likely to cooperate than those who do not.

B. Successful collective action occurs more frequently when initiative is taken than when it is not.

The point of initiative is to persuade the other party to cooperate, to follow the path suggested by the leader. In our setting, the leader has only the power of rhetoric to achieve this end. She cannot offer favors, either now or in the future, for compliance nor can she punish non-compliance. Our next result examines the effectiveness of rhetoric on changing partner's behavior.

Result 3.

When the rewards for loyalty are high, a partner is more likely to cooperate following initiative than when no initiative is taken. Under low rewards, initiative has no (statistical) effect on partner's behavior.

In the PD treatment, among all of the instances in which a given individual does not suggest cooperation, 35% (121 of 342) of the partners end up cooperating, while in instances in which a subject does suggest cooperation, 45% (106 of 234) of the partners cooperate (chi-squared p-value=0.017). In the SH treatment however, the proportion of partners cooperating after observing initiative does not change significantly. With no initiative, 82% (201 of 244) of the partners cooperate, while with initiative, 81% (269 of 332) of the partners cooperate (chi-squared p-value=0.679). In Appendix A-F we show the same result when clustering standard errors at the individual level and when we control for the covariates. In sum, this result highlights an important feature that differentiates these two contexts: In the PD treatment, initiative is a costly signal of intention to cooperate that may convince others to follow suit, while in the SH treatment, initiative is taken by almost everybody (it is not costly) because there are no conflicting interests in the decision to cooperate.

Why is speech ineffective at inducing cooperation in SH treatment? Part of the reason is that, in many instances, both players are leaders. There is no scope to persuade a party who has already been persuaded. Moreover, given how widespread leadership is in that setting, even when one does not initiate, there is a great chance that one's partner does. In short, when there is already widespread agreement about the correct course of action, there is little scope for leaders to persuade.

Taken together, our results highlight the importance of considering leadership in context rather as an abstract quality to be called upon (or not) in all situations. Leaders were rarer in the high rewards regime. Leaders also differed in their characteristics as well depending on the situation. Leadership represented an inward commitment to cooperate as well as an outward attempt to persuade. Finally, leadership proved effective in inducing collective action.

Next we turn to the another key aspect of leadership—attracting followers. Absent followers, one can hardly be said to be a leader at all. We study the characteristics and behavior of followers in the next section.

1.6.3.2 Followership

Before proceeding, it is important to be clear about how we identify followers. A person is coded as a follower if, after a leader takes initiative, the subject indicates agreement with the leader's proposition or point of view in the chat box. In circumstances where there are co-leaders, we exclude either player from consideration as a follower. This is to avoid the awkward situation where the same subject is both a leader and a follower.

The good of initiative is to persuade the partner to pursue the same course of action. We earlier saw evidence of the success of this persuasion. The decision to become a follower (i. e. to agree) seems to represent a signal that the initiators attempt of persuasion was successful. Our next result confirms that that is indeed the case.

Result 4.

Among non-initiators, followers are more likely to cooperate than non-followers.

After partner's initiative, followership occurs in 78% (259 of 332) of the games in the SH treatment and in 71% (165 of 234) of the games in the PD treatment. In the SH game, followers cooperate 86% of the time while non followers do it only 62% of the time (chi-square p-value=0.000). Likewise, in the PD game, 58% of the followers cooperate, while only 16% of the non-followers do so (chi-square p-value=0.000).

Next, we compare the credibility of the promises made by leaders and followers. We saw that both roles tended to stand by their word more than a neutral comparison group, but how do they compare with each other?

Result 5.

A leader is more likely to keep her word and to cooperate than a follower.

We consider the games in which one individual takes the initiative and the other follows (i.e. agrees). In those cases, 95% of the leaders and 83% of the followers end up cooperating in the SH game. In the PD game, the account is similar: 65% of the leaders and 57% of the followers adhere.

What accounts for this difference in behavior? To examine this, notice that followers represent a subset of all of those who do not take initiative. From Table 1.6, we observe that leaders should be more reciprocal altruists and less lying averse only in the PD game. The results in Table 1.7 are somewhat consistent with this. The difference in reciprocal altruism between leaders and followers, is positive although significant at conventional levels only for the proportion of high reciprocal altruistic types in the PD treatment. Followers are more lying averse than leaders, although barely insignificant in the PD treatment.[12]

1.6.3.3 Agreement and cooperation

Finally, we turn to the effects of agreement on collective social action. We classify an interaction as leading to agreement if: a) Both parties initiated; or b) One party initiated

[12]Tables A.3 and A.4 in Appendix A-G present the t-test for difference in means for all the characteristics. Leaders are slightly more Agreeable and Conscientious in the SH and in the PD respectively.

	Leader N	Leader mean	Follower N	Follower mean	diff	se	p
SH							
Reciprocal altruism	174	16.30	174	16.63	-0.33	2.13	0.88
High reciprocal altruism	174	0.51	174	0.50	0.01	0.05	0.75
Lying Aversion	174	0.40	174	0.44	-0.04	0.05	0.45
PD							
Reciprocal altruism	143	21.35	143	17.91	3.44	2.51	0.17
High reciprocal altruism	143	0.67	143	0.53	0.14	0.60	0.02**
Lying Aversion	143	0.31	143	0.41	-0.09	0.06	0.11

Table 1.7: t-tests, difference in means between leaders and followers, by treatment

and the other party become a follower. Is this "informal agreement" related to mutual cooperation?

Result 6.

Agreement increases the chances of cooperation in both games.

In the SH treatment, cooperation obtains in 43% (46 of 106) of the instances if agreement is not reached, but in a 90% (424 of 470) if agreement obtains (chi-square p-value=0.000). For the PD treatment, the pattern is similar, although perhaps more salient. Individuals in groups that do not observe agreement after the communication round cooperate in 13% (33 of 254) of the cases, while after agreement they do it in 60% (194 of 322) of the cases (chi-square p-value=0.000).

1.6.3.4 Types, agreement and cooperation

In the high rewards regime, initiative is related to reciprocal altruism and lying aversion. If different types display different content in their messages, then we may risk to confound the effect of types and content on cooperation. Moreover, these differential content can manifest itself in a dialog differently depending on the content the partner provides to the conversation. We do not attempt to explain these difference here, rather, we aim at providing correlations that hint that types matter for cooperation only in the PD treatment. In effect, in order to minimize this endogeneity problem, we provide evidence on cooperation by groups.

Result 7.

Reciprocal altruism and lying aversion positively correlate with cooperation only in the PD treatment.

Table 1.8 shows a normal probability model of both individuals cooperating on the characteristics of the members of that group. The variables used as control are the number

	SH Pr{Both C}	PD Pr{Both C}
Group reached agreement	1.58***	1.65***
	(0.27)	(0.33)
Group average reciprocal altruism	0.00	0.04***
	(0.01)	(0.01)
At least one lying averse in group	-0.04	0.61**
	(0.21)	(0.30)
_cons	0.99	-11.04***
	(2.80)	(2.74)
N	273	286
CONTROLS	YES	YES
pseudoR^2	0.357	0.460

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.8: Group members' reciprocal altruism and cooperation, by treatment

of individuals in the pair who are selfish, Asian or White and the average score of the pair in the Internal locus of control, big 5, Cognitive Reflection Test and risk aversion measures. The first column exhibit the results for the SH treatment, and the second for the PD treatment.

We find that reciprocal altruism and lying aversion are related to mutual cooperation only in the PD treatment. Support for this result can be seen in Table 1.8. As we saw before, in both treatments, agreement is positively related to mutual cooperation. In the SH treatment, the effects of reciprocal altruism and lying aversion are not statistically different from zero, while in the PD treatment, lying aversion and reciprocal altruism are significantly related to cooperation. This is only suggestive evidence, however, because agreement is endogenous.

These results are in line with our third hypothesis: 1) Cooperation is more common in SH game than in the PD game and 2) reciprocal altruism and lying aversion are related to cooperation only in the PD treatment.

1.6.3.5 Banning Communication

We have seen that some countries, such as Iran or China, web based social networks have been intervened or directly banned. Companies have also freedom in deciding the

Cooperate	SH-Chat	SH-NoChat
No one	20 (7%)	172 (60%)
One	66 (23%)	101 (35%)
Both	208 (70%)	15 (5%)

a.

Cooperate	PD-Chat	PD-NoChat
No one	128 (44.4%)	237 (82.3%)
One	93 (32.3%)	47 (16.3%)
Both	67 (23.3%)	4 (1.4%)

b.

Table 1.9: Cooperation with and without communication, by treatment

type of communication they want to encourage between individuals. Our last set of results shows that when communication is not allowed, collective action decreases. In the SH treatment without communication, 23% (131 of 576) of the time individuals cooperate (with communication this proportion is 82%). Cooperation occurs only in 10% of the individual decisions in the PD treatment without communication (the proportion with communication is 39%).

Result 8.

Collective action hardly occurs absent communication in both treatments.

Table 1.9 presents evidence to support this result. The first columns in panel a. and b. replicate the results for the SH and PD treatments with communication. The second columns show the results for the treatments without communication. In the SH game, mutual cooperation reduces from 70% with chat to 5%; in the PD treatment, mutual cooperation decreases from 23% to 1% (chi-square p-value=0.000 for each of both tables).

As expected, these result support our fourth hypothesis that communication facilitates (and is almost a necessary condition for) collective action in both games.

1.7 Conclusion

In leaderless groups, coordination requires a special member to take the initiative to foster cooperation. We find that, in groups without formal authority, the traits that matter for leadership depend on context. When the incentives to remain loyal when others dissent are low, no special talent is required from leaders—anyone initiates and anyone follows suit. When the rewards to loyalists are high, leaders are indeed special. Individuals who exhort others to demonstrate out or to push for change have higher intrinsic value for mutual cooperation than those who remain silent. Thus, this intrinsic value interacts with lying aversion. If individuals have low intrinsic motivation for mutual cooperation, then high lying aversion makes them less likely to initiate. If this intrinsic motivation is high, high lying aversion is positively associated with leadership. As an implication, if the members of a society or a firm consist of mutually cooperative individuals, the chances of demonstration and change are higher than when those members do not specially care about mutual cooperation—no matter how honest they are. A special taste for mutual cooperation provides the intrinsic incentives to push for change, a necessary condition for change to occur.

Moreover, governments seeking to undermine social movements, or companies attempting to hinder innovation should put obstacles to communication, regardless of the current policy of rewards to loyalists in the face of dissent. Communication allows the emergence of leaders who signal inward commitment in order to persuade others to implement change.

From this study we may learn that communication leads to coordination when a leader steps up. In this sense, this analysis contributes to unpack the black box of communication by providing evidence on: 1) the interplay of leaders' characteristics and the context they are immersed in and 2) the role of initiative-taking (and agreement) on collective action.

Chapter 2

Social Preferences and Collusion: Experimental Evidence on the Responses to Relative Performance Pay

2.1 Introduction

Workers are often incentivized based on relative performance. This can come in the form of explicit relative pay such as bonuses or performance tournaments. Alternatively, incentives can be implicitly provided through the opportunity of promotion or advancement within an organization. A large literature, starting with the work of [55] and others, sheds light on the benefits and drawbacks of this kind of compensation relative to fixed and piece rate pay. One problem with relative incentive pay is its proneness to collusion by workers. For instance, [6] use personnel data to show that when compensation was based on relative incentive pay compared to piece rate pay, fruit pickers that are able to monitor each others' performance had 1/3 less output. When performance could not be monitored, however, both incentive schemes fared equally well. For fruit pickers working alongside "friends" this effect was strongest. This last finding raises the possibility of another potential culprit of thwarted performance schemes: social preferences. However, social preferences can also possibly help performance in the workplace. In a related paper, [58] find grocery store checkers increase their productivity when they are observed by more productive workers. Interestingly, this effect is intensified when the worker is being watched by someone with whom they have regular contact. Thus, increased productivity could be coming solely from peer pressure but it could also be motivated by "reciprocal altruism," as those with more frequent contact have more intense opportunity for reciprocal prosocial behavior (see [35] and cites therein for this notion of social preferences). In other work settings, it could also be the case that some workers are unconditionally altruistic: workers internalize the negative externality imposed on others under relative incentives and the positive externality provide in team production. In fact, this force could also be at work in the past two mentioned studies. However, due to not being able to measure unconditional social preferences, any such effect was instead averaged across outcomes. Whether non-competitive efforts come from cooperating through collusion or social preferences provides different implications for a principal or policy maker. In the case of the former, making it more difficult for agents to monitor or punish one another will stem non competitive efforts. For the latter case, however, it could instead mean screening out workers with certain preferences. We approach the problem of the impact of different social preferences on performance outcomes by focusing on the case of relative incentives. Unfortunately, in the field, it is difficult to separate social preferences and selfishly motivated cooperation. Therefore, we turn to the laboratory to explicitly disentangle these forces. In such a setting, we can distinguish between classical collusion (i.e., due to purely selfish coordination) and low, non-competitive efforts arising from social preferences. To this end, we conduct an experiment where we measure social preferences via dictator menus a la [1] and then link these different preferences back to behavior when facing relative performance schemes. As theory suggests, we find that social preferences quite generally decrease the effort invested under relative incentives. Across all of the treatments, a subject we classify as other-regarding decreases effort by about 15% relative to a subject we classify as selfish. In addition to individual social preferences, we

also find some evidence that the composition of a group is important: One additional other-regarding group member decreases a subject's own effort by about 12% on average when communication is possible. Thus, also interaction effects between group members' social preferences seem to be important. To further explore how social preferences can limit competitive efforts, we consider the forces of communication and observability of effort, which are identified as crucial ingredients for classical collusion. First, we assess the extent to which non-competitive efforts arise in a communication environment (i.e., via chat) and in an observability environment. We find ample evidence of systematic effort reduction. Consistent with the literature on coordination and communication [21], the availability of communication reduces efforts by some 50%. In many groups we observe convergence to minimum effort or alternating minimum effort. In fact, 62% of the groups can be classified as coordinating on noncompetitive, minimum efforts. By examining the chat messages of subjects we can identify leaders who proposes non-competitive efforts to the group. Interestingly, leaders who suggest the strategies for such outcomes tend to be selfish. Thus, we find that when communication is possible, groups that are most effective at decreasing efforts contain one selfish member suggesting collusion and 2 other regarding group members following. When communication is not possible, on the other hand, collusive outcomes are much rarer, regardless of observability of effort. In these settings, groups consisting solely of other regarding subjects are most successful at coordinating on low(er) efforts. In an attempt to turn off, or at least mute social preferences, we conduct an additional treatment, where subjects face computerized players instead of human subjects. The computerized subjects are programmed using the actual behavior of past human subjects. In this treatment, a subject's behavior does not have any consequences for other subjects' payoffs, and therefore we expect the subjects' behavior to be the same whether we categorize them as selfish or other-regarding. We indeed find that behavior of all social preference types is indistinguishable by the end of the game, providing further evidence to our earlier results. We also study the relationship between gender and social preferences and subjects' decisions. Consistent with earlier studies, females are more often other regarding than males. Therefore, studies controlling for gender but not social preferences in the analysis of behavior may be unwittingly using gender as a noisy proxy for social preferences. In our study, conditional on one's social preferences, we only find a gender effect in the communication treatment: males are disproportionately more likely to emerge as leaders suggesting non-competitive behavior to their group. The paper is organized as follows. In Section 2 we review the relevant literature. In Section 3 we describe our experimental design. Section 4 provides our results. In Section 5 we conclude.

2.2 Literature

Our paper is most closely related to [6]. As discussed above, they find effort is depressed under the relative performance scheme. This suggests that under relative incen-

tives workers at least partially internalize their externality imposed on their co-workers. This raises the question of how much of the effect is due to selfish collusion with the threat of punishment, and how much of it is due to other regarding preferences. Our analysis complements these findings by directly testing that question. We implement a similar relative incentive scheme to [6] in the laboratory where we can measure individual social preferences. Our design allows us to answer how much of non-competitive efforts is driven by social preferences and how much is purely due to "classical collusion." In order to measure social preferences and categorize subjects we draw on work by (hereafter AM) as well as [37] (hereafter FKM). These papers analyze the rationality of individual giving behavior in dictator games under different budget sets. They find that giving choices of most subjects are consistent with the generalized axiom of revealed preferences [74] and can be represented by a CES utility function. We adopt the categorization of AM of subjects into three groups: Selfish, Substitutes (Utilitarian) and Complements (Rawlsian). This categorization is also consistent with the model proposed by [19]. Another paper our work is closely related to is [33]. They experimentally study players who compete for relative performance pay and then allocate (some) of their earnings to their competitors. First place performers are less likely to give to their group members compared to lower performers. They thus suggest other regarding players are more likely to exert less effort. Our work complements this paper in that we directly test this notion. In particular, we categorize players according to social type before they compete. Thus, we remove any potential confounds of income effects, competitive preferences, reciprocity, or any other effect of one's competitive experience and opportunity to compensate one's competitors ex-post. Additionally, we explore the role of communication and leadership, which then can become an avenue for more selfish players to exert less effort. Finally, we also consider how the group composition of social preferences shapes competitive outcomes as opposed to only individual preferences. Other papers considering social preferences' relation to cooperative behavior include [35, 72, 52, 31, 29]. The latter paper is closest in spirit to our analysis. [29] relates social preferences, elicited through a two person dictator game, to cooperative behavior in an infinitely repeated, noisy prisoner's dilemma. They do not find any significant effect of social preferences on cooperative behavior. Our work differs in that we measure social preferences before and not after the "cooperation" stage. In addition we provide a game environment in which cooperation can evolve more gradually (effort choices between 1 and 12 versus "cooperate" and "defect"). Lastly, and importantly, we do not have noisy output in our game, which makes any conditional strategy more viable. In contrast to [29], we do find that social preferences predict cooperative behavior under relative incentives. More generally our paper also speaks to the literature of consistency of individual social preference types across different economic settings. Some examples in this vein include [27] and [11]. However, the question this literature is interested in is whether individual social preferences are stable across a range of economic games. The assumption underlying our research is that the subjects' preferences are stable between the elicitation and the effort provision stage. If this is not the case, we should not find any significant relation between behavior and types. Thus our experiment

Type	Preferences
Selfish	π_i
Complements (Rawlsian)	$\min_{j \in N}\{\pi_j\}$
Substitutes (Utilitarian)	$\Sigma_{j \in N}\pi_j$

Table 2.1: Overview of social preference types. π_i represents the pecuniary payoff of individual i (self), N denotes the set of individuals, in our case $\{1, 2, 3\}$.

provides additional evidence on this fundamental question.

2.3 Experimental Design

In total, we conducted 8 experimental sessions with 168 subjects. Participants were students from UC Berkeley, enrolled in the X-lab subject pool. Sessions lasted approximately 60 minutes from reading instructions to subject payment, which averaged approximately $16 per subject. Participants were not allowed to take part in more than one session. The treatments were programmed and conducted using *z-Tree* developed by [36].

We had the double purpose of identifying people's social preferences and measuring their choices when facing a relative performance incentive scheme. In order to achieve this, the experiment was divided into three stages. In the first stage, we randomly matched subjects into anonymous groups of three individuals. Participants were then given 100 tokens each for 9 periods and played a dictator game with their group members (including themselves). In each period participants faced different "prices" or token exchange rates of giving to each group member. Prices varied such that we could both identify individuals' willingness to give to others and individuals' willingness to give between others when facing different prices of giving.[1] We use these 9 periods to classify our subjects. In periods 10 and 11 we conducted allocation decisions with positive sloped budget sets as in AM where subjects are given an allocation and decide on the overall exchange rate. We will use these decisions as robustness controls to see whether aversion to disadvantageous inequality matters also separately.

Subjects did not learn their other group members' choices to avoid uncontrolled learning. Participants were told that for 5 out of a total of 11 allocation decisions one of the group members' choices would be randomly selected to compute payoffs.

We use this first stage, in particular decision 1 to 9, to classify participants as "Selfish", "Complement" (Rawlsian) or "Substitute" (Utilitarian), as shown in table 2.1.[2] To do so,

[1] FKM uses a slightly different nomenclature to describe distributional preferences. They call *preferences for giving* the fundamentals that rule the trade-off between individiual and others' payoffs and *social preferences* the ones that govern the allocation between others. Our study does not attempt to dwell on that distinction, therefore we endorse the more traditional terminology: We use "social preferences" or "other regarding concerns" indistinctly to represent non-selfish behavior.

[2] From now on we use the capitalized form of selfish, complement, substitute and other-regarding to

we first compute the relative giving rates of an archetypal Selfish, Utilitarian and Rawlsian individual according to the preferences in Table 1. We denote player i's monetary payoff as π_i and the total number of players n.Thus, an archetypal Selfish type, is only interested in his own monetary payoff. In contrast, an archetypal Rawlsian player only values the minimal monetary payoff of all of her group member's payoffs. Finally, an archetypical Substitute simply maximizes his group's total monetary payoff.

To categorize subjects, we then measure the (Euclidean) distance of each of the participants' decisions to these archetypes' decisions. We compute such distance for each choice and then we compare the average distance across periods to each archetype's decision. We classify subjects as the archetype whose decision is closest to the subject's decision.[3] Consistent with AM we classify 22% of subjects as (perfectly) Selfish, whereas AM find 23% of subjects are perfectly Selfish. 6.5% of our subject can be classified as perfect Substitutes, while AM find 6.2%. In contrast to AM we only classify one subject as a perfect Complement, while they have 14.2%. Different from AM, we do not have any "weak" Selfish types, as we categorize all other regarding subjects (i.e., subjects that give to others) as either Complement or Substitute types.[4]

For the second stage, participants were again randomly matched with two other players for the remainder of the experiment. They participate in a relative performance game modeled after [6]. The purpose of this stage was to give players the possibility to collude by jointly providing low levels of effort. Thus, we simulated an infinitely repeated game with continuation probability of $\delta = 95\%$. In order to gain consistency across treatments, we randomly drew the number of periods before running the sessions as in Fudenberg, Rand and Dreber (Forthcoming).

We also varied factors considered important for creating and sustaining collusion. In particular, in the first treatment ("Chat and Observability") we allowed chat via computer terminals *during* each period and observability of choices and payoffs *after* every period. In the second treatment ("Observability") we did not allow for chat but continued with observability after each period. In the third treatment ("No Observability") neither chat nor observability was allowed. In this treatment, subjects only learned their own payoff after each period. Thus, this treatment serves as a baseline to identify effort levels when coordination is not reasonably possible.

If we were able to mechanically switch on and off subject's social preferences, we could directly identify the effect of social preferences on effort. Unfortunately, this is not generally possible. We conducted a final treatment where we approximate this idea.

refer to our categorization. Thus we do not imply that a subject we categorize as selfish necessarily always acts in a selfish manner, but only that given our three categories, he or she most closely resembles this type.

[3]As we only use relative giving rates between the other two group members our classification does not account for the intensity of social preferences. We can control for intensity separately by including the overall giving rate of a subject.

[4]We also analyze an alternative specification of allowing imperfect Selfish subjects and find the results are qualitatively similar.

Treatment	Subjects
Chat & Observability	63
No Chat & Observability	63
No Chat & No Observability	21
Robot	21
Total	168

Table 2.2: Summary of Treatments

Instead of facing human subjects, a subject played against the computer, which simulated the play of past subjects' decisions ("Robot"). This treatment attempted to "switch off" social preferences by making it clear to subjects that even though they faced the same consequences for their choices as if playing human subjects, their effort decisions no longer affected any person's payoffs.

Table 2.2 provides a summary of these treatments.

A subject's payoff was calculated as follows. Note these figures are in Berkeley Bucks \$, converted at \$66.6 Berkeley Bucks to 1 US\$, as this is how it was presented to subjects. Each participant received an endowment of \$12 (Berkeley Bucks \$) each period from which they could choose costly effort. Effort costs \$1 for each unit of effort. Total payoff was then

$$\pi_i = 12 + \frac{x_i}{\overline{x}} 15 - x_i$$

where $\bar{x} = \sum x_j/3$ is the average effort across i´s group and i chooses effort $x_i \in [1, 12]$.[5] Hence, each participant's effort is discounted by the average effort, so a higher average effort will reduce payoffs, ceteris paribus. This is the relative performance evaluation similar to [6]. The stage game (or one-shot) Nash equilibrium for homogeneous and selfish (and risk neutral) players is to play $x_i = 10$ for all i, which is below the upper bound of the action space. Coordinating on $x_i = 1$ is sustained by a continuation probability $\delta > 60\%$ (optimal deviation from Pareto Dominant outcome is to play $x_i = 7.5$). Therefore, our $\delta = 95\%$ should support collusion for utility maximizing rational, risk-neutral selfish agents. For the final stage, subjects were again given the same allocation price menus as in the first stage. Critically, subjects did not know they were going to have this final allocation opportunity. Instead, they were told at the beginning of the experiment they would have a final stage with some additional opportunities to increase their payoffs. We do not consider this data in this study. After the allocation decisions, subjects completed a risk aversion test a la [48], and a basic demographic questionnaire. We note we did not attempt to elicit beliefs. However, soliciting beliefs in an infinitely repeated game would have only provided updated beliefs and possibly contaminated our results. In addition, our robot treatment provides a benchmark comparison for what happens when subjects do

[5] Although subjects were not told to do so, almost all entered effort choices as an integer. We had an effort lower bound of 1 to create an upper bound for payoffs. The effort upper bound of 12 came from the periodic endowment of \$12.

not have the belief they are benefiting or hurting subjects by their own strategic decisions. We now turn to our hypotheses.

2.4 Experimental Hypotheses

In this section we informally derive testable hypotheses regarding the effect of social preferences on effort provision in the relative performance game. In order to illustrate the main trade-offs, we consider a subject's optimal response to his group members' efforts in a one shot interaction. Assume a player believes his group members efforts are x_{1o} and x_{2o}. The following best responses result:

Selfish A Selfish subject's best response to his group members' efforts is equal to $x_i = max1, \sqrt{3w(x_{1o} + x_{2o})} - (x_{1o} + x_{2o})$. If the other group members put in the minimal effort of 1, his optimal response (in a one-shot interaction) is to put in $x_i = \max\left\{1, \sqrt{6w} - 2\right\} = \sqrt{15 \times 6} - 2 = 7.5$. Thus, absent any long-run incentives, a commonly low-effort will not be sustained if there is a Selfish group member.

Complements A Complement type by our definition maximizes the payoff of the worst-off group member. As we assure by the choice of parameters in our experiments, if total wages weakly exceed total cost of efforts, the player with the lowest effort receives the lowest payoff. Thus a Complement's best response is to put in $x_i = \min\{x_{1o}, x_{2o}\}$ which equalizes his or her pecuniary payoff with that of the worst-off group member. [6] This implies that a Complement individual constantly faces a coordination problem.

Substitutes A Substitute subject maximizes the (weighted) sum of the group members' utilities. Because his effort decreases the other group members' utilities, he faces an additional cost of effort relative to a Selfish individual. Thus relative to a Selfish type, a Substitute depresses effort. In fact, if he puts sufficient weight on his group members' utility, a Substitute always chooses minimum effort, regardless of x_{1o} and x_{2o}. This includes the case of equal weights, as given in Table 1.

In a one shot game, it is an equilibrium for groups of Complements and Substitutes to coordinate on minimal efforts of 1, whereas this is not the case for a group with at least one Selfish player. Of course, when we move to an infinitely repeated game, even a group of all Selfish players can sustain an equilibrium of all providing efforts of 1. However, in practice, such an outcome will likely take time. Meanwhile, it is still an equilibrium for these Selfish players to play the high-efforts of a stage Nash equilibrium. Hence, on whole,

[6] Strictly speaking we require $\sqrt{3w(x_{1o} + x_{2o})} - (x_{1o} + x_{2o}) \geq \min\{x_{1o}, x_{2o}\}$. For $\sqrt{3w(x_{1o} + x_{2o})} - (x_{1o} + x_{2o}) < \min\{x_{1o}, x_{2o}\}$ a Complement behaves identical to a Selfish and chooses $x_i = \max\left\{1, \sqrt{3w(x_{1o} + x_{2o})} - (x_{1o} + x_{2o})\right\}$.

it seems likely Selfish players will provide more effort on average and have a harder time coordinating on low efforts compared with other regarding players. This intuition results in our first hypothesis.

Hypothesis 1. *Subjects with Other Regarding preferences exert less effort on average than Selfish subjects*

We are assuming the canonical Other Regarding model where a player maximizes peoples' net payoffs and not utility per se. Hence, with this assumption, we can also make a prediction about the composition of group social preferences and effort choices.

Hypothesis 2. *Average group effort choices are increasing in the fraction of group members that are Selfish.*

Another issue we explore is leadership. Infinitely repeated games like ours typically allow for a multiplicity of equilibria, at least for the standard case of perfectly Selfish individuals, and there is no reason a priori to consider the Pareto optimal outcome as the focal one. However, as [53] points out, a leader could determine what equilibrium to play. Once the equilibrium is established, there is no incentive for anyone to deviate. Our first treatment allows us to consider leadership in a controlled environment. We analyze the chat messages of the subjects to identify leaders and relate them to outcomes. In particular we are not only interested in whether communication generates cooperative outcomes as in [21], but also how this relates to social preferences. Note the best possible payoff, even for a Selfish player, is for all players to coordinate on minimal effort of 1. For Other Regarding players, there is already an incentive to depress efforts through social preferences. However, for Selfish players, their natural tendency is higher efforts. If players believe their opponents are of their own social type (e.g., see[49]), the presence of a coordination device—e.g., communication—is particularly valuable to a Selfish player. In addition, as outlined in their best responses, Complement types also have a coordination problem. Hence, we conjecture this makes it more likely either a Complement or a Selfish agent will attempt to lead a group.

Hypothesis 3. *When communication is available, Selfish and Complements are more likely to become leaders.*

We now turn to our empirical analysis to test these hypotheses.

2.5 Empirical Analysis

2.5.1 Examples of Decisions

We begin with some examples of actual giving and effort rates of particular groups to illustrate subjects' behavior. Figures 1 to 4 illustrate the patterns of decisions across time. In the first stage (periods 1 to 9), we can observe the number of tokens each player in the group keeps for him or herself. In the second stage, (periods 12 to 40) we

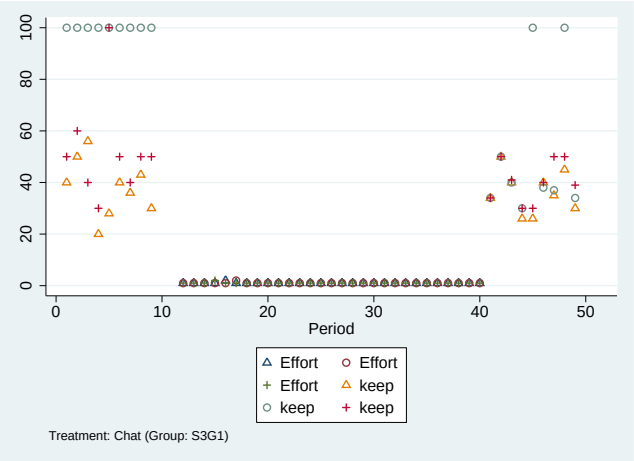

Figure 2.1: An example of non-competitive efforts in Treatment 1.

observe the choice of effort ranging from 1 to 12.[7] In the third stage, (periods 41 to 50) we again observe the number of tokens subjects kept for themselves, but this time the allocation counterparts are the other two members of their respective groups from stage 2. From Figure 2.1 we observe heterogeneous patterns of keeping in the first stage: One subject keeps everything to himself, while the others share almost equally. Perhaps more interestingly, it provides an example (Session 3, Group 1) of "perfect collusion" in the chat treatment: Subjects coordinate on minimal effort during the whole second stage.

Figure 2.2 shows another group from the chat treatment (Session1, Group 5). In this case, behavior in the second stage is surprising: Participants play a strategy that is not the Pareto-dominant one. Subjects alternate between providing maximal and minimal effort. In each period a different subject reaps the rents of outperforming the other subjects With the help of the chat, they perfectly coordinate on this synchronized play. Although this does not allow the subjects to reach the maximal group payoff, this form of collusion still leads to relatively high payoffs.

Figure 2.3 shows an example (Session 5, Group 3) of successful collusion in the No Chat/ Observability treatment. Subjects very slowly coordinate on lower efforts. Also in this group the giving rates are highly heterogeneous.

Our last example, Figure 2.4 illustrates failed collusion. In this group from the Chat treatment (Session 1, Group 3) subjects almost in all rounds choose the maximal efforts. Only one subject tries to deviate from this strategy once, without success. Note that two of these subjects are perfectly Selfish and keep everything to themselves.

[7]Notice that period 10 and 11 are not described in the experimental results. In these periods we elicited the propensity to act spitefully by destroying output to reduce inequality. We use this as a control variable only and do not find any additional predictive power.

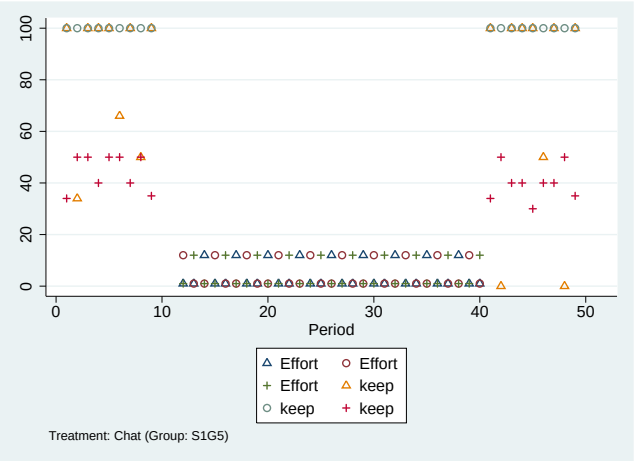

Figure 2.2: An example of a form of non-competitive efforts in Treatment 1.

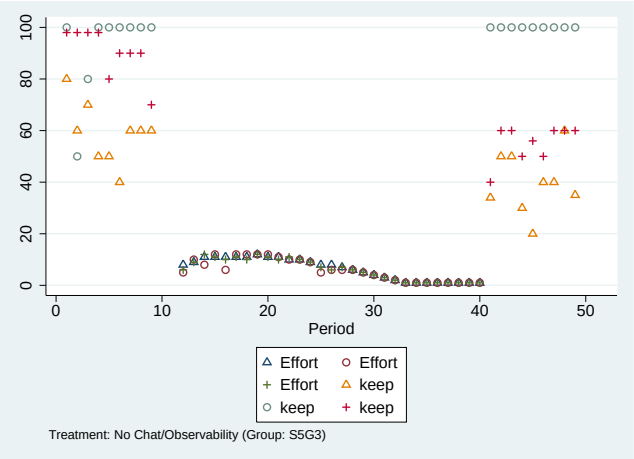

Figure 2.3: An example of non-competitive efforts in Treatment 2.

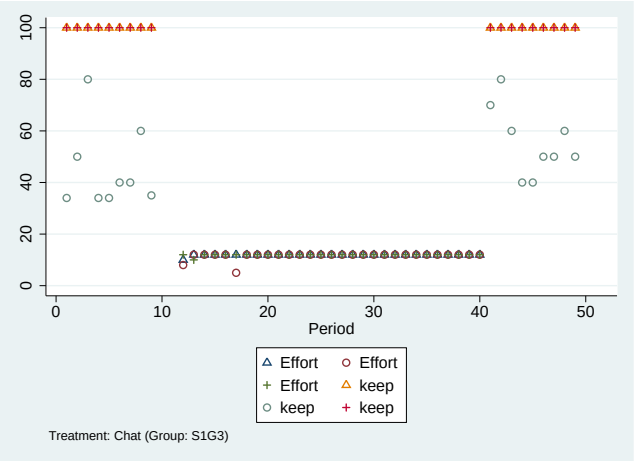

Figure 2.4: An example of competitive efforts in Treatment 1.

2.6 Categorizing Social Preference Types from Giving Menus

Table 2.3 summarizes the mean choices of our subjects under all 9 price vectors in treatments 1 to 3. We will analyze treatment 4 in section 2.9.

We see that regardless of the price of giving, subjects keep on average around 70% of their endowment. Using these choices, we sort our subjects into social preference type categories as described in Section 2.3. Figure 2.5 shows the distribution of social preference types in our subject population in treatments 1 to 3 (Chat/Observability, No Chat/Observability and NoChat/No observability). Most of the participants (63%) are categorized as Complements (or Rawlsian) since their giving behavior aims at equalizing payoffs across the members of their groups. Selfish and Substitutes (Utilitarian) are almost equally frequent in our sample, with 20% and 17% of the total, respectively.

Figure 2.6 shows the distribution of Selfish subjects across groups. Since subjects were allocated randomly and Selfish subjects are relatively rare we do not observe groups with only Selfish group members in treatments 1 and 2. Otherwise we do observe random variations across groups in the number of Selfish subjects which we will use to identify the effect of group composition in the next sections.

Figure 2.7 illustrates giving behavior under our categorization of social preferences types. We see that Selfish types never give anything to their group members. In contrast, Substitutes and Complements give positive amounts on average for every price vector. When the price of giving increases, Substitutes typically react by decreasing their giving rate, while Complements do the opposite. This is most easily seen for periods 6 to 9 where the price of giving to individual 2 is always lower than the price of giving to individual 1. Thus, as archetypal types would do, Complements react by allocating more to individual 1 while Substitutes react by allocating more to individual 2.

Period	Price vector	Keep (min, max)	Give to 1	Give to 2
1.	$(1, 1, 1)$	70.66 (33,100)	15.21	14.13
2.	$\left(1, \frac{1}{2}, \frac{1}{2}\right)$	73.39 (0,100)	13.24	13.37
3.	$\left(1, \frac{3}{4}, \frac{3}{4}\right)$	71.82 (0,100)	13.98	14.20
4.	$\left(1, \frac{5}{4}, \frac{5}{4}\right)$	72.29 (20,100)	14.13	13.59
5.	$\left(1, \frac{3}{2}, \frac{3}{2}\right)$	71.03 (20,100)	14.67	14.30
6.	$\left(1, 1, \frac{2}{3}\right)$	71.80 (0,100)	15.90	12.30
7.	$\left(1, 1, \frac{3}{4}\right)$	73.46 (0,100)	15.03	11.51
8.	$\left(1, \frac{3}{4}, \frac{1}{2}\right)$	77.09 (0,100)	12.33	10.58
9.	$\left(1, \frac{5}{4}, \frac{3}{4}\right)$	72.72 (0,100)	16.18	11.10

Table 2.3: Giving Rates.

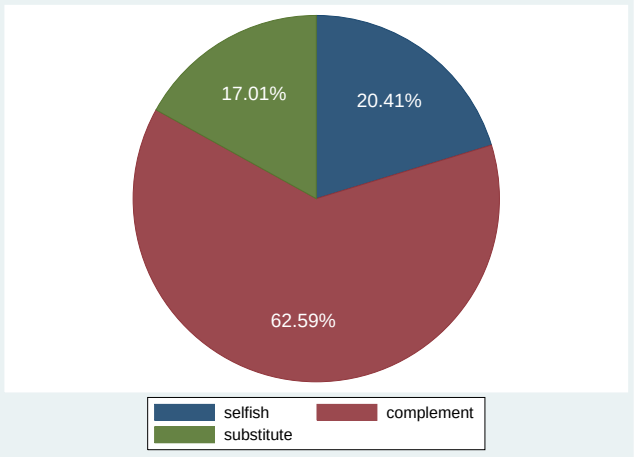

Figure 2.5: Distribution of social preferences.

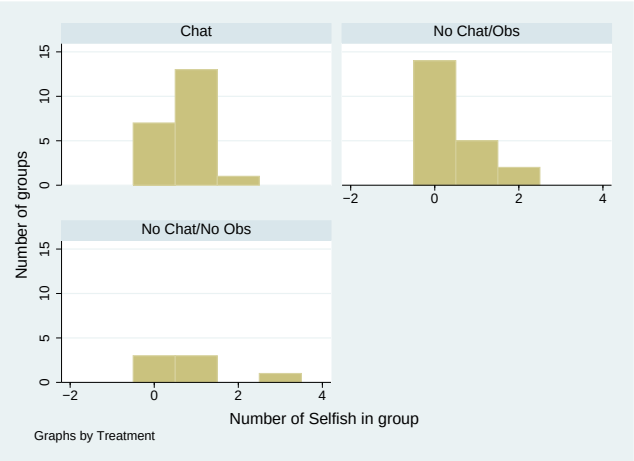

Figure 2.6: Allocation of Selfish across groups.

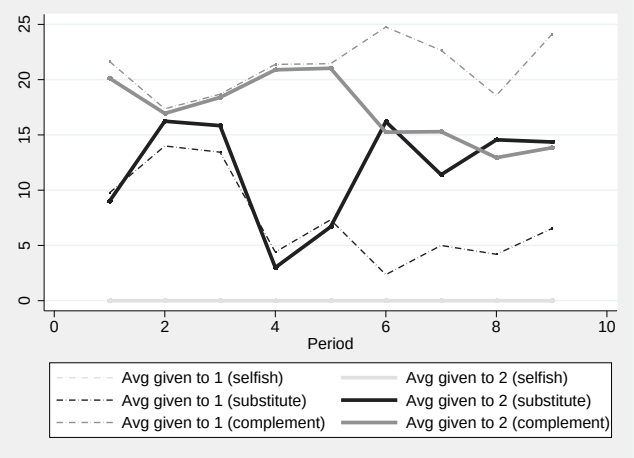

Figure 2.7: Giving rates by social preference types.

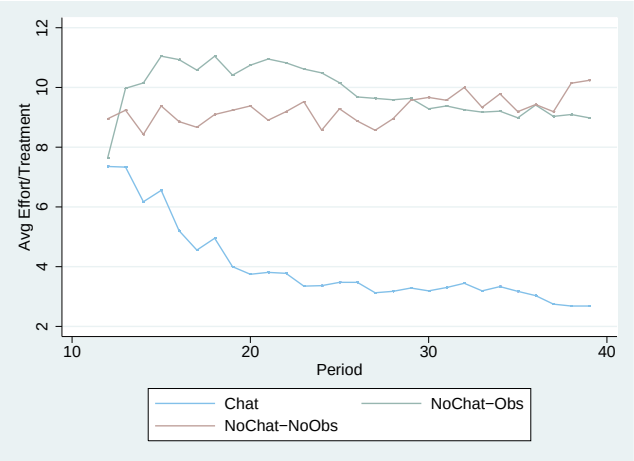

Figure 2.8: Average effort by treatment over time.

2.7 Social Preferences and Effort

Figure 2.8 provides a summary of effort choices over time by treatment. As expected, there is a strong tendency to coordinate on lower efforts over time when subjects are able to communicate and to observe past behavior in the Chat/Observability treatment. When communication is not available, observability by itself does little in sustaining lower levels of effort overall.

How do individual social preferences and group composition relate to efforts? To give an answer to this question we exploit the random allocation of subjects into groups. We compare behavior of groups with different numbers of Selfish and Other-Regarding individuals. A group solely composed of Selfish individuals should represent what we know from neoclassical economics: The case of collusion driven by pecuniary payoffs and "the shadow of the future" [26]. Comparing this to the behavior of groups whose members are Other-Regarding should therefore show us the effect of social preferences on effort in relative performance environments. Figure 2.9 gives a first overview of our findings. Consider first panel a) in the upper left corner. Here we directly address Hypothesis 1 and compare the average effort of subjects categorized as Selfish with the average effort of subjects categorized as Other-Regarding. We see that for all three treatments, average effort is higher for subjects categorized as Selfish, which is in line with our hypothesis. Panel b) in the upper right corner addresses Hypothesis 2. Here we consider average group effort as a function of the number of Selfish players within a group. In the two treatments where communication was not possible, we observe that each additional Selfish group member increases average group effort. In contrast, when communication is possible, we observe a non-monotonicity. Average group effort is lowest when there is only one Selfish individual in the group. Communication enables a subject to suggest a course of action for the group. If the other group members follow his or her advice, a subject

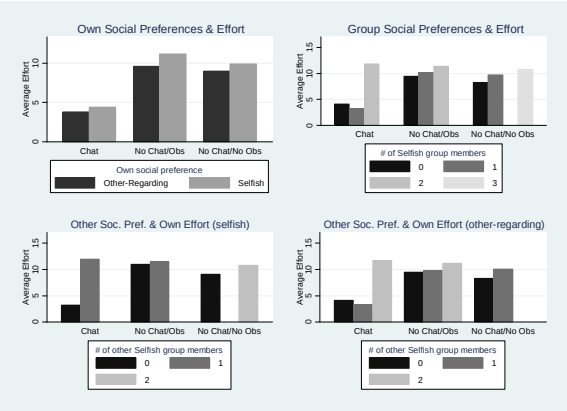

Figure 2.9: Overview of Effects of Social Preferences on Effort.

can be considered a leader. The non-monotonicity in treatment 1 thus points towards the possibility that social preferences affect the propensity to take the lead (through communication) differently than the propensity to reduce efforts. We will analyze this link in detail in the next section by analyzing the chat messages and categorizing subjects using the chat in the described way as leaders. Panel c) (resp. d)) shows how individual effort of a Selfish (resp. Other-Regarding) subject varies with the social preferences of the other group members. Thus this comparison looks at interaction effects between group members' social preferences. For a subject categorized as Selfish we find that effort is increasing in the number of other Selfish group members for all treatments. In contrast, for an Other-Regarding subject we find monotonicity of efforts only in the non-communication treatments. In the chat treatment we find as before that effort is lowest when there is one group member who is categorized as Selfish. These findings might suggest that it is not only a subject's own social preferences that determines the effort decision, but also interaction effects through the social preferences of the group members.

To explore this further in our data we first construct a dependent variable of group effort averaged over all rounds of play (at stage 2). Since groups were randomly assigned these averages are independent of treatment assignment. As a result, variation in the group members' social preferences allows us to interpret the coefficient on Selfish in an OLS regression as the causal effect of group composition on effort. Table 2.4 reports the results of regressing average group effort on the number of Selfish individuals in a group. Column 1 shows the results for treatment 1, our Chat/Observability treatment, column 2 and column 3 report the results for treatment 2 (No Chat /Observability) and 3 (No Chat/No Observability) respectively. In treatment 1 we do not find a significant effect of Selfish group members. This is to be expected given the impressions gained from Figure 2.9, where we identified a non-monotonic relationship. We will focus on explaining this non-monotonicity in the next section. On the contrary, when communication is not possible (treatment 2) and when neither communication nor observability are allowed

	(1) Treatment 1 AvgEffort	(2) Treatment 2 AvgEffort	(3) Treatment 3 AvgEffort
# Selfish	1.06	0.87**	0.860***
	(1.26)	(0.379)	(0.163)
_cons	3.18***	9.453***	8.540***
	(1.02)	(0.440)	(0.266)
N	21	21	7
adjusted R^2	-0.012	0.081	0.656

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.4: Effect of the number of Selfish group members on average group effort per session by treatment.

(treatment 3), each Selfish group member increases effort by approximately .9 units on average, which equals a 12% increase over our baseline mean effort of roughly 7.5 per period. Interestingly, when there is no observation of other group members' efforts (i.e., treatment 3), Other-Regarding players still put in significantly less effort. This is in contrast to the findings of [6] where relative incentives performed poorly in contrast to a piece rate only when monitoring of co-workers was possible.

Table 2.5 reports the effects by social preference type. We consider Other-Regarding types categorized as Complements as well as Substitutes. Again, in treatment 1 (Chat/ Observability) we find no significant effect of social preferences on the average effort level. In treatments 2 (No Chat/Observability) and 3 (No Chat/ No Observability) each Complement group member decreases effort by approximately .9 units. The effect for Substitutes is of similar magnitude, but only marginally significant in treatment 3. Recall though that there are roughly three times the number of Complement to Substitute types and thus this estimate is derived using fewer observations.

To disentangle the effect of one's own social preference from group interaction effects we estimate a random effects model for treatment 2, clustering standard errors on the group level.[8] We exclude treatment 1 as we will devote the next section to this treatment. Table 2.6 reports our results. We find evidence for hypothesis 1: a subject of Complement or Substitute type puts in significantly less effort, even controlling for group composition. A subject classified as Complement puts in 1.4 units less while a subject classified as Substitute reduces effort by 1.7 units relative to a Selfish subject. Complement or Substitute

[8]Throughout the paper when using a random effects regression we cluster at the group level. Results are qualitatively unchanged when clustering at the individual level.

	(1) Treatment 1 AvgEffort	(2) Treatment 2 AvgEffort	(3) Treatment 3 AvgEffort
# Complements	-0.593	-0.873**	-0.919***
	(1.582)	(0.389)	(0.154)
# Substitutes	-1.742	-0.856	-0.604*
	(2.009)	(0.685)	(0.265)
_cons	5.952	12.06***	11.02***
	(4.017)	(0.942)	(0.388)
N	21	21	7
adjusted R^2	-0.036	0.030	0.637

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.5: Effect of the number of Complement and Substitute group members on average group effort per session by treatment.

group members do not seem to significantly affect a subjects effort decision. Although not included in this table, we have checked whether one's reaction to the social preferences of one's group members depends on one's own social preference type. We did not find any significant relationship. We do not include lagged effort choices due to the issue of inconsistent estimates. Nonetheless, when doing so, our results are qualitatively the same. In addition, since effort choices are constrained to be between 1 and 12,we re-run our analysis using a Tobit panel model. We find these results are qualitatively the same. We also conducted our individual level analyses controlling for gender, education major, and risk preferences, and find the results qualitatively unchanged. Finally, rather than using social preference types as regressors, we conduct individual-level regressions using instead the average amount of endowment kept by a subject to examine if subjects' intensity of social preferences matters. We find little explanatory power using this measure of giving intensity. Instead, simply differentiating those that are perfectly Selfish (keep 100%) and those that are not has significant explanatory power. We also consider an alternative classification of social types. In particular, we now classify Selfish subjects as those that keep on average at least 90% of their endowment (as opposed to 100%). Using this less stringent definition of Selfish subjects we find that the magnitude of the coefficient estimates on Selfish types decreases, but are still significant. However, for the group level regressions although the sign is still correct, the coefficient estimates are no longer significant.

	(1) Effort Coeff.	S.E.	(2) Effort Coeff.	S.E.
Period	-0.0538*	(0.0294)	-0.0538*	(0.0294)
Selfish	1.478***	(0.401)		
# Other-Selfish	0.569	(0.412)		
Complement			-1.410***	(0.386)
Substitute			-1.714**	(0.854)
# Other Substitutes			-0.427	(0.669)
# Other Complements			-0.604	(0.411)
Constant	10.85***	(0.502)	13.46***	(1.188)
N	1827		1827	
R^2within/between	0.0322/0.0954		0.0322/0.0994	

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.6: Effect of own and others social preferences on own effort (treatment 2).

Overall we can conclude that the results from Treatment 2 lend support for hypotheses 1 and 2: Our group regressions show that groups consisting of more Other-Regarding members exhibit lower efforts relative to more Selfish groups when communication is not possible. In particular, group members who are of the Complement type drive group efforts down. Looking at individual effort we find that a subject's own social preference determines the amount of effort invested. In particular, Complements as well as Substitutes exhibit lower effort than their Selfish counterparts. We cannot reject the null hypothesis that Complements and Substitutes depress effort by the same magnitude (p-value 0.7102).

A subject's group members do not significantly affect his effort decision. Surprisingly, when communication is possible social preferences do not seem to affect efforts in a linear fashion. We found that groups with only one Selfish group member seem to be best at coordinating on non-competitive efforts. In the next section we investigate deeper into the reason behind this non-monotonicity. In order to do this we differentiate between 1) the initiation of low efforts through chat and 2) the general tendency of choosing low efforts. Social preferences might relate differently to these two aspects of non-competitive efforts. To disentangle these effects we analyze the chat messages of each group and identify leaders, or "collusion initiator" and their social preferences.

2.8 Leadership

In treatment 1, a subject can take the initiative through chat, asking the group members to jointly exert low effort. This channel was absent in all other treatments. We use the chat messages to identify this form of "leadership". We differentiate between two kinds of leaders: "First Leader" and "Right Leader".[9] A First Leader is the first subject to propose coordination on low efforts, without consideration for the actual level proposed. Thus this is a relative broad category. A Right Leader, on the other hand, is the first to propose coordinating on the Pareto-Dominant minimum effort (i.e., all providing effort of 1). We identify 18 First Leaders (29%) and 13 Right Leaders (21%) among the 63 subjects (21 groups) in the Chat treatment (11 subjects are both a Right Leader and a First Leader). We start by providing a breakdown of the social preferences of the subjects we identified as leaders. Figure 2.10 shows the distribution of social preference types in the population of First Leaders and in the population of Non-First Leaders. Figure 2.11 shows the distribution for Right Leaders and Non-Right Leaders.

For First Leaders as well as Right Leaders we observe that Selfish and Substitute subjects seem more likely to take the initiative, while Complements are less likely to lead. A Fisher's exact test shows though that the distribution of social preferences is not statistically different between First Leaders and Non-First Leaders (p-value 0.151) while it is

[9] We initially collected a third category: "Failed Leader" for a subject who called on his group members to decrease efforts but was not listened to/followed. This is a rare event in our study and thus we do not include this variable in our analysis.

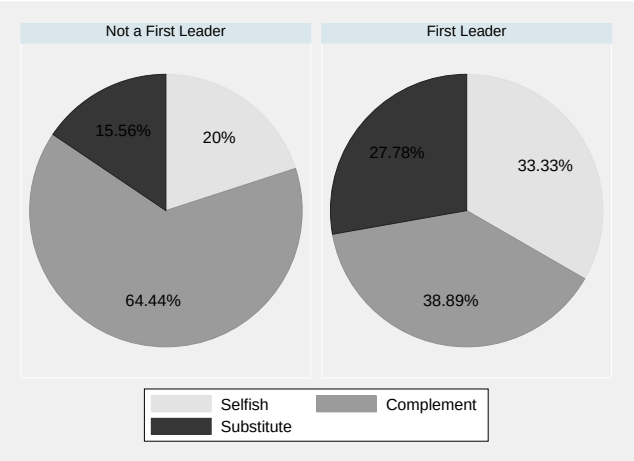

Figure 2.10: Distribution of social preferences: First Leader.

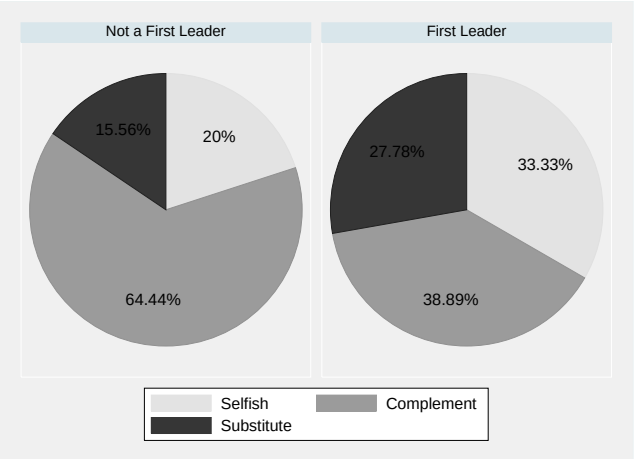

Figure 2.11: Distribution of social preferences: Right Leader.

statistically different for Right Leaders and Non-Right Leaders (p-value 0.066). That is, while we do not find significant differences between First Leaders and Non-First Leaders in terms of social preferences, we do find differences between those leaders suggesting the Pareto-Dominant outcome and those who don't. In particular, when comparing the two types of Selfish and Complements, we find Right Leaders tend to be Selfish types over Complement types (two-sided Fisher's exact test, p-value 0.047) . However, when comparing the fraction of Right Leaders that are Selfish to Substitute types, we cannot distinguish a difference. In short, subjects with Complement type social preferences are unlikely to suggest the right group strategy when leading. Thus, it turns out that hypothesis 3 is not supported in the data.

Given that we filter out the effect of social preferences that runs through leadership in suggesting low efforts, is it still true that low efforts are related to group members' social preferences similarly as in treatment 2 and 3? Table 2.7 reports the results of a random effects model for treatment 1. Column 1 replicates our findings without regard for leader emergence. In column 2 we add as a control whether a Right Leader has emerged (dummy that takes on a value of one once a Right Leader emerged in the given group) and whether the subject itself is a Right Leader (time-invariant dummy that takes on value of one for all subjects who are classified as Right Leader). We only consider the variable Right Leader because First Leader is found not to be related to social preferences.[10] Notice that the coefficients of own social preference as well as group members' social preferences are highly significant and larger in magnitude once controlling for leadership in this way. This means that controlling for leadership, social preferences lead to significantly lower group efforts. The effect is slightly larger in magnitude than in treatment 2. We find that a Selfish subject puts in 2 units effort more per period than an Other-Regarding subject. Furthermore the presence of a Selfish group member increases a subject's own effort by 2 units per period. We cannot reject the null hypothesis (t-test, p-value 0.8940) that those two coefficients are equal. Thus in treatment 1 interaction effects seem to be more important while in treatment 2 only one's own social preference type was a significant predictor of effort.

Column 3 investigates further into the timing of the effect of social preferences. We include interactions of social preference measures and the emergence of a leader. We find that social preferences depress efforts before a Right Leader emerges in a group. Once a leader emerges there is no difference between Selfish and Other-Regarding choices. Selfish are thus no more likely to deviate and we conclude that communication is a powerful collusion device in our setting. Finally, note that the coefficient of Right Leader is insignificant. Thus, Right Leaders do not lead also by good example, i.e. putting in lower effort but only purely through cheap talk. For brevity we report here only a dichotomous measure of social preferences. Replicating the analysis with Selfish, Complement and Substitute typography yields similar results and the coefficients on both Complement

[10]Our results are robust though to including controls for First Leader emergence and type as well though coefficients on social preferences become slightly smaller in magnitude.

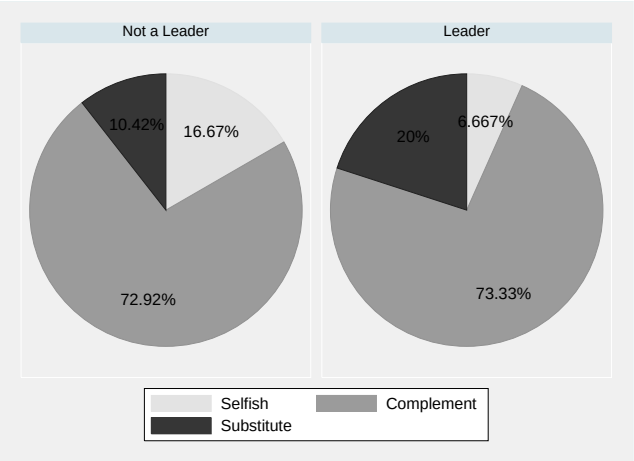

Figure 2.12: Distribution of social preferences: leader by example.

and Substitute as well as # other Complements and # other Substitutes are of similar magnitude and all highly significant.

One might conjecture at this point that Selfish individuals in our sample are smarter than Other-Regarding ones, or just are better able to understand the game and optimal strategy. Thus, naturally they will be the ones suggesting non-competitive efforts, not because of their social preference, but because of their better understanding of the game. This argument is flawed though. It does not explain why we find that, controlling for leadership, Other-Regarding subjects systematically put in lower efforts on average than Selfish ones. In addition, in treatment 2, when communication is not possible, we would expect that a subject understanding the game better would try to lead by example in order to induce the other group members to follow his or her lead. Thus we categorize subjects as attempting to be leaders when expending an effort less than four given that in the round before his or her group members expended efforts larger than 9. We do not systematically observe Selfish subjects leading "by example" with reduced efforts to communicate the optimal strategy to their group members (if anything, we observe Substitute and Complement types trying out low efforts, however, we do not find a statistical difference in the two distributions (Fisher's exact test, p-value = 0.439).). Figure 2.12 shows the distribution of social preferences between attempted leaders by example and non-leaders in treatment 2. Finally, we do not find that subjects with a background in Economics or Business are more likely to be Right Leaders. Thus we conclude that differences in leadership are unlikely to be caused by differences in the ability to understand the optimal strategy.

We conclude that Other-Regarding preferences seem important in sustaining lower levels of effort throughout the game also in treatment 1. When a leader can lead by communication he/she tends to be a Selfish individual and low effort followers tend to be Other Regarding. This suggests the ideal group for creating and sustaining collusion

	(1)	(2)	(3)
	Effort	Effort	Effort.
Period	-0.133***	-0.0725***	-0.0728***
	(0.0276)	(0.0250)	(0.0245)
Selfish	1.069	2.054***	2.797***
	(1.596)	(0.7370)	(0.6870)
# Other Selfish	1.0600	2.067***	2.864***
	(1.5810)	(0.6940)	(0.6000)
Right Leader Exists		-5.709***	-3.661***
		(0.6370)	(0.4230)
Right Leader		0.0784	0.107
		(0.3500)	(0.3380)
RLeader*Selfish			-2.729***
			(0.6780)
RLeader*OthSelfish			-2.800***
			(0.5620)
Constant	6.628***	7.353***	6.911***
	(1.4710)	(0.7410)	(0.7890)
N	1827	1827	1827
R^2within/between	0.1000/0.0379	0.2117/0.7465	0.2184/0.7848

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.7: Effect of social preferences on individual effort controlling for leadership (Treatment 1).

is a Selfish leader and Other-Regarding followers. However, absent communication, a fully Other-Regarding group is best at coordinating on low effort. Thus, the relationship between social preferences and non-competitive efforts seems to be a nuanced one. The most convincing case that social preferences matter in depressing efforts would be to randomly turn such preferences on and off across subjects and compare the outcomes. Of course, this is not possible in practice. However, our final treatment attempts to approximate just such a procedure.

2.9 Robot Treatment

This treatment is similar to treatment 2 (No Chat/Observability) in the sense that subjects cannot communicate but get to observe the efforts and payoffs of their group members after each period. The crucial difference is that in stage 2, instead of randomly pairing subjects to other subjects we paired them to two simulated subjects we call robots. In particular, we programmed 42 "robot" subjects who react to past effort decisions just like real subjects did in earlier treatments. Each "robot" chooses current period effort based on last period's effort choices of the other two subjects in the same way the real subject did on which it is based on. Critical in this treatment is that it is no longer the case a subject's effort choices impose a negative externality on other players, as the robots receive no payoffs. Thus the fundamental difference between treatment 2 and treatment 4 (i.e., the robot treatment) is the latter attempts to "turn off" subjects' social preferences since their actions no longer affect any other player. Note, however, that social preferences are not completely absent, as the robot's choices simulate decisions by participants whose social preferences did matter. Thus, subject's decisions can reflect beliefs about the past subjects' social preferences. This in fact is helpful for us, as it allows us to distinguish an alternative hypothesis: "Selfish" subjects differ in their beliefs about their group members' (re-)actions from "Other-regarding" subjects. If this were the case, we should still see a difference between Selfish and Other-regarding effort choices in this treatment. While if our categorization instead captures social preferences, differences in effort should vanish in this treatment.

We first compare subject behavior for treatment 2 and 4 graphically. Figure 2.13 depicts the effort profiles over the 29 periods of play by treatment for Selfish and Other-Regarding individuals.We find that in the first half of the relative performance stage (Periods 12-27) the effort of Selfish and Other-Regarding subjects in Treatment 4, the robot treatment, is not statistically different (t-test, p-value 0.2122). There is some effort divergence in the intermediate term though—however, by the end of the relative performance stage, efforts of different social types converge back to similar effort levels. In fact, in the last 5 rounds a t-test cannot reject equality of efforts (p-value 0.1578). Interestingly, efforts of all social preference types in Treatment 4 converge towards the efforts of Selfish subjects in Treatment 2. For the last 5 periods a t-test cannot reject equality of efforts of any social preference type in treatment 4 compared to Selfish in treatment 2 (i.e.,

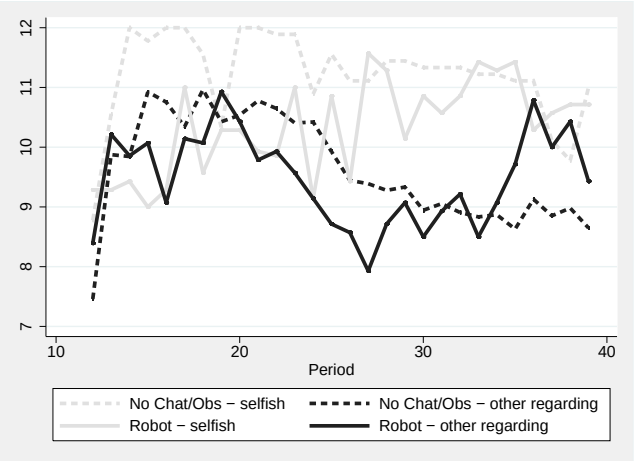

Figure 2.13: Comparing efforts between selfish and other regarding types over time.

Selfish treatment 2 vs. Selfish treatment 4, p-value .7315; Selfish treatment 2 vs. Other-Regarding treatment 4, p-value .1578; Other-Regarding treatment 4 vs Other-Regarding treatment 2, p-value .0016). However, as suggested by the chart, if we include the final ten periods of effort, there is a statistical difference in effort between Other-Regarding and Selfish players (p-value .002). Whatever the case, it is unclear whether subjects are behaving similarly towards the end of the game as they did during the first half of the game. Thus, we do not find convincing evidence of equal behavior between Selfish and Other-Regading players for the last half of the relative performance game. Perhaps, subjects forget that they are playing "robot" subjects and began behaving as if they are playing "real" subjects. We did attempt to minimize this possibility by reminding subjects on each effort-entry screen that their effort choice will not affect the payoffs of any participants. Unfortunately, we cannot rule out that subjects disregarded this message after a while. It does seem these results suggest beliefs are not driving the difference in choices for different types of players: beliefs should loom largest in creating differences at the beginning of the relative-performance game before they converge based on experience. However, we observe just the opposite pattern.

If instead analyzing individual rather then average effort choices, we find a similar pattern of similar effort choices across social preference types. Table 2.8 reports the results of regressing individual effort on own and group members' social preference types for treatments 2 and 4. The coefficient estimate for Selfish is half the value as in treatment 2 and is no longer significant, though we do note the sample size is smaller.

Overall, the robot treatment provides further evidence that social preferences (and not beliefs) matter in creating and sustaining non-competitive efforts.

	(1) Treatment 2 Effort	(2) Treatment 4 Effort
Period	-0.053*	0.0168
	(0.0294)	(0.0285)
Selfish	1.478***	0.824
	(0.4010)	(0.8130)
# Other Selfish	0.569	-0.28
	(0.4120)	(0.9960)
Constant	10.85***	9.152***
	(0.5020)	(0.6850)
N	1827	609
R^2within/between	0.032/0.095	0.003/0.049

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.8: Effect of social preferences on individual effort treatment 2 vs. treatment 4.

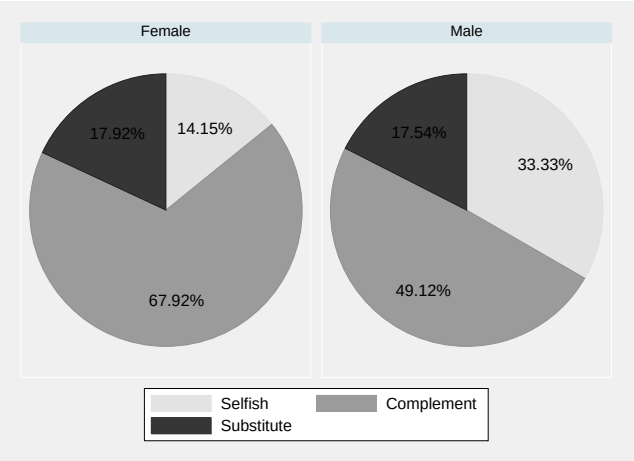

Figure 2.14: Distribution of social preferences by gender.

2.10 Gender

In this section we explore the interaction of gender and social preferences. Many past studies, from empirical to experimental, find it important to control for gender. For example women shy away from competition or women free-ride less in public goods games (for a survey see [24]). In our study, gender matters both in terms of the likelihood of becoming a leader and also in determining social types. In this section, we relate our social preference measure to gender and compare the predictive power of both together. Figure 2.14 shows the distribution of social preference types by gender. We find that, in accordance with the literature, females are more likely to be categorized as a Complement compared to males. Males, meanwhile, are more likely to be a Selfish type. These differences are statistically significant (Fisher's exact, p-value = 0.013). However, there is not a statistically significant difference between the likelihood of males and females being Substitute types.

Tables 2.9 and 2.10 report individual level regressions separately for treatment 1 (Chat/Observability) and 2 (No Chat/Observability), respectively, now with an additional control for gender. The dummy Male takes on the value one if the subject is a male and 0 otherwise. There was one subject who chose not to identify their gender. We omitted this subject from the regressions.

Overall, gender effects are small relative to individual and group members' social preference types. The individual gender coefficient is only statistically significant in Treatment 2, where male subjects exert an average of 0.89 less effort per period. This is significant at the modest 10% level The coefficient on the number of other males in a group is highly significant in treatment 1 but inconsequential in magnitude. In terms of propensity to lead, gender matters. Whether leading-by-example (i.e., treatment 2) or leading-through-speech (i.e., treatment 1), males are more than twice as likely to be leaders. For treatment

	(1)	(2)	(3)
	Effort	Effort	Effort.
Period	-0.0725***	-0.0674***	-0.0664***
	(0.03)	(0.02)	(0.02)
Selfish	2.054***		2.098***
	(0.74)		(0.61)
# Other Selfish	2.067***		2.145***
	(0.69)		(0.58)
Right Leader Exists	-5.709***	-5.523***	-5.620***
	(0.64)	(0.67)	(0.61)
Right Leader	0.078	0.301	0.237
	(0.35)	(0.39)	(0.29)
Male		-0.155	-0.066
		(0.14)	(0.17)
# Other Males		-0.011	-0.0130**
		(0.01)	(0.01)
Constant	7.353***	9.160***	7.722***
	(0.74)	(0.75)	(0.63)
N	1827	1798	1798
R^2within/between	0.212/0.746	0.213/0.648	0.213/0.788

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.9: Social Preferences vs. Gender, Treatment 1.

	(1)	(2)	(3)
	Effort	Effort	Effort.
Period	-0.0538*	-0.0544*	-0.0544*
	(0.029)	(0.030)	(0.030)
Selfish	1.478***		1.402***
	(0.401)		(0.423)
# Other Selfish	0.569		0.17
	(0.412)		(0.429)
Male		-0.699	-0.891*
		(0.485)	(0.498)
# Other Males		0.0196**	0.015
		(0.009)	(0.010)
Constant	10.85***	10.36***	10.44***
	(0.502)	(0.689)	(0.696)
N	1827	1798	1798
R^2within/between	0.0322/0.0954	0.0324/0.105	0.0324/0.155

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table 2.10: Social Preferences vs. Gender, Treatment 2.

1, males are more likely to be a First Leader and Right Leader (t-test p values of .0573 and .0626, respectively). For treatment 2, males are more likely to attempt to lead by example (t-test p value of .0059). In short, to the extent gender matters, it seems to operate only indirectly through females being more likely to be Other-Regarding and males more likely to lead. Nonetheless, this implies past studies that simply control for gender may be nosily proxying social preferences and leadership.

2.11 Conclusion

Agents coordinating on low efforts can be driven by both collusion, as traditionally defined, and social preferences. In practice, these effects are difficult to disentangle, as

both of these mechanisms may lead to the same observed behavior. We turn to economic experiments to identify the dominant mechanism and mediating factors in generating noncompetitive efforts.

We find other regarding preferences substantially explain noncompetitive efforts but in surprising ways. First, players categorized as selfish are more likely to initiate collusive behavior when communication is available. Second, controlling for the existence of leaders, players categorized as other regarding exert lower levels of effort. Thus, when communication is available, a group with one selfish and otherwise other regarding members can most successfully create and sustain noncompetitive efforts. When communication is not available, groups of other regarding players produce the lowest levels of effort. In terms of gender, females are more likely to be other regarding. However, males are more likely to lead, both by speech and action.

It seems other regarding players unconditionally provide lower levels of effort. In all treatments, other regarding players tend to provide a similar 1-2 unit reduction in effort, about 15% of average effort. Even when other group members' efforts are not observable the effect is similar. In treatment 1 though, when communication and observability are possible, we do find evidence of an interaction effect. A subject reduces his or her effort significantly when paired up with other-regarding as compared to selfish group members.

We also attempted to "switch off" subjects' social preferences through our robot treatment, which simulated the responses of human subjects via machine, thus removing a player's negative externality of higher efforts on other players' payoffs. By the end of the treatment, subjects of any social preference type acted just like the selfish subjects in the comparable treatment with groups of all human subjects. This provides further evidence that other regarding people are depressing their efforts as they internalize the negative externality of high effort.

Our findings have policy implications for relative performance settings. In organizations with more other regarding workers (e.g., nonprofits or firms engaged in corporate social responsibility), relative performance schemes are likely to not be as effective as in other organizations. When compensation naturally induced a negative externality on one's co-workers, for example through technological constraints, screening of workers can increase performance. Generally, when workers are closely engaged so that communication flows freely and output is easily observed, relative performance schemes are more likely to encourage noncompetitive behavior. However, if a firm consists of mostly selfish workers, relative incentive schemes should elicit very high levels of effort, as we found such a setting yielded above Nash Equilibrium predicted efforts.

Although we only tested the possibility of valuing negative externalities, to the extent workers also value their positive externalities, other regarding preferences are likely to help the free rider problem amongst teams. That is, a team of workers with other regarding preferences that receive a share of the common output are more likely to provide higher outputs, as they further value their efforts positive effects on their team members. Thus suggests with homogeneous groups of workers, whereas selfish groups should be paid in relative performance incentives, other regarding workers should be placed in team

production pay. This would be an interesting avenue for further research.

We note we did not consider the case where workers might value their firm's payoff. Thus, our results can be seen as applying to settings where ownership is dispersed or the worker is removed from the top of the hierarchy.

Finally, our measure of leadership is endogenous to the effort exerted in each group. It is an interesting challenge to design an experiment in which leadership varies with incentives and analyze how it relates to social preferences. We leave these topics for future research.

Chapter 3

Paths to Order and Prosperity: State Formation with Endogenous Coercive and Productive Capacities

3.1 Introduction

The relation between political order and economic prosperity has been a perennial concern in the social sciences. According to an intellectual tradition that can be traced back to Thomas Hobbes, order is a necessary, albeit not sufficient, condition for prosperity. A central claim in the Hobbesian tradition is that the state is the only source of political order. In a stateless society, war and preparation for war become a permanent condition that consumes all human effort and material resources. Nothing is left to make the economy grow. On the other hand, reversing the Hobbesian equation, recent research in economics and political science has argued that the consolidation of political order is the outcome, rather than the pre-condition, of prosperity. A key motivating stylized fact is that poor countries are more prone to civil wars (see inter alia [20]).[1]

This paper advances a formal model for understanding the relation between order and prosperity. We build the argument in three steps. First, we take an agnostic approach to the causal relation between order and prosperity. Instead of positing effects of one on the other, we investigate under what conditions, of military and economic technology, order and prosperity can be jointly achieved, and when neither or only one of them is possible. We do this by postulating the existence of an "incumbent" facing the potential attacks of a challenger. Examples of this situation are historical cases of civilized urban centers (a port city home to traders, or the central administration of an agricultural settlement) facing the predatory attacks of nomadic tribes or plundering warlords. In that setting, the incumbent has the potential to invest in expanding its productive capacities to grow future income flows, but may prefer to spend its resources in arming itself and consuming now if future flows maybe lost to a successful attack. Two aspects are key to the incumbent's calculus: one is that if attacks will take place, the effective rate of return to productive investments may not be high enough to justify investment. The other one is that by expanding the future income flow, productive investment may render predatory challenges even harsher. When investment is curtailed, the polity will be trapped in a conflictive and economically stagnant situation.

But orderless stagnation is not the only possible outcome of the model. A key finding is that all four combinations are theoretically possible under reasonable assumptions for military and productive technology, including both the existence of order without prosperity and of prosperity without order. Although anti-Hobbesian, the possibility of prosperity without order is consistent with a widespread occurrence in the history of humanity: populations that prefer to grow their economies rather than building substantial military protection despite permanent threats of predation from neighboring plunderers, like the Chinese with Mongolians and the Saxons with Vikings in the 10th century, as well as the Americans with the Sioux and the Argentines with the Quilmes in the 19th century. Also, we find that economies may be locked in situations of conflict without growth because growing the economy would trigger even harsher predatory challenges

[1] Others argue that some types of wealth, especially the one derived from mineral resources, are detrimental to political order. See for example [67], and the survey by [12].

from neighbors.

Second, we trace the effects of shocks to military and productive capacities in terms of transitions from one area to another in the order/prosperity space. A rich picture of paths where effects of shocks depend on initial conditions emerges. For instance, a small shock to productive capacities will have a different effect depending on whether it occurs in a context of political disorder and economic stagnation or a context of order without prosperity. A key finding is that whereas big shocks to military capacity always have positive effects in terms of placing a country closer to order and prosperity, big shocks to productive capacity can have negative effects, in particular, worse effects than small socks to productive capacity. When it occurs in a context of order without prosperity, a large enough shock to productive capacity triggers the voracity of challengers, and results in growth without order. A smaller shock is preferable.

If exogenous shocks can shift a polity from one combination of order and prosperity to another, foreign policy interventions in failed states that recreate those shocks can do the same. Thus, we can apply our model to draw some lessons for state-building in modern societies featuring failed states. The defense community of the United States has reached a consensus around the idea that development initiatives can be an important element in counterinsurgency and state-building (see for instance the Counterinsurgency Manual from the Headquarters of the Army, 2008, [63]). An important question concerns the relative weight of development vs military build up initiatives in countries like Iraq or Afghanistan. Our model shows that enhanced military capacities are a necessary condition for achieving order and prosperity, while expanding productive capacities is not. Moreover, under certain conditions, an imbalanced mix may worsen outcomes.

The third part of our argument involves extending our the original model to allow for endogenous upgrades in military capacity. In this context, it is possible to trace how shocks to the initial levels of wealth shape political order. An anti-Hobbesian finding is that, starting from a situation of political disorder, prosperity is the driver of state consolidation. The new wealth is allocated to expanding military capacity in order to enable political order and protect the economy. However, once political order is in place, a Hobbesian effect is observed: the new levels of defense prevent predatory challenges, and the enhanced return to productive investment ushers in faster growth.

3.2 Literature

Our paper contributes to the understanding of how a polity may move from a stateless, subsistence economy to one that enjoys order, understood as the presence of a monopoly on violence, and prosperity, understood as having the ability to grow its income, and thus escape subsistence. The productive asset controlled by our incumbent player could be land or a trading infrastructure, but also a state apparatus that helps expand (and tax) the economy's revenue. In this sense, the productive investments in our model can be interpreted as investments in state capacity as studied by [10]. Their Chapter 4 studies

the incentives of a controlling party to expand state capacity when there is conflict. An important difference in our model is that it adds the possibility to study study a key tradeoff governing the dynamics of state capacity: the probability and intensity of conflict in our model is endogenous to the investments made by those controlling the state. In other words, our model helps understand the dynamics of state capacity when expanding those capacities make the state a more attractive booty for challengers.

Our paper is also related to the formal study of state consolidation. [65] analyzes state consolidation when it happens exogenously and endogenously. The key difference is that in our model consolidation is studied in relation to the evolution of the economy.

The rich array of consequences that economic shocks may have for peace and growth outcomes underscores the complexity of the interrelation between order and prosperity. This message of our paper joins an emerging emphasis on the conditionality of effects in conflict research. For example, [25]identified theoretically the diverging effects on conflict of economic shocks depending on the relative factor intensity of the productive and conflict activities. These effects have received empirical validation in the Colombian context (see [30]). Using a cross-country approach, [10] showed that price shocks may have different effects on conflict depending on whether they affect exports vs imports.

3.3 The Basic Model

Our baseline model features the incentives to build an army to protect wealth from usurpers at the cost of detracting from the resources that are available for consumption or investment. Later on we introduce the decision to invest in military capacity.

Players

There is an "incumbent" who controls a productive asset that yields a non-storable flow $v_t > 0$ every period. The incumbent may be seen as the merchant elite of a port city who control the port infrastructure and the gains from trade. Alternatively, the incumbent can be seen as the ruling caste of a city controlling an agricultural hinterland. Thus, the productive asset in these examples is alternatively the port and an area of productive land. There is also a "challenger" who receives an exogenous income flow from nature that we normalized to zero, and who is interested in wrestling control of the productive asset away from the incumbent. This challenger may be seen as nomadic tribes that threaten with invading the city (as with the Mongolians in China) or relatively idle men led by provincial warlords (as with "caudillos" in 19th century Argentina who threatened the wealthy port city of Buenos Aires).

Actions, resources and technology

In each period the incumbent can spend its flow v_t in consumption, productive investment i_t or mobilizing resources to defend its asset. One dollar of productive investment i_t costs one dollar of consumption and it adds $\beta > 1$ dollars to the yield of the productive asset in the future. That is, the asset yield evolves according to the relation $v_{t+1} = v_t + \beta i_t$; we abstract from depreciation for simplicity.

The effectiveness of the incumbent's army is denoted a_t and such an army costs the challenger an amount $\frac{a_t}{c}$ where $c \geq 0$ is the value of the incumbent's military capacity. The higher the military capacity of the incumbent, the higher the "firepower" a_t attained by a given war effort $\frac{a_t}{c}$ (or alternatively, given a war effectiveness a_t, the lower is the war effort $\frac{a_t}{c}$ when the military capacity c is higher). In this section c is exogenous and we will derive implications for conflict and growth stemming from different values of c. The expanded version of the model in section 3.4.2 will be devoted to endogeneizing c. Thus, in period t the incumbent must observe a budget constraint

$$v_t - i_t - \frac{a_t}{c} \geq 0. \tag{3.1}$$

The challenger observes the choices of a_t and i_t by the incumbent and chooses its own war effort b_t.[2] If victorious in the first period the challenger captures control of the productive asset in the second period. Whenever the challenger attacks ($b_t > 0$), it prevails with probability $\frac{b_t}{a_t+b_t}$ and it gains nothing with the complementary probability; in other words, we adopt the typical Tullock contest success function. If the incumbent is defeated it obtains an outside payoff normalized to zero; the challenger becomes the new incumbent and in the following period faces a new challenger. If the challenger selects $b_t = 0$ we say the incumbent has successfully deterred the challenger, and this lack of challenge to the authority of the incumbent is the outcome we associate with state consolidation.

Timing

In each period the incumbent selects a_t and i_t. After observing (a_t, i_t) the challenger selects b_t. If $b_1 = 0$, the players retain their positions in period 2. If $b_1 > 0$, then there is a war at the end of period 1. The winner of the war becomes the incumbent in period 2, and faces a new challenger then.

Payoffs - problems for the challenger and incumbent

Both challenger and incumbent are risk neutral and care linearly about income and units of effort. The incumbent acts as a Stackelberg leader, choosing a_1 and i_1 to maximize the value of the game for an incumbent V_t:

$$V_t = v_t - \frac{a_t}{c} - i_t + \frac{a_t}{a_t + b_t} V_{t+1}. \tag{3.2}$$

The challenger chooses b_t to maximize the expression

$$W_t = \frac{b_t}{a_t + b_t} V_{t+1} - b_t, \tag{3.3}$$

where $V_{t+1} = v_t + \beta i_t$.

We will solve for a Subgame Perfect Nash Equilibrium by backward induction.

[2] Assume that the challenger's war expense is basically effort is equivalent to assuming that the challenger's income is sufficient to finance the optimal war effort b_t^*. Because the effects of interest are not driven by a budget constraint on the challenger being binding, we follow the most parsimonious approach of not making explicit a resource constraint on the challenger.

Second period

In the second period, there is nothing the challenger would want to fight for, as there is no future in which to enjoy the productive asset if stolen. Thus, $b_2 = 0$. The incumbent then chooses i_2 and a_2 to maximize the value of consumption in the second period $V_2 = v_2 - i_2 - \frac{a_2}{c}$. Since there is no use for an army and investment would only pay in a nonexistent third period, $i_2 = a_2 = 0$, yielding $V_2 = v_2$.

First period

The challenger observes the pair (a_1, i_1) and chooses b_1 to maximize W_1 as given by expression (3.3). Since the first order condition is $\frac{a_1}{(a_1+b_1)^2} v_2 = 1$, and $v_2 = v_1 + \beta i_t$, the best response function of the challenger is immediately seen to be,

$$b_1(a_1, V_2) = \begin{cases} \sqrt{a_1 (v_1 + \beta i_1)} - a_1 & if\, a_1 < V_2 \\ 0 & otherwise \end{cases}. \tag{3.4}$$

This expression exhibits a key trade-off of the model: productive investments i_1 raise the value of the productive asset. Thus, conditional on maintaining control of the asset, investment is a good idea for the incumbent since $\beta > 1$; however, the future control of the asset is not a forgone conclusion. Investment raises the incentives of the challenger to arm itself since it makes it more attractive to become the incumbent. Therefore, while productive investments increase the value of future incumbency, they may lower the chance that the current incumbent gets to reap that value. The lack of state consolidation may be an obstacle for productive investment and growth. The fundamental problem is to understand whether there are any parameter values v_1, c, and β that map into a path of state consolidation and growth. To answer this question we must study the problem of the incumbent.

The incumbent maximizes V_1 as given by (3.2) subject to the budget constraint (3.1) and anticipating the challenger's best response in (3.4). The latter indicates that if $a_1 \geq v_1 + \beta i_1$ the challenger will choose not to fight, and therefore the incumbent would never choose a_1 beyond the point $v_1 + \beta i_1$, which attains deterrence. This can be incorporated into the incumbent's problem as an additional deterrence constraint. The incumbent's problem in period one can then be written as,

$$\max_{a_1, i_1} v_1 - \frac{a_1}{c_1} - i_1 + \frac{a_1}{a_1 + b_1}(v_1 + \beta i_1) \tag{3.5}$$

subject to

$$v_1 - \frac{a_1}{c} - i_1 \geq 0 \; (BC) \tag{3.6}$$

$$v_1 - a_1 + \beta i_1 \geq 0 \; (DC) \tag{3.7}$$

$$a_1 \geq 0$$

$$i_1 \geq 0.$$

Let us call λ_{BC}, λ_{DC}, λ_a and λ_i the Lagrange multipliers for each restriction. The Lagrangian, which expresses the expected utility of the incumbent, is:

$$\begin{aligned}\mathcal{L} = v_1 - \frac{a_1}{c} - i_1 + \frac{a_1}{a_1 + b_1}(v_1 + \beta i_1) \\ + \lambda_{BC}(v_1 - \frac{a_1}{c} - i_1) + \lambda_{DC}(v_1 - a_1 + \beta i_1) + \lambda_a a_1 + \lambda_i i_1. \end{aligned} \tag{3.8}$$

We will characterize the solution $(a_1, i_1, \lambda_{BC}, \lambda_{DC}, \lambda_a, \lambda_i)$ to this problem for each parameter combination (β, c, v_1).

The first order and complementary slackness conditions that characterize the optimum are given by,

$$\frac{\partial \mathcal{L}}{\partial a_1} = \frac{1}{2}\sqrt{\frac{v_1 + \beta i_1}{a_1}} - \frac{1}{c_1} - \frac{\lambda_{BC}}{c_1} - \lambda_{DC} + \lambda_a = 0; a_1 \geq 0, \lambda_a \geq 0, \lambda_a a_1 = 0 \text{ c.s.} \tag{3.9}$$

$$\frac{\partial \mathcal{L}}{\partial i_1} = \frac{\beta}{2}\sqrt{\frac{a_1}{v_1 + \beta i_1}} - 1 - \lambda_{BC} + \lambda_{DC}\beta + \lambda_i = 0; i_1 \geq 0, \lambda_i \geq 0, \lambda_i i_1 = 0 \text{ c.s.} \tag{3.10}$$

$$\lambda_{BC}(v_1 - \frac{a_1}{c_1} - i_1) = 0 \text{ c.s.}, \ \lambda_{DC}\left(v_1 - a_1 + \beta i_1\right) = 0 \text{ c.s.} \tag{3.11}$$

Solving the program (3.8) requires checking which combinations of values for the endogenous variables $(a_1, i_1, \lambda_{BC}, \lambda_{DC}, \lambda_a, \lambda_i)$ constitute the optimum for different regions of the parameter space $(c_1, \beta, v_1) \in \mathbb{R}^3_+$. Note from (3.9) that the marginal benefit of a_1 goes to infinity as a_1 goes to zero (a typical feature of contests), so the optimum must feature $a_1 > 0$ and $\lambda_a = 0$. Beyond this, the method for solving the problem is tedious: it requires checking which combinations of values for the endogenous variables are consistent with the constraints for each parametric region and then identifying the ones that yield the highest value for the program. The details of the solution are contained in the appendix. A summary of the solution is offered in the following,

Proposition 1. *Optimal behavior by the incumbent yields a division of the parameter space $(c_1, \beta, v_1) \in \mathbb{R}^3_+$ into four distinct regions:*

- ***Region 1 (R1):*** $\left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | c_1 \geq \beta, \ \beta \geq c_1/(c_1 - 1) \text{ and } c_1 \geq 1\right\}$ *Consolidation and growth.*

In R1 the solution is: $\left\{a_1 = v_1 \frac{c_1(1+\beta)}{c_1+\beta}, i_1 = v_1 \frac{1}{2}\left(1 - \frac{1}{\beta}\right), \mathcal{L} = v_1 \frac{1}{2}\left(1 + \frac{1}{\beta}\right)\sqrt{\beta c_1}\right\}$

- ***Region 2 (R2):*** $\left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | \beta > c_1, \ \beta \geq 4/c_1 \text{ and } \beta \geq 1\right\}$ *Unstable growth. In R2 the solution is:* $\left\{a_1 = v_1 \frac{c_1}{2}\left(1 + \frac{1}{\beta}\right), i_1 = v_1 \frac{c_1 - 1}{c_1 + \beta}, \mathcal{L} = v_1 \frac{1}{2}\left(1 + \frac{1}{\beta}\right)\sqrt{\beta c_1}\right\}$

- ***Region 3 (R3):*** $\left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | 2 \geq c_1 \text{ and } \beta < 4/c_1\right\}$ *Conflict and stagnation. In R3 the solution is:* $\left\{a_1 = v_1\left(\frac{c_1}{2}\right)^2, i_1 = 0, \mathcal{L} = v_1\left(1 + \frac{c_1}{4}\right)\right\}$

- ***Region 4 (R4)***:$\left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | c_1 \geq 2,\ \beta < c_1/(c_1 - 1)\right\}$ *Stagnant consolidation. In R4 the solution is:* $\left\{a_1 = v_1, i_1 = 0, \mathcal{L} = v_1\left(2 - \frac{1}{c_1}\right)\right\}$.

Proof: See appendix.

The following figure contains a graphical representation of the solution.

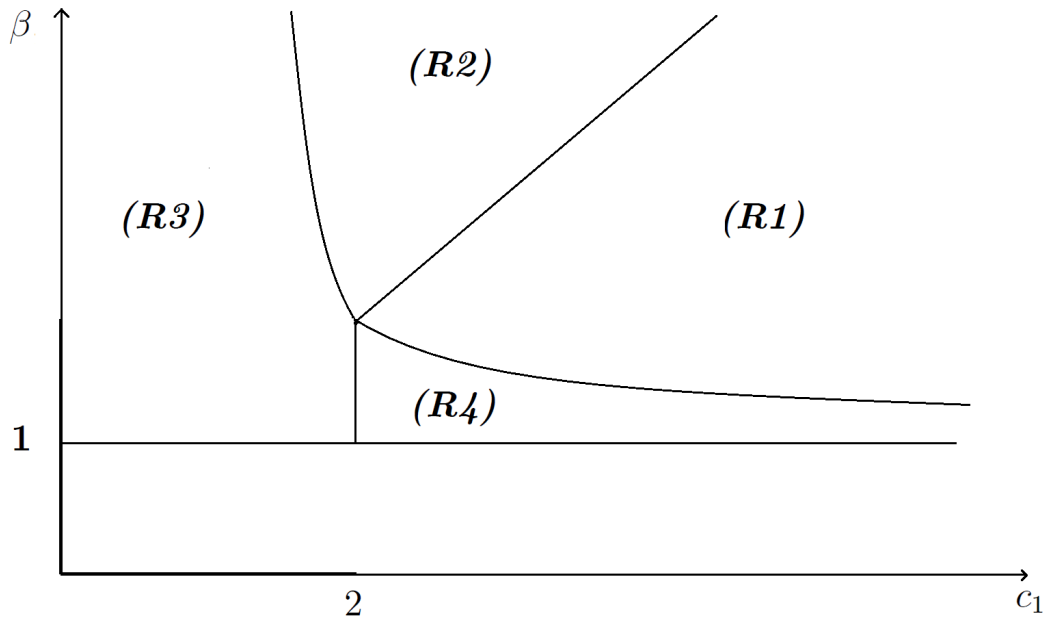

Figure 3.1: Characterization of partition of parameter space, Proposition 1

The payoffs that the incumbent obtains in each region are easily computed by noting that his expected utility at the beginning of period 1 is $v_0 - \frac{a_0}{c_0} - i_0 + \frac{a_0}{a_0+b_0}V_1$.

3.4 Discussion

3.4.1 Properties of equilibrium and lessons for state building

A brief preamble on parameter interpretation

The parameter v_1 tracks properties of the environment (e.g., weather, quality of the soil, topography) that affect the quantity of goods that the economy can produce. If the polity trades, v_1 will also be affected by the price fetched by the goods sold. A convenient feature of this model is that v_1 is a scale parameter and that the optimal decisions by the

incumbent on defense a_t and productive investment i_t are invariant in v_1. (In later stages the parameter v_1 can be studied as a determinant of decisions to invest in improving productivity parameters, c or β). This feature greatly simplifies the characterization of emerging "regimes" as we can restrict attention to the bi-dimensional space (c, β). A brief preamble on how to interpret these parameters is worthwhile at this point.

The two parameters $(c\,, \beta)$ track respectively the productivity of expenditures in defense and productive investment. Thus, β captures anything that increases the returns to productive investments in the asset controlled by the incumbent. For example, β could, like v_1, respond to climatic conditions and other features of the environment, or to the price of goods sold.[3] As for c, it captures anything that yields the incumbent an advantage at producing military firepower at a given expense, such as better military technology or expertise. Note however that changes in infrastructure (e.g., introducing railroads) may affect both β and c: a railroad may increase the returns to investing in a port, and it may also make the incumbent's army more effective. But note that the effects on c may be dependent on context: in some circumstances a railroad may primarily help the incumbent reach rebel strongholds, but it in others it may mostly help rebels reach the capital city. We do not intend to make categorical statements about whether a concrete element–such as a railroad–will help or hinder state consolidation. Rather, we intend to make general but conditional statements on whether changes in β or c can help state building. The mapping of these parameters to concrete forms of investment must be made with reference to specific historical situations.

Characteristics of the equilibrium: the diverging effects of economic shocks and lessons for state building

Given that for $\beta < 1$ investment is never worthwhile, failure to obtain it in equilibrium is obvious and uninteresting. We will focus on the area where $\beta > 1$ where investment is a possibility and where it is the anticipation of conflict and the relative costs of arms and investment that may discourage investment. The main feature of the solution is that all four combinations of order and prosperity can be observed depending on the values of the parameters (c, β). For low values of both military capacity of the incumbent and yield of investment, the polity will be stuck in a situation of economic stagnation and conflict ($R3$). In $R3$ the prospect of conflict lowers the rate of return to investment preventing growth from occurring. If military capacity c is low, an increase in the yield of productive investment is not guaranteed to help much (the region $R3$ extends upwards). If large enough, increasing the yield to productive investment may move the polity to a region ($R2$) where productive investments, and economic growth, occur, but conflict remains: investment yield β may be high enough that despite the possibility of conflict the incumbent wants to invest, but the appeal of seizing the asset fuels the challenger's incentives to fight. Only an even larger military capacity c_1 would move the polity from $R2$

[3]If $v_1 = p.q$, i.e. the value of what the polity produces, and we write $v_2 = v_1 + \beta.p.i = p.q + \beta.p.i$, then changes in p can be captured in our model as changes in both v_1 and β. Changes in the baseline physical capacity of production q will be captured through changes in v_1 exclusively, and changes in the physical returns to investment as changes in β only.

to $R1$, where peace is gained and higher investments and faster growth will occur.[4] Thus, gaining order in addition to the incipient prosperity of $R2$ further enhances prosperity.

An interesting aspect of the model is that shocks to prices of technology that map into increases in β may have very different consequences depending of the value of military capacity. The following remark is a corollary to Proposition 1.

Corollary 1. *If $c_0 \leq 2$ and the polity is coping with disorder and stagnation in $R3$, an increase in β will keep the polity in the same region or make it transition into disorder and growth in $R2$. But if the polity has larger state capacity $c_0 > 2$ and faces peace with stagnation in $R4$, an intermediate increase in β could attain both order and prosperity (in $R1$), while a larger increase in β could make the polity jump to a regime of growth with disorder.*

The last corollary shows that too much of an improvement in the rate of return to investment may be a bad thing if military capacity is large, while it would only help if military capacity is low.

Starting from $R3$, the only way to conquer peace is to augment the military capacity of the incumbent. For moderate increases, peace may be conquered (moving into $R4$), but the cost of the arms that attain deterrence is still high enough that the incumbent cannot channel resources toward investment. Thus, peace is attained but the polity remains stuck in a no-growth regime. In $R3$ the fact that the incumbent may lose power acts as a tax on the returns to investment. In $R4$ the effect that prevents investment is more subtle: the incumbent realizes that investments will expand the voracity of the challenger, and this will trigger too high costs of maintaining the peace. Only large enough increases in military capacity, or a balanced increase in military capacity and investment yield, allow the incumbent to arm to levels that attain deterrence while leaving resources available for promoting investment and growth (a move from $R3$ to $R4$ and then to $R1$).

Note shocks may increase or decrease parameters like β and c_1. A polity that enjoys order and prosperity in $R1$ with a $\beta < 2$ could, through a reduction in c_1, be plunged into stagnation and disorder in $R3$. A reduction in c_1 could be thought of as a negative shock to the incumbent's military technology or as a positive shock to the military technology of the challenger. An interesting example described by [59] is that of the narrowly aristocratic "chariot" civilizations which successfully maintained order and the creation of surplus until iron-made weaponry became available around 1500BC. Because of the ease with which iron could be obtained and worked with, access to weaponry got "democratized." Nomadic herdsmen from the plains obtained weapons, helmets, and shields of iron that could give them a chance against the arrows of the chariot archers. The result was the invasion of the chariot centers by the nomadic tribes and the ensuing of a period of disorder.

This simplified model does not incorporate the destruction of resources brought on by conflict and it represents the limit solution to a more general model where only a fraction

[4] Productive investment is higher in $R1$ than in $R2$ whenever $v\frac{c-1}{c+\beta} > \frac{v}{2}\left(1-\frac{1}{\beta}\right)$ or when $\beta c - \beta + c - \beta^2 > 0$, which is always the case for $c_1 > \beta$, a condition characterizing $R1$.

σ of the asset survives the war. The solution to the expanded model is similar and one can show that for σ low enough, an increase in β that moves the polity from a point of consolidation ($R4$ or $R1$) to one with conflict (in $R2$) may leave the polity worse off. The fact that the "sweet spot" $R1$ is wedge-shaped yields a central insight for state-building contained in the following two remarks.

Remark 1. *From a situation of stagnation and conflict, a large enough increase in military capacity c is a necessary and sufficient condition for successful state-building featuring consolidation and growth.*

We also have,

Remark 2. *Increases in the yield to productive investment β are not necessary nor sufficient condition for successful state building. However, if military capacity is large enough to ensure stagnant state consolidation (to be in R4), the smallest parametric improvement needed to ensure growth involves the returns to productive investment β.*

These remarks help think about a crucial problem in state-building. The consensus in the military community is that counterinsurgency and reconstruction of broken states require substantial development initiatives, in addition to reinstating a military and policy capacity.[5] What should be the balance between the two? If development initiatives are associated with improvements in β, and military assistance is associated with improvements in c, these remarks tell us that improvements in c alone can yield state consolidation and growth, but the cheapest way might be to add improvements in β once a minimum domestic military capacity has been established. The reason is that once the polity is in $R4$, the shortest route to $R1$ is to increase β. This follows from the fact that the frontier between these two regions has slope $\frac{-1}{(c-1)^2}$ which is smaller than -1 whenever $c > 2$.

3.4.2 Endogenous military capacity and the paths to order and prosperity

Setting

Let us now introduce a period zero, before the periods 1 and 2 that we have analyzed so far. This allows us to endogenize military capacity. We will model this by allowing the incumbent to spend resources in one period to increase its military effectiveness in the next period. Since the challenger will never fight in period 2, the incumbent will never spend in expanding military capacity in period 1. Thus, the decision to augment military capacity will be relevant only in period 0. We postulate that in period 0 the incumbent has a military capacity c_0, and can spend an amount m_0 that will take military capacity in the next period to $c_1 = c_0 + \gamma m_0$. To make things interesting, we assume c_0, β are such that if things were left unchanged, in period 1 the incumbent would find himself in region $R3$, which means he can expect disorder and stagnation. In particular, we impose the following,

[5] A famous example of this thinking is the US Army manual on counterinsurgency (Headquarters of the Army 2006).

Assumption 1: $\beta c_0 < 4$ and $c_0 < 2$.

All other aspects of the interaction between challenger and incumbent remain as before.

Timing

In period 0, the incumbent starts by selecting m_0. Then, in each period the incumbent selects a_t and i_t.[6] After observing (a_t, i_t) the challenger selects b_t. If $b_t = 0$, the players retain their positions in the next period. If $b_t > 0$, then there is a war at the end of period t. The winner of the war becomes the incumbent in the next period, and faces a new challenger then.

Payoffs

The fact that there is a new type of expenditure changes the incumbent's budget constraint to $v_0 - m_0 - \frac{a_0}{c_0} - i_0 \geq 0$. And the fact that there is an extra period now implies that a zero arming decision by the challenger in period 1 could open the challenger to vulnerability. So in this three-period model an additional assumption is that the challenger cannot be eliminated.[7] This matches the historical cases of settlers dealing with nomadic raiders, who have vast steppes on which to run away from the forces of the civilized, settled, center.

As before, we solve the model through backward induction. The solution for periods 1 and 2 is given by our analysis in the previous section. That analysis tells us the expected payoff for being an incumbent in period 1 is given by,

$$V_1(i_0, m_0) = (v_0 + \beta i_0) \times \begin{cases} \frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta} & (c_0 + \gamma m_0,\ \beta) \in \mathbf{R1} \\ \sqrt{\frac{(c_0+\gamma m_0)}{\beta}\frac{(1+\beta)}{2}} & (c_0 + \gamma m_0,\ \beta) \in \mathbf{R2} \\ \left(1 + \frac{c_0+\gamma m_0}{4}\right) & (c_0 + \gamma m_0,\ \beta) \in \mathbf{R3} \\ \left(2 - \frac{1}{c_0+\gamma m_0}\right) & (c_0 + \gamma m_0,\ \beta) \in \mathbf{R4} \end{cases} \equiv (v_0 + \beta i_0) S(m_0)$$

Given this continuation value, we can solve for decisions in period 0. After the incumbent has selected m_0, a_0 and i_0, the challenger decides whether to arm himself. Using the same logic as in the previous section, we see that the challenger's best response function is given by,

$$b_0(a_0, m_0, i_0) = \begin{cases} \sqrt{a_0 V_1(i_0, m_0)} - a_0 & if\ a_0 < V_1(i_0, m_0) \\ 0 & if\ a_0 \geq V_1(i_0, m_0) \end{cases}$$

This notation embeds the four regions over which $V_1(i_0, m_0)$ is defined into the calculus of the challenger. Given this best response function, the incumbent has to choose a_0, i_0 after it chose m_0 such that she maximizes her expected utility.

[6] The assumption that m_0 is decided before a_0 and i_0 is immaterial and will just simplify the exposition. It is equivalent to assume that the incumbent selects all three variables simultaneously. What is of course important is that the incumbent makes his choices before the challenger.

[7] If the challenger can be eliminated when selecting zero arming, then it would not be an equilibrium for the challenger to desist from arming itself.

The incumbent maximizes,

$$\max_{a_0,i_0\geq 0} v_0 - m_0 - \frac{a_0}{c_0} - i_0 + \frac{a_0}{a_0 + b_0(a_0, i_0, m_0)} V_1(i_0, m_0)$$

subject to

$$\begin{aligned} v_0 - m_0 - \frac{a_0}{c_0} - i_0 &\geq 0 \ (BC) \\ (v_0 + \beta i_0)S(m_0) - a_0 &\geq 0 \ (ND) \\ a_0 &\geq 0 \\ i_0 &\geq 0 \end{aligned}$$

Notice this problem is similar to the one with two periods in the previous section, except now the continuation value depends explicitly on m_0 (which is fixed at this stage) through $S(m_0)$. The objective function is differentiable in a_0 and i_0. As before, the first order and complementary slackness conditions that characterize the optimum are given by

$$\frac{\partial \mathbf{L}}{\partial a_0} = \frac{1}{2}\sqrt{\frac{(v_0 + \beta i_0)\, S(m_0)}{a_0}} - \frac{1}{c_0} - \frac{\lambda_{BC}}{c_0} - \lambda_{ND} + \lambda_a = 0 \tag{3.12}$$

$$\frac{\partial \mathbf{L}}{\partial i_0} = \frac{\beta}{2}\sqrt{\frac{a_0}{(v_0 + \beta i_0)\, S(m_0)}} - 1 - \lambda_{BC} + \lambda_{ND}\beta S(m_0) + \lambda_i = 0 \tag{3.13}$$

$$\lambda_{BC}(v_0 - m_0 - \frac{a_0}{c_0} - i_0) = 0,\ \lambda_{ND}\left((v_0 + \beta i_0)S(m_0) - a_0\right) = 0,\ \lambda_a a_0 = 0,\ \lambda_i i_0 = 0 \tag{3.14}$$

As before, $\lambda_a = 0$, so $a_0 > 0$, so there are in principle eight possible cases depending on whether the Lagrange multipliers are positive or zero. The following Lemma shows that, given our Assumption 1 there are two feasible cases in period 0.

Lemma 1. *If Assumption 1 holds, then in period 0 the incumbent chooses:*
i) $i_0 = 0$ *and* $a_0 = \frac{c_0^2}{4} v_0 S(m_0)$ *when* $\frac{v_0 S(m_0)}{(v_0 - m_0)} < \frac{4}{c_0}$*; or*
ii) $i_0 = 0$ *and* $a_0 = c_0(v_0 - m_0)$ *when* $\frac{v_0 S(m_0)}{(v_0 - m_0)} \geq \frac{4}{c_0}$.
Proof: See Appendix.

Lemma 1 reveals that no productive investment is carried away on period 0 and that the army size depends on the value of m_0.With this result, we are now equipped to study the incentives of the incumbent to make changes in military capacity m_0. We can trace how those changes will affect period 0 investment and army decisions for both incumbent and challenger and, consequently, future investment, army sizes, peace and prosperity outcomes.

Endogenous military capacity

In our case of interest, for any value of m_0 we are able to compute the incumbent's present expected utility, given our result in Lemma 1 . The effect of m_0 on the incumbent's

utility depends on the initial conditions in period 0. If the maximum utility comes from extremely low m_0 then the incumbent will be in the zone with no growth and war (**R3**) in period 1. On the contrary if the optimal m_0 is extremely high, investment and peace will obtain in period 1. Notice, however, that the path to prosperity (via investments in m_0) depends on β–the return on productive investment. In particular, when $\beta < 2$, the path to peace and prosperity requires going from **R3** to **R1** through **R4**, and when $\beta \geq 2$, it requires going from **R3** to **R1** through **R2** (see Figure 3.1).We analyze these cases in turn.

Case $\beta < 2$

In this case, the incumbent starts at **R3** and might either stay at **R3**, move to **R4** or evolve to **R1** depending on the budget.

Proposition 2. *Under Assumption 1 and provided that* $\beta < 2$, *there exist cutoffs* τ_L, τ_M *and* τ_H, $\tau_L < \tau_M \leq \tau_H$ *such that*

1. If $\gamma v_0 < \tau_L$, *the polity stays in R3 (stagnation and conflict);*

2. If $\tau_H < \gamma v_0$, *the polity moves to R1 (conquers peace and prosperity); and*

3. If $\tau_M < \gamma v_0 < \tau_H$, *the polity moves to R4 (conquers peace)*

Proof: See Appendix.

This proposition tells us that, given the effectiveness γ of expenditures on military capacity, the initial military capacity c_0 and the productivity of investment β, the path followed by the polity will be very different depending on the initial level of prosperity v_0. If v_0 is very low, the polity will remain trapped without order or prosperity. If v_0 lies in an intermediate region, the polity will move into **R4** where it will attain peace, but given the low $\beta < 2$ it will not be able to grow. The reason is that even though it attains a higher military capacity c_1 in the next period, which gives giving the incumbent the ability to fend off attacks at a lower cost, the benefit from consumption will still be higher than the present value from investing. If v_0 is very high, however, the subsequent military capacity c_1 will allow the incumbent to free resources for both a deterrent army and a large-scale investment at **R1**.

Case $\beta \geq 2$

In this case, the incumbent starts at **R3** and might either stay at **R3**, move to **R2** or evolve to **R1**. As before, the following result establishes the conditions on the return of the productive investment and on the productivity of the military capacity investment such that this transitions obtain.

Proposition 3.*Under Assumption 1 and provided that* $\beta \geq 2$, *there exist cutoffs* σ_L, σ_M *and* σ_H, $\sigma_L \leq \sigma_M \leq \sigma_H$ *such that*

1. If $\gamma v_0 < \sigma_L$, *the polity stays in R3 (stagnation and conflict);*

2. If $\sigma_H < \gamma v_0$, *the polity moves to R1 (conquers peace and prosperity); and*

3. If $\sigma_M < \gamma v_0 < \sigma_H$, *the polity moves to R2 (conquers growth)*

Proof: See Appendix.

Thus, the two cases $\beta < 2$ and $\beta \geq 2$ yield a picture with a commonality and a difference. The commonality is that if v_0 is small the polity will remain stagnant and violent while if the initial income is large enough the polity will have enough resources to conquer peace and prosperity. However, if the windfall is intermediate the path of the polity will be different. For $\beta \geq 2$ the polity will conquer growth in period 1 but remain violent, while for $\beta < 2$ the polity will conquer peace in period 1 but remain poor. To summarize, while large enough initial income guarantees order and prosperity through sufficient accumulation of military capacity, intermediate levels will allow to attain either order *or* prosperity. Which one is attained depends on the value of β. Thus, the emerging picture is one that situates the Hobbesian argument in very specific place in the process of attainment of order and prosperity. Not only is order not always a precondition of prosperity (in $R2$ we can have without the other) but it is initial income (as captured by v_0) which allows the polity to augment the military capacity to the level that could ensure order.

3.5 Conclusion

The dismal performance of contemporary "failed states" highlights the challenges facing the process of state formation and consolidation. Failed states can be seen as the symptom of a process of state formation that has not quite taken off. Since the early dawn of history, states have emerged in relation to the accumulation of surplus, either from agriculture or trade. The key test for a state having formed successfully is that it should be able to guarantee a stable productive environment through its ability to monopolize violence internally, and ensure external defense. However, a fundamental challenge must be met. The monopolization of violence requires the dissuasion or suppression of predatory challenges to, or violent contestation for, the authority of the state. The fundamental tension in the process of state formation is the fact that the more successful a state is in promoting the creation of wealth (and capturing a part of it), the stronger the incentives for predatory parties to challenge the authority of the state—this should in turn discourage incumbent authorities from taking actions that promote growth. But then how do states ever form and promote growth? Every process of state formation must break the "trap" of the predatory incentives that become amplified by its own success.

We cast this problem in the context of a simple model where an incumbent authority must decide whether to invest in growing a productive asset or not, but where a challenger may attack the incumbent to gain control of that asset. One might think that a natural candidate to resolve the fundamental tension is the fact that as a state gets wealthier it gains a financial advantage over the challenger that must make the latter eventually desist, as its chances of prevailing in conflict become lower. However, contest settings cleanly isolate the nature of the fundamental tension: the wealthier the incumbent gets, the stronger the incentive for a challenge, unless the incumbent develops such a mili-

tary power that the challenger refrains from contesting altogether. The pivotal element behind the resolution of the fundamental tension in our model is the possibility of investing in military capacity. This is an investment that alters the relative productivities of resources applied to conflict between the parties. This possibility is a necessary but not sufficient condition for states to consolidate, however. We characterize conditions for the initial value of the productive asset, the costs of investments in military capacity, and the returns to productive investments, such that states may accumulate military and productive capacities, conquer peace, and "take off." Conversely, we can also characterize conditions for "failed states" where accumulation of productive assets is hindered and conflict persists.

Bibliography

[1] Andreoni, J., and Miller, J. Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica 70*, 2 (2002), 737–753.

[2] Andreoni, J., and Miller, J. H. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal 103*, 418 (1993), 570–585.

[3] Arbak, E., and Villeval, M.-C. Endogenous leadership: selection and influence.

[4] Avolio, B. J. Promoting more integrative strategies for leadership theory-building. *American Psychologist 62*, 1 (2007), 25.

[5] Avolio, B. J., Sosik, J. J., Jung, D. I., and Berson, Y. Leadership models, methods, and applications. *Handbook of psychology* (2003).

[6] Bandiera, O., Barankay, I., and Rasul, I. Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics 120*, 3 (2005), 917–962.

[7] Bárány, I. Fair distribution protocols or how the players replace fortune. *Mathematics of Operations Research 17*, 2 (1992), 327–340.

[8] Bass, B. M. The leaderless group discussion. *Psychological Bulletin 51*, 5 (1954), 465.

[9] Bass, B. M., and Stogdill, R. M. Handbook of leadership. *Theory, Research & Managerial Applications. New York, The free press* (1990).

[10] Besley, T., and Persson, T. *Pillars of prosperity: The political economics of development clusters*. Princeton University Press, 2011.

[11] Blanco, M., Engelmann, D., and Normann, H. T. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior 72*, 2 (2011), 321–338.

[12] Blattman, C., and Miguel, E. Civil war. *Journal of Economic Literature* (2010), 3–57.

[13] Boehm, C., Barclay, H. B., Dentan, R. K., Dupre, M.-C., Hill, J. D., Kent, S., Knauft, B. M., Otterbein, K. F., and Rayner, S. Egalitarian behavior and reverse dominance hierarchy [and comments and reply]. *Current Anthropology 34*, 3 (1993), 227–254.

[14] Bolton, G. E., and Ockenfels, A. Erc: A theory of equity, reciprocity, and competition. *American economic review* (2000), 166–193.

[15] Bruttel, L. V., and Fischbacher, U. *Taking the initiative: what motivates leaders?* Bibliothek der Universität Konstanz, 2010.

[16] Burns, J. M. leadership. ny, 1978.

[17] Calvert, R. Leadership and its basis in problems of social coordination. *International Political Science Review 13*, 1 (1992), 7–24.

[18] Charness, G., and Dufwenberg, M. Promises and partnership. *Econometrica 74*, 6 (2006), 1579–1601.

[19] Charness, G., and Rabin, M. Understanding social preferences with simple tests. *The Quarterly Journal of Economics 117*, 3 (2002), 817–869.

[20] Collier, P., and Hoeffler, A. On economic causes of civil war. *Oxford economic papers 50*, 4 (1998), 563–573.

[21] Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. Communication in coordination games. *The Quarterly Journal of Economics 107*, 2 (1992), 739–771.

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

[23] Crawford, V. A survey of experiments on communication via cheap talk. *Journal of Economic theory 78*, 2 (1998), 286–298.

[24] Croson, R., and Gneezy, U. Gender differences in preferences. *Journal of Economic Literature* (2009), 448–474.

[25] Dal Bó, E., and Dal Bó, P. Workers, warriors, and criminals: social conflict in general equilibrium. *Journal of the European Economic Association 9*, 4 (2011), 646–677.

[26] Dal Bo, P. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *The American Economic Review 95*, 5 (2005), 1591–1604.

[27] De Oliveira, A., Croson, R. T., and Eckel, C. Are preferences stable across domains? an experimental investigation of social preferences in the field. *An Experimental Investigation of Social Preferences in the Field (January 26, 2009)* (2011).

[28] DeRue, D. S., Nahrgang, J. D., Wellman, N., and Humphrey, S. E. Trait and behavioral theories of leadership: An integration and meta-analytic test of their relative validity. *Personnel Psychology 64*, 1 (2011), 7–52.

[29] Dreber, A., Fudenberg, D., and Rand, D. Who cooperates in repeated games? *Available at SSRN 1752366* (2011).

[30] Dube, O., and Vargas, J. F. *Are All Resources Cursed?: Coffee, Oil and Armed Conflict in Columbia*. Center for International Development at Harvard University, 2006.

[31] Duffy, J., and Muñoz-García, F. Patience or fairness? analyzing social preferences in repeated games. *Games 3*, 1 (2012), 56–77.

[32] Ellingsen, T., and Johannesson, M. Promises, threats and fairness*. *The Economic Journal 114*, 495 (2004), 397–420.

[33] Erkal, N., Gangadharan, L., and Nikiforakis, N. *Relative earnings and giving in a real-effort experiment*. Department of Economics, University of Melbourne, 2010.

[34] Eysenck, H. J. The concept of intelligence: Useful or useless? *Intelligence 12*, 1 (1988), 1–16.

[35] Fehr, E., and Fischbacher, U. Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The economic journal 112*, 478 (2002), C1–C33.

[36] Fischbacher, U. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics 10*, 2 (2007), 171–178.

[37] Fisman, R., Kariv, S., and Markovits, D. Individual preferences for giving. *Yale Law & Economics Research Paper*, 306 (2005).

[38] Forges, F. Can sunspots replace a mediator? *Journal of Mathematical Economics 17*, 4 (1988), 347–368.

[39] Foss, N. J. Leadership, beliefs and coordination: An explorative discussion. *Industrial and Corporate Change 10*, 2 (2001), 357–388.

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

[41] Gächter, S., and Renner, E. Leading by example in the presence of free rider incentives. In *a Conference on Leadership* (2003).

[42] Gneezy, U. Deception: The role of consequences. *The American Economic Review 95*, 1 (2005), 384–394.

[43] Gough, H. G. Testing for leadership with the california psychological inventory.

[44] Güth, W., Levati, M. V., Sutter, M., and Van Der Heijden, E. Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics 91*, 5 (2007), 1023–1042.

[45] Hemphill, J. K. *Leader behavior description*. Ohio State University, 1950.

[46] Hermalin, B. Toward an economic theory of leadership: Leading by example. *Available at SSRN 15570* (1997).

[47] Hermalin, B. Leadership and corporate culture. *Handbook of Organizational Economics* (2013), 432–478.

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

[49] Iriberri, N. Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Available at SSRN 1374287* (2009).

[50] John, O. P., Naumann, L. P., and Soto, C. J. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research 3* (2008), 114–158.

[51] Judge, T. A., Bono, J. E., Ilies, R., and Gerhardt, M. W. Personality and leadership: a qualitative and quantitative review. *Journal of applied psychology 87*, 4 (2002), 765.

[52] Kato, T., and Shu, P. Competition, group identity, and social networks in the workplace: Evidence from a chinese textile firm.

[53] Kreps, D. M. *Corporate culture and economic theory*. Oxford Management Readers, Oxford University Press, Oxford and New York, 1996.

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

[55] Lazear, E. P., and Rosen, S. Rank-order tournaments as optimum labor contracts, 1979.

[56] Ledyard, J. O. Public goods: A survey of experimental research. Tech. rep., EconWPA, 1994.

[57] Lewis, H. S. *Leaders and followers: Some anthropological perspectives*. No. 50. Addison-Wesley, 1974.

[58] Mas, A., and Moretti, E. Peers at work. Tech. rep., National Bureau of Economic Research, 2006.

[59] McNeill, W. The pursuit of power: Technology, armed force, and society since ad 1000.

[60] Meidinger, C., Villeval, M.-C., et al. Leadership in teams: Signaling or reciprocating?

[61] Miettinen, T., and Suetens, S. Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution 52*, 6 (2008), 945–960.

[62] Moxnes, E., and Van der Heijden, E. The effect of leadership in a public bad experiment. *Journal of Conflict Resolution 47*, 6 (2003), 773–795.

[63] Nagl, J. A., Amos, J. F., Sewall, S., Petraeus, D. H., et al. *The US Army/Marine Corps Counterinsurgency Field Manual*. No. 3-24. University of Chicago Press, 2008.

[64] Potters, J., Sefton, M., and Vesterlund, L. After you: endogenous sequencing in voluntary contribution games. *Journal of Public Economics 89*, 8 (2005), 1399–1419.

[65] Powell, R., Oliveros, L. S., and Shapiro, J. Deterring and defending against strategic attackers: Deciding how much to spend and on what. Tech. rep., Technical Report April, Travers Department of Political Science, University of California Berkley, 2008.

[66] Rivas, M. F., and Sutter, M. The benefits of voluntary leadership in experimental public goods games. *Economics Letters 112*, 2 (2011), 176–178.

[67] Ross, M. L. The natural resource curse: How wealth can make you poor. *Natural resources and violent conflict: options and actions* (2003), 17–42.

[68] Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied 80*, 1 (1966), 1–28.

[69] Sally, D. Conversation and cooperation in social dilemmas a meta-analysis of experiments from 1958 to 1992. *Rationality and society 7*, 1 (1995), 58–92.

[70] Stogdill, R. M. Manual for the leader behavior description questionnaire-form xii. *Columbus: Ohio State University, Bureau of Business Research* (1963).

[71] Urbano, A., and Vila, J. E. Computational complexity and communication: Coordination in two–player games. *Econometrica 70*, 5 (2002), 1893–1927.

[72] Van den Assem, M. J., Van Dolder, D., and Thaler, R. H. Split or steal? cooperative behavior when the stakes are large. *Management Science 58*, 1 (2012), 2–20.

[73] Van Vugt, M. Evolutionary origins of leadership and followership. *Personality and Social Psychology Review 10*, 4 (2006), 354–371.

[74] Varian, H. R. *Microeconomic analysis*, vol. 2. Norton New York, 1992.

Appendix A

On the Origins of Leadership Through Communication: The Role of Context and Social Preferences

A. Instructions

This is an experiment in the economics of decision-making. If you follow these simple instructions carefully and make good decisions, you could earn a considerable amount of money. The currency we will use throughout the instructions and the experiment is the Berkeley Buck. We will denote it as "$" and the exchange rate is $ 12 per US$ dollar. Please be aware that we do not expect any particular behavior from you or any other participant.

This session will be divided in 4 blocks, each comprising a series of situations in which you will have to make decisions. In what follows we will describe these blocks in chronological order as they will appear in your computer screen.

In block 1, you will face 4 situations. In each situation you will have to allocate a total of 10 tokens between yourself and another participant. The tokens may have different values in each situation. The other participant will be selected randomly at the end of the experiment. Also, neither you nor the other participant will receive information about each other's identity and about your decision in each situation. The computer will randomly select 1 of the 4 situations (with equal chance) to compute your payout based on your decisions. Symmetrically, at the end of the experiment you will be randomly matched to the decision made by another participant. Thus, you will have two ways of earning money from these situations. The first is from your allocation decision, and the second is from the allocation decision of another randomly matched participant.

Block 2 consists of one situation. You will have to send a message to another participant. This message will be available to the other randomly assigned participant at the end of the experiment. After he/she reads the message in his/her screen, he/she will make a choice. Your payout from this situation depends on the choice of this other participant.

Block 3 consists of one situation. You will have to allocate a sum of money between yourself and another participant. The other participant will be assigned randomly at the end of the experiment. In this situation your allocation may be reversed by the computer with some probability specified in the corresponding screen. Neither you nor the other participant will be told about each other's identity and about your decision. Thus, you will have two ways of earning money from this situation. The first is from your allocation decision (depending on whether is reversed or not), and the second is from the allocation decision of the other randomly matched participant. <STOP READING HERE>

In block 4 you will face 12 situations. Each situation will contain two types of scenarios: a non-interactive and an interactive scenario. In the non-interactive scenario, your decisions will be matched at the end of the experiment. In the interactive scenario your decisions will be matched immediately with a different participant in each round.

In the non-interactive scenario you will choose between playing two alternatives, A or B. The following screen shot shows an example:

[ADD SCREENSHOT OSPD HERE]

Each of these scenarios will feature different combinations of possible payoffs. In every scenario, you will be the Row player so your payoffs are the ones on the left of each cell. Both you and the other participant will have two possible choices. You can choose A or you can choose B. In this example:

- If you both choose A you will both get a payoff of $5.
- If you both choose B you will both get a payoff of $9.

• If you choose A, but the other participant chooses B, you will get a payoff of $15, but the other player will receive $4.

• Likewise, if you choose B, but the other participant chooses A, then you receive $4 and the other participant receives $15.

When choosing your move, you will not know the decision of the other participant. The other participant will not know your decision. For each scenario, your decision will be matched with the decision of another randomly selected participant at the end of the experiment. <STOP READING HERE>

After you decide between A and B, you will have to make a prediction about other participants' decisions. You will have to forecast how many of the other 23 participants in the experiment will choose either A or B in each scenario. You will be rewarded for the accuracy of your predictions. The formula to compute your reward is:

Maximum{0 , 8 – 1 * (Distance Between Your Prediction and Actual Decisions Other Participants)}

To explain the formula, we will focus on the prediction about A (because the prediction about B is 23 minus the Prediction about A). If your prediction coincides with the actual value you will get $8. An amount of $1 will be deducted for each unit above or below the actual number of participants who chose A (or equivalently B). Thus, if your prediction is more than 7 units away from the actual number of participants choosing A, you will receive $0. This calculation will be performed for each of the 12 scenarios. <STOP READING HERE>

From these 12 choices and predictions of each scenario, the computer will randomly select 6 to calculate the payouts. We emphasize that all the decisions described so far will be matched at the end of the experiment, so your actions will not affect the actions of other participants.

In the interactive scenarios, you will be randomly matched with a new participant in each of the 12 scenarios. We call each of these scenarios an Interaction. In each Interaction you will be matched with a different participant, so you will interact with one participant only once. Also, this participant will never be matched with any of the participants you will be matched with.

Each Interaction consists of two screens. In the first screen you will see the payoffs of the game and you may use the chat box on the left to communicate with the participant you are matched with in that Interaction. This screen will be shown for 30 seconds:

[ADD SCREENSHOT OF CHAT HERE]

In the second screen, you will have the opportunity to select your action. There will be no chat and you will have 30 seconds to make your decision. This is an example of a screen shot:

[ADD SCREENSHOT OF SIMULTANEOUS GAME HERE]

After each Interaction you will observe your and the other participants' choice (A or B) as well as your and other participant's payoffs. Your payoff in each Interaction will be based on your decision and the decision of the matched participant you interacted with. Your final payout will be calculated by the computer using 6 randomly selected Interactions (with equal chance) out of the 12. Recall that each situation in this block 4, comprising one non-interactive and one interactive scenario, will be repeated 12 times. <STOP READING HERE>

Finally, at the end of the experiment, you will read the message a randomly matched participant sent you, and you will have to make a choice. You will also be given questionnaires that can yield some additional payoffs. After all questionnaires are completed, the computer will match the decisions of all the situations except for the Interactions (which were already

matched in each of the 12 scenarios), and final payments will be made to each of you.

Each screen you see throughout the experiment has all the instructions necessary for the decision on that screen. Recall, during the session, all payoffs are expressed in terms of Berkeley Bucks. However, at the end of the session, all of your Berkeley Bucks will be converted at 12 Berkeley Bucks to 1 US$. Thus, in US dollars, your final payment will be between $5 and $30, depending on how you do.

Recall that at no time your true identity nor your final payout will be revealed to the other participants in this experiment.

Thank you very much and good luck!

B. Screenshot Lying aversion elicitation

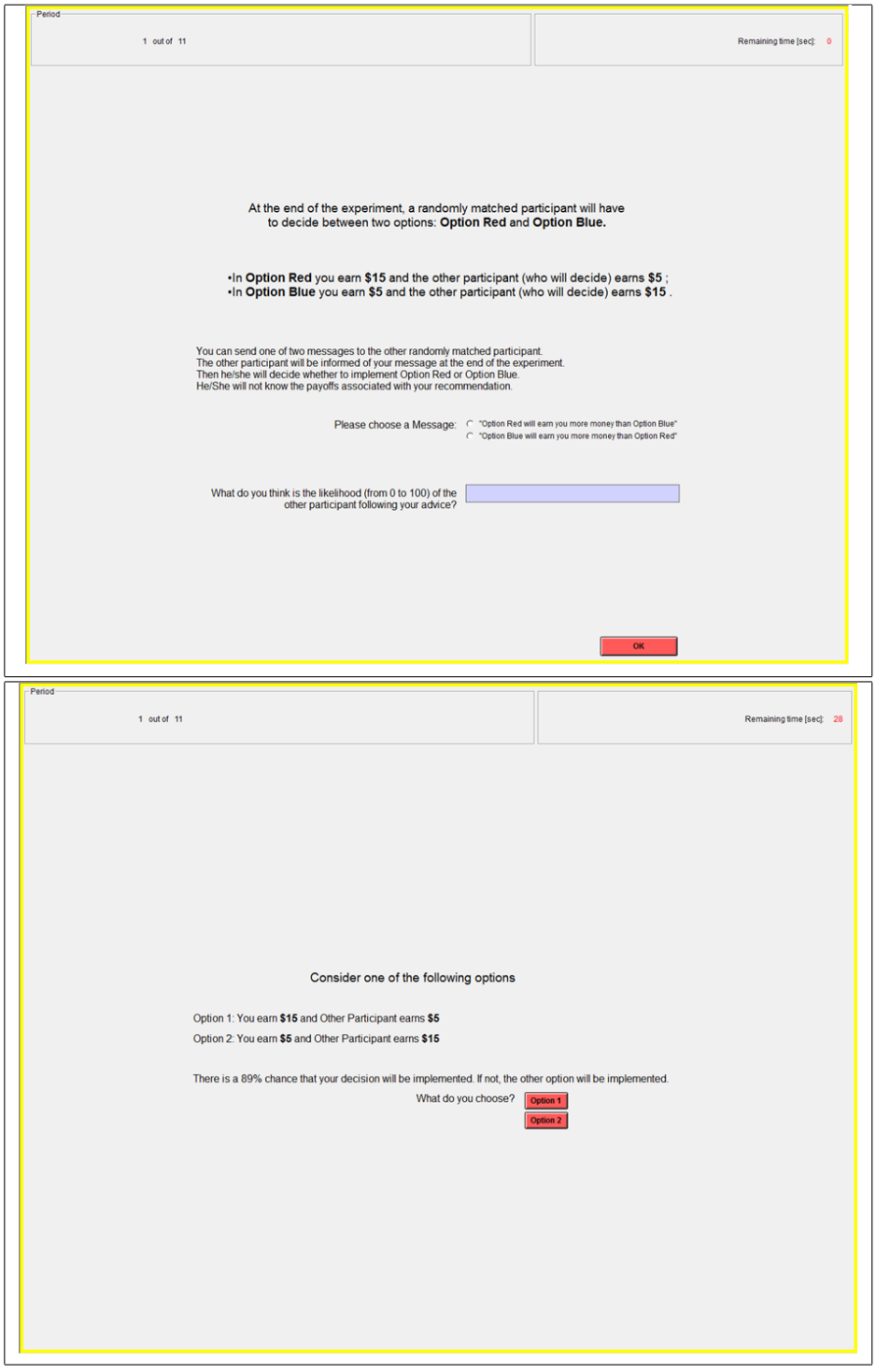

Figure A.1: Screenshots lying aversion elicitation.

C. Treatment screen shots

Figure A.2: Screen shots, treatments.

D. Screenshots Risk Aversion test

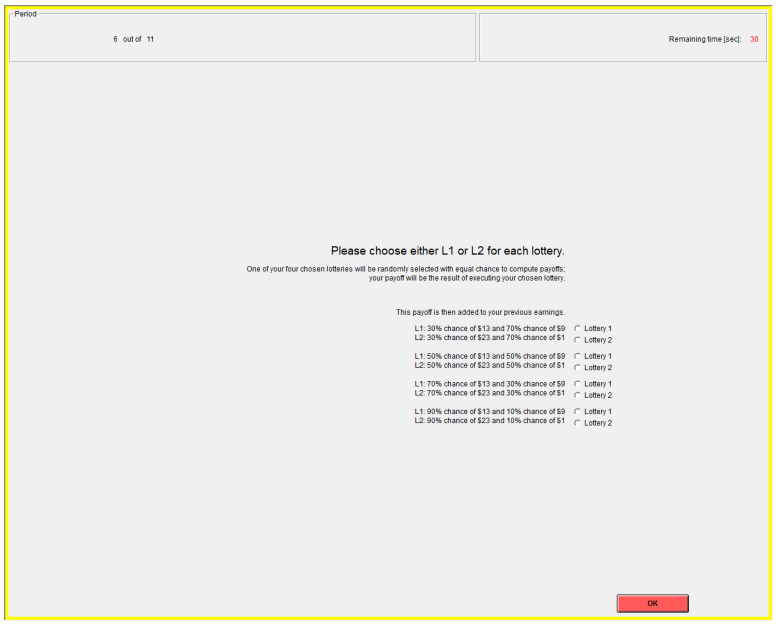

Figure A.3: Screen shot, risk-aversion test.

E. Cooperation and own initiative

	(1) SH Pr{Coop.}	(2) SH Pr{Coop.}	(3) PD Pr{Coop.}	(4) PD Pr{Coop.}
Initiate	0.82***	0.71***	0.64***	0.68***
	(0.19)	(0.16)	(0.15)	(0.14)
High LA	-1.50	-1.17	-0.29	-0.17
	(1.25)	(0.75)	(0.30)	(0.33)
High RA	0.01*	-0.01	0.02***	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
RA x LA	0.21	0.09	0.03***	0.04***
	(0.16)	(0.10)	(0.01)	(0.01)
_cons	0.31	3.62	-0.92***	-5.40***
	(0.27)	(2.81)	(0.19)	(1.38)
N	576	540	576	540
CONTROLS	NO	YES	NO	YES
pseudo R^2	0.122	0.319	0.169	0.311

S.E. in parentheses
* $p < 0.10$, ** $p < 0.05$,
***$p < 0.01$

Table A.1: Reduced form of cooperation on initiative, reciprocal altruism, lying-aversion and controls.

F. Cooperation and other's initiative

	(1) CG Pr{d_i=C}	(2) CG Pr{d_i=C}	(3) PD Pr{d_i=C}	(4) PD Pr{d_i=C}
Other Initiate	-0.05	0.19	0.30***	0.23**
	(0.15)	(0.16)	(0.10)	(0.11)
Low Reciprocal-A. High Lying-A.	0.08	-0.57**	-0.30	-0.09
	(0.42)	(0.28)	(0.38)	(0.45)
High Reciprocal-A. Low Lying-A.	0.05	-0.37	0.49*	0.19
	(0.37)	(0.43)	(0.27)	(0.35)
High Reciprocal-A. High Lying-A.	0.07	-0.60*	1.11***	1.46***
	(0.48)	(0.35)	(0.35)	(0.42)
_cons	0.88***	4.88*	-0.82***	-5.03***
	(0.25)	(2.87)	(0.19)	(1.54)
N	576	540	576	540
CONTROLS	NO	YES	NO	YES
pseudo R^2	0.001	0.278	0.099	0.242

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Reduced form of cooperation on other's initiative, reciprocal altruism, lying-aversion and controls.

G. Difference in attributes, leaders and followers

	Leader N	Leader mean	Follower N	Follower mean	diff	se	p
Reciprocal altruism	174	16.30	174	16.63	-0.33	2.13	0.88
High reciprocal altruism	174	0.51	174	0.50	0.01	0.05	0.75
Lying Aversion	174	0.40	174	0.44	-0.04	0.05	0.45
PerfectlySelfish	174	0.22	174	0.20	0.02	0.04	0.60
InternalLocusofControl	174	6.61	174	6.69	-0.07	0.25	0.76
Extraversion	174	3.06	174	3.07	-0.01	0.09	0.87
Agreeableness	174	3.71	174	3.60	0.11	0.05	0.05**
Conscientiousness	174	3.43	174	3.37	0.07	0.07	0.32
Neuroticism	174	2.83	174	2.83	-0.00	0.07	0.97
Openness	174	3.42	174	3.50	-0.08	0.06	0.23
ScoreCRT	174	1.45	174	1.55	-0.10	0.12	0.40
RiskAversion	164	3.27	157	3.35	-0.08	0.10	0.44
female	174	0.76	174	0.70	0.07	0.05	0.15
Asian	174	0.74	174	0.67	0.06	0.05	0.20
White	174	0.16	174	0.22	-0.07	0.04	0.10
OtherEthnicity	174	0.11	174	0.10	0.01	0.03	0.86

Table A.3: t-tests, difference in means, SH treatment

	Leader N	Leader mean	Follower N	Follower mean	diff	se	p
Reciprocal altruism	143	21.35	143	17.91	3.44	2.51	0.17
High reciprocal altruism	143	0.67	143	0.53	0.14	0.60	0.02**
Lying Aversion	143	0.31	143	0.41	-0.09	0.06	0.11
PerfectlySelfish	143	0.29	143	0.34	-0.04	0.06	0.45
InternalLocusofControl	143	6.37	143	6.26	0.11	0.28	0.69
Extraversion	143	3.18	143	3.21	-0.03	0.09	0.73
Agreeableness	143	3.76	143	3.71	0.04	0.07	0.53
Conscientiousness	143	3.64	143	3.49	0.15	0.08	0.06*
Neuroticism	143	2.75	143	2.85	-0.10	0.09	0.25
Openness	143	3.51	143	3.44	0.07	0.07	0.31
ScoreCRT	143	1.23	143	1.44	-0.21	0.14	0.12
RiskAversion	143	3.48	141	3.54	-0.06	0.13	0.66
female	138	0.66	139	0.65	0.00	0.06	0.93
Asian	143	0.61	143	0.66	-0.05	0.06	0.39
White	143	0.24	143	0.25	-0.01	0.05	0.89
OtherEthnicity	143	0.15	143	0.09	0.06	0.04	0.14

Table A.4: t-tests, difference in means, PD treatment

H. Equilibrium with pre-play communciation and purely monetary payoffs, SH and PD treatments

In order to prove Proposition 1, we first introduce one useful construct and state and prove two lemmas. In technical terms, the subset of correlated equilibria corresponding to the convex combinations of pure strategy Nash equilibrium in the one shot game is the correlated equilibria in which the underlying distribution is "diagonal." A diagonal distribution means that knowledge of the corresponding action by one player perfectly determines the outcome of the game. We will use this concept, so we define it formally.

Definition 1 (Barany 1992).

In a two player game, let S_1 and S_2 the action space in a normal form game. A distribution π is said to be diagonal if for each $a_1 \in A_1$ not strictly dominated, there is only one $a_2 \in A_2$ with $\pi(a_1, a_2) > 0$; and for each $a_2 \in A_2$ not strictly dominated, there is only one $a_1 \in A_1$ with $\pi(a_1, a_2) > 0$.

A diagonal distribution can also be interpreted as a distribution in which the signals each player receives in a correlated equilibrium are perfectly correlated: conditional on

being prescribed by the mediator to play a_i, individual i knows with certainty what individual j is prescribed to do, $i, j = 1, 2$.

In order to apply Barany's result, we need to characterize the subset of correlated equilibria with diagonal distributions for each of our simple games. We will focus on direct mechanisms, in the sense that the set of states correspond to the outcomes of the game (C for Cooperation, D for defection): $S = \{C, D\} \times \{C, D\}$, and the equilibrium partition associated to each individual $i = 1, 2$ is $P_i = \{\{(C, C), (C, D)\}, \{(D, C), (D, D)\}\}$. We first characterize the equilibrium for our SH game.

Lemma 1.

In the SH game without pre-play communication, the correlated equilibria with diagonal distribution is characterized as follows:

1. A probability space (S, π), such that $\pi(C, C) = p$, $\pi(D, D) = 1 - p$, and $\pi(C, D) = \pi(D, C) = 0$, for $p \in [0, 1]$.

2. An information partition for each player $i = 1, 2$

$$P_i = \Big\{\{(C, C), (C, D)\}, \{(D, C), (D, D)\}\Big\}.$$

3. Equilibrium actions for each player $i = 1, 2$: $\sigma_i(C, C) = \sigma_i(C, D) = C$ and $\sigma_i(D, C) = \sigma_i(D, D) = D$.

Proof. Fix $p \in [0, 1]$. This is a correlated equilibrium because when player 1 observes signal $\{(C, C), (C, D)\}$, the conditional probability player 2 plays C is equal to 1 and the conditional probability player 2 plays D is 0 (the distribution is diagonal, so the signals are perfectly correlated). Player 1's best response is to play C. When player 1 observes signal $\{(D, C), (D, D)\}$, then the conditional probability of player 2 playing C is 0 and the conditional probability of player 2 playing D is 1. Player 1's best response in this case is D. If $p = 0$, then player 1 will be informed only on $\{(D, C), (D, D)\}$, so his best response is D and if $p = 1$ player 1 will only receive the signal corresponding to $\{(C, C), (C, D)\}$, so his best response is C. Similarly for player 2.

The distribution π is diagonal by construction. The only thing we need to check is whether there is another diagonal distribution $\pi' \neq \pi$ for all $p \in [0, 1]$ that is also part of a correlated equilibrium. Let us assume there is one. For $\pi' \neq \pi$ for all $p \in [0, 1]$ to be diagonal, there must be $q \in [0, 1]$ such that $\pi'(C, D) = q$, $\pi'(D, C) = (1 - q)$, and $\pi'(C, C) = \pi'(D, D) = 0$ (that is, only the outcomes on the anti-diagonal occur with positive probability). If that is the case, then the best response of player 1 to the signal $\{(C, C), (C, D)\}$ is D, because the probability that player 2 plays C conditional on the signal prescribing player 1 to play C is equal to 0. This holds for any $q \in (0, 1)$. If $q = 0$, then player 1's best response to signal $\{(D, C), (D, D)\}$ is C instead of D and if $q = 1$, player 1's best response to signal $\{(C, C), (C, D)\}$ is D instead of C. This contradicts the fact π' is part of a correlated equilibrium.□

The next Lemma characterizes the correlated equilibrium with diagonal distribution for our PD game. In this case, only mutual defection obtains.

Lemma 2.

In the PD game without pre-play communication, the correlated equilibria with diagonal distribution is characterized as follows:

1. A probability space (S, π)*, such that* $\pi(D, D) = 1$*, and* $\pi(C, C) = \pi(C, D) = \pi(D, C) = 0$*, for* $p \in [0, 1]$*.*

2. An information partition for each player $i = 1, 2$

$$P_i = \Big\{\{(C, C), (C, D)\}, \{(D, C), (D, D)\}\Big\}.$$

3. Equilibrium actions for each player $i = 1, 2$ $\sigma_i(C, C) = \sigma_i(C, D) = D$ *and* $\sigma_i(D, C) = \sigma_i(D, D) = D$*.*

Proof. This outcome coincides with the Nash equilibrium outcome and any Nash equilibrium is a correlated equilibrium. The distribution π is diagonal by construction. This is the unique correlated equilibrium (with respect to a direct mechanism), therefore is the unique correlated equilibrium (with respect to a direct mechanism) with diagonal distribution.□

Proposition 1.

- *In the SH game with pre-play communication:*
  - *The probability of mutual cooperation* (C, C) *is* p *and of mutual defection* (D, D) *is* $1 - p$*, with* $p \in [0, 1]$*. Outcomes* (D, C) *and* (C, D) *occur with* 0 *probability.*
  - *Every time player* $i = 1, 2$ *is prescribed to play* C *she plays* C *and when player* i *is prescribed to play* D *she plays* D*.*
- *In the PD game with pre-play communication:*
  - *The probability of mutual defection* (D, D) *is* 1*. Outcomes* (C, C)*,* (D, C) *and* (C, D) *occur with* 0 *probability.*
  - *Player* $i = 1, 2$ *plays defection* D *regardless of what she is prescribed to do.*

Proof. According to Barany (1992, Theorem 3 and Theorem 6), Lemma 1 implies the result for the SH game with pre-play communication and Lemma 2 the result for the PD game with pre-play communication.□

Notice that pre-play communication in our SH game may yield to any convex combination of the equilibrium outcomes in the one shot game, in particular mutual cooperation. Lemma 2, however, implies that in our PD game pre-play communication cannot yield mutual cooperation.

Appendix B

Paths to Order and Prosperity: State Formation with Endogenous Coercive and Productive Capacities

Proofs

Proof of Proposition 1 : Given that $\lambda_a = 0$, we have eight possible cases given by whether the three remaining Lagrange multipliers $\lambda_{BC}, \lambda_{DC}$, and λ_i are zero or positive. We analyze each one of them.

1. Case $\lambda_{BC} > 0$ (BC binds), $\lambda_{DC} > 0$ (DC binds, consolidation),and $\lambda_i = 0$ $(i_1 > 0)$

It proves useful to determine for which subset $\{(c_1, \beta, v_1)|c_1, \beta, v_1 \geq 0\}$ all the conditions (3.9), (3.10) and (3.11) hold. This is

$$\begin{aligned} \frac{1}{2}\sqrt{1} - \frac{1}{c_1} - \frac{\lambda_{BC}}{c_1} - \lambda_{DC} &= 0 \\ \frac{\beta}{2}\sqrt{1} - 1 - \lambda_{BC} + \lambda_{DC}\beta &= 0 \\ v_1 - a_1 + \beta i_1 &= 0 \\ v_1 - \frac{a_1}{c_1} - i_1 &= 0 \end{aligned}$$

In this case, investment and army are given by

$$\begin{aligned} i_1 &= v_1 \frac{(c_1 - 1)}{(c_1 + \beta)} \\ a_1 &= v_1 \frac{c_1(1 + \beta)}{(c_1 + \beta)} \end{aligned} \tag{B.1}$$

As a result $\lambda_i = 0$ (or $i_1 > 0$) is supported by $c_1 > 1$. After some algebra (using the first two equations) we find that $\lambda_{DC} > 0$ if and only if $c_1 > \beta$ and $\lambda_{BC} > 0$ if and only if $\beta > c_1/(c_1 - 1)$. Therefore the parameter set such that this is the solution to the incumbent's problem in period 1 is given by

$$\mathbf{R1} = \left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | c_1 > \beta,\, \beta > c_1/(c_1 - 1)\, and\, c_1 > 1\right\},$$

and in this area there is investment and deterrence of the challenger, yielding consolidation.

The expected utility in period 1 is in each case easily computed by substituting the solutions into the Lagrangian. In this first case expected utility is,

$$E[U_1] = v_1 \frac{c_1(1 + \beta)}{(c_1 + \beta)}.$$

2. Case $\lambda_{BC} > 0$ (BC binds), $\lambda_{DC} = 0$ (DC does not bind, conflict),and $\lambda_i = 0$ $(i > 0)$

Again we check the first order and complementary slackness conditions to see for which parameter set this case contains the solution. The relevant conditions are,

$$\frac{1}{2}\sqrt{\frac{v_1+\beta i_1}{a_1}}-\frac{1}{c_1}-\frac{\lambda_{BC}}{c_1}=0$$
$$\frac{\beta}{2}\sqrt{\frac{a_1}{v_1+\beta i_1}}-1-\lambda_{BC}=0$$
$$v_1-\frac{a_1}{c_1}-i_1=0$$

Investment and army solutions are respectively given by,

$$i_1=\frac{v_1}{2}\left(1-\frac{1}{\beta}\right)$$
$$a_1=\frac{c_1 v_1}{2}\left(1+\frac{1}{\beta}\right).$$

This solution is consistent with $\lambda_{DC}=0$ (satisfies (DC) with strict inequality) and $\lambda_i=0$ if and only if $\beta>c_1$ and $\beta\geq 1$. It is consistent with $\lambda_{BC}>0$ if and only if $\beta\geq 4/c_1$ (this comes from checking the conditions such that $\lambda_{BC}>0$ in the first two equations). As a result, the parameter set for which these are the optimal army and investment is given by

$$\mathbf{R2}=\left\{(c_1,\beta,v_1)\in\mathbb{R}^3_+|\beta>c_1,\ \beta\geq 4/c_1\ and\ \beta\geq 1\right\}$$

Expected utility for the incumbent in this case is,

$$E[U_1]=\frac{v_1}{2}\left(1+\frac{1}{\beta}\right)\sqrt{\beta c_1}$$

3. Case $\lambda_{BC}=0$ (BC does not bind), $\lambda_{DC}=0$ (DC does not bind, conflict),and $\lambda_i=0$ ($i_1>0$)

Non-generic, in the sense that it is consistent only over a subset of the space (β,c_1,v_1) that has measure zero. Proof to be included.

4. Case $\lambda_{BC}=0$ (BC does not bind), $\lambda_{DC}>0$ (DC binds, consolidation),and $\lambda_i=0$ ($i_1>0$)

Non-generic. Proof to be included.

5. Case $\lambda_{BC}>0$ (BC binds), $\lambda_{DC}>0$ (DC binds, consolidation),and $\lambda_i>0$ ($i_1=0$)

Non-generic. Proof to be included.

6. Case $\lambda_{BC}>0$ (BC binds), $\lambda_{DC}=0$ (DC does not bind, conflict),and $\lambda_i>0$ ($i_1=0$)

Not feasible. This case yields $a_1=v_1c_1$ and $i_1=0$. This case is consistent if and only if $c_1>4$ (which is the case with $\lambda_{BC}>0$). However, for $c_1>4$, the (DC) constraint is

violated ($v_1 - v_1 c_1 > 0$ iff $c_1 < 1$). That is, there is no profile of parameter values such that the optimum satisfies the conditions in this case.

7. Case $\lambda_{BC} = 0$ (BC does not bind), $\lambda_{DC} = 0$ (DC does not bind, conflict), and $\lambda_i > 0$ ($i_1 = 0$)

In this case $a_1 = v_1 c_1^2/4$ and $i_1 = 0$. This solution is consistent with $\lambda_{BC} = 0$ and $\lambda_{DC} = 0$ if and only if $c_1 \leq 2$. Also for $\lambda_i > 0$ we need $1 - \beta c_1/4 > 0$ (from the FOC of i_1). Thus, this holds for any triple $(\beta, c_1, v_1) \in \mathbb{R}^3_+$ such that $c_1 \leq 2$ and $\beta < 4/c_1$. In other words, the parameter set for which this region contains the solution to the incumbent's problem is

$$\mathbf{R3} = \left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | 2 \geq c_1 \, and \, \beta < 4/c_1\right\}.$$

Expected utility in this case is given by,

$$E[U_1] = v_1 \left(1 + \frac{c_1}{4}\right)$$

8. Case $\lambda_{BC} = 0$ (BC does not bind), $\lambda_{DC} > 0$ (DC binds, consolidation),and $\lambda_i > 0$ ($i_1 = 0$)

In this case the system of conditions is given by

$$\begin{aligned} \frac{1}{2}\sqrt{1} - \frac{1}{c_1} - \lambda_{DC} &= 0 \\ \frac{\beta}{2}\sqrt{1} - 1 + \lambda_{DC}\beta + \lambda_i &= 0 \\ v_1 - a_1 + \beta i_1 &= 0. \end{aligned}$$

Investment $i_1 = 0$ and therefore from (DC) we get $a_1 = v_1$. For this to be consistent with $\lambda_{DC} > 0$, we must have from the first equation that $c_1 > 2$, and to be consistent with $\lambda_i > 0$ we need $\beta < c_1/(c_1 - 1)$. Therefore the region such that this represents the optimal solution is given by

$$\mathbf{R4} = \left\{(c_1, \beta, v_1) \in \mathbb{R}^3_+ | c_1 > 2, \, \beta < c_1/(c_1 - 1)\right\}.$$

The expected utility in this case is,

$$E[U_1] = v_1 \left(2 - \frac{1}{c_1}\right).$$

■

Proof of Lemma 1 :

Case $\lambda_{BC} = 0$, $\lambda_{ND} = 0$, and $\lambda_i > 0$

FOC

$$\frac{\partial \mathbf{L}}{\partial a_0} = \frac{1}{2}\sqrt{\frac{v_0 S(m_0)}{a_0}} - \frac{1}{c_0} = 0 \tag{B.2}$$

$$\frac{\partial \mathbf{L}}{\partial i_0} = \frac{\beta}{2}\sqrt{\frac{a_0}{v_0 S(m_0)}} - 1 + \lambda_i = 0 \tag{B.3}$$

This implies:

$$i_0 = 0$$

$$a_0 = \frac{c_0^2}{4} v_0 S(m_0)$$

$$\lambda_i = 1 - \frac{\beta c_0}{4}$$

The necessary and sufficient conditions for this case to hold are

$$\lambda_{BC} = 0 \iff \frac{v_0 S(m_0)}{c_0(v_0 - m_0)} < \frac{4}{c_0^2}$$

$$\lambda_{ND} = 0 \iff c_0 < 2$$

$$\lambda_i > 0 \iff \beta < \frac{4}{c_0}$$

The first one holds for values of m_0 low enough and the second and third hold by Assumption 1.

Case $\lambda_{BC} > 0$, $\lambda_{ND} = 0$,**and** $\lambda_i > 0$

$$\frac{\partial \mathbf{L}}{\partial a_0} = \frac{1}{2}\sqrt{\frac{v_0 S(m_0)}{a_0}} - \frac{1}{c_0} - \frac{\lambda_{BC}}{c_0} = 0 \tag{B.4}$$

$$\frac{\partial \mathbf{L}}{\partial i_0} = \frac{\beta}{2}\sqrt{\frac{a_0}{v_0 S(m_0)}} - 1 - \lambda_{BC} + \lambda_i = 0 \tag{B.5}$$

$\lambda_{BC} > 0$ implies $c_0 (v_0 - m_0) = a_0$ which is always true.

From FOC, we obtain

$$\lambda_{BC} = \frac{c_0}{2}\sqrt{\frac{v_0 S(m_0)}{a_0}} - 1$$

$$\lambda_i = \frac{c_0}{2}\sqrt{\frac{v_0 S(m_0)}{a_0}} - \frac{\beta}{2}\sqrt{\frac{a_0}{v_0 S(m_0)}}$$

As before, the conditions on the parameters for this to be a solution are

$$\lambda_{BC} > 0 \iff \frac{v_0 S(m_0)}{c_0(v_0 - m_0)} > \frac{4}{c_0^2}$$

$$\lambda_{ND} = 0 \iff \frac{v_0 S(m_0)}{c_0(v_0 - m_0)} > 1$$

$$\lambda_i > 0 \iff \frac{v_0 S(m_0)}{\beta(v_0 - m_0)} > 1$$

By Assumption 1 the first case obtains when $\frac{v_0 S(m_0)}{(v_0-m_0)} < \frac{4}{c_0}$ and the second case when $\frac{v_0 S(m_0)}{(v_0-m_0)} \geq \frac{4}{c_0}$ .The proof that the other cases are infeasible (that is, the conditions on the parameters that support them do not hold under Assumption 1) is tedious but straightforward.■

Proof of Proposition 2: Recall that under Assumption 1 there are only two feasible cases:

Case $\lambda_{BC} = 0$, $\lambda_{ND} = 0$, **and** $\lambda_i > 0$

$$\lambda_{BC} = 0 \iff \frac{v_0 S(m_0)}{(v_0 - m_0)} < \frac{4}{c_0}$$
$$\lambda_{ND} = 0 \iff c_0 < 2$$
$$\lambda_i > 0 \iff \beta < \frac{4}{c_0}$$

And the expected utility

$$EU = v_0 - m_0 + \frac{c_0}{4} v_0 S(m_0)$$

Case $\lambda_{BC} > 0$, $\lambda_{ND} = 0$,**and** $\lambda_i > 0$

$$\lambda_{BC} > 0 \iff \frac{v_0 S(m_0)}{(v_0 - m_0)} > \frac{4}{c_0}$$
$$\lambda_{ND} = 0 \iff \frac{v_0 S(m_0)}{(v_0 - m_0)} > c_0$$
$$\lambda_i > 0 \iff \frac{v_0 S(m_0)}{(v_0 - m_0)} > \beta$$

And the expected utility

$$EU = \sqrt{c_0 \left(v_0 - m_0\right) v_0 S(m_0)}$$

From these two sets of the parameter space, we can compute the values of EU in period $t = 0$ for each m_0 fixing all the other parameters.

Let us call $\bar{m}$ the value of m_0 that satisfies this equation: $\frac{v_0 S(\bar{m})}{c_0(v_0-\bar{m})} = \frac{4}{c_0^2}$. $\bar{m}$ is the value of m_0 at which regimes change in period $t = 0$.

Let us call $m_{R3|R4}$ and $m_{R4|R1}$ the values of m_0 such that regimes change in period $t = 1$: $m_{R3|R4} = \frac{2-c_0}{\gamma}$ and $m_{R4|R1} = \frac{1}{\gamma}\left(\frac{\beta}{\beta-1} - c_0\right)$, $m_{R3|R4} < m_{R4|R1}$. $\bar{m}$ depends on whether it is above or below $m_{R3|R4}$ or $m_{R4|R1}$, because $\bar{m}$ is an implicit function of $S()$.

Before proving Proposition **2** we need a technical result. The following Lemma establishes the conditions of the parameters that determines the value of $\bar{m}$ relative to $m_{R3|R4}$ and $m_{R4|R1}$.

Lemma 2. *Under Assumption 1,*

i) If $0 < \gamma v_0 < \frac{8(2-c_0)}{8-3c_0}$, *then* $\bar{m} < m_{R3|R4}$.

ii) If $\frac{8(2-c_0)}{8-3c_0} < \gamma v_0 < \frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}}$, *then* $m_{R3|R4} < \bar{m} < m_{R4|R1}$

iii) If $\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}} < \gamma v_0$, *then* $m_{R4|R1} < \bar{m}$

Proof. To determine whether $\bar{m}$ lies within $[0, m_{R3|R4}]$, $[m_{R3|R4}, m_{R4|R1}]$ or $[m_{R4|R1}, \infty]$ first notice that $\frac{v_0 S(m_0)}{(v_0-m_0)}$ is increasing in m_0. Therefore, the conditions on the parameters for each of these cases to hold are:

For $\bar{m} < m_{R3|R4}$ This is the case if $\frac{v_0 S(m_{R3|R4})}{c_0(v_0-m_{R3|R4})} > \frac{4}{c_0^2} \iff v_0\gamma < \frac{8(2-c_0)}{8-3c_0}$

For $m_{R3|R4} < \bar{m} < m_{R4|R1}$ From above $m_{R3|R4} < \bar{m} \iff \frac{8(2-c_0)}{8-3c_0} < v_0\gamma$

Now we need to unveil the condition for $\frac{v_0 S(m_{R4|R1})}{c_0(v_0-m_{R4|R1})} > \frac{4}{c_0^2} \iff v_0\gamma < \frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}} = \frac{4\beta(\beta-c_0(\beta-1))}{(\beta-1)(4\beta-c_0(\beta+1))}$. Therefore, this case occurs when

$$\frac{8(2-c_0)}{8-3c_0} < v_0\gamma < \frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}}$$

For $m_{R4|R1} < \bar{m}$ It follows directly from before

$$\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}} < v_0\gamma$$

■

Lemma 2 provides us conditions on the parameters that fully describe the regimes in period 0 and period 1. The optimal m_0 therefore can be computed by simply computing the m_0 that maximizes EU in period 0 in each of the three cases in Lemma 2. For the proof of part 1 in Proposition 2 we only need to find a cutoff such that the polity stays in $R3$.We propose $\tau_L \equiv \frac{8(2-c_0)}{8-3c_0}$.

In this case, $v_0\gamma < \frac{8(2-c_0)}{8-3c_0}$ is equivalent to a regime described by $\bar{m} < m_{R3|R4}$.Note that $\frac{8(2-c_0)}{8-3c_0}$ is strictly decreasing in c_0 so its maximum value is at $c_0 = 1$. In this case $\frac{8(2-1)}{8-3\times1} = \frac{8}{5} < 2$. Let us analyze the expected utility by segments:

Segment $[0, \bar{m}]$ Expected utility in period $t = 0$ is

$EU = v_0 - m_0 + \frac{c_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4} v_0 \left(1 + \frac{c_0 + \gamma m_0}{4}\right) = v_0(1 + \frac{c_0}{4} + \left(\frac{c_0}{4}\right)^2) + m_0 \left(\frac{c_0 v_0 \gamma}{16} - 1\right)$

Since $\bar{m} < m_{R3|R4} \iff v_0\gamma < \frac{8(2-c_0)}{8-3c_0}$ then $\frac{c_0 v_0 \gamma}{16} - 1 < 0$, To see why, replace $v_0\gamma = \frac{8(2-c_0)}{8-3c_0}$ in $\frac{c_0 v_0 \gamma}{16}$ so $\frac{c_0 v_0 \gamma}{16} = \frac{c_0 \left(\frac{8(2-c_0)}{8-3c_0}\right)}{16} \leq \frac{c_0 \frac{8}{5}}{16} < 1$.So the optimal choice is $m_0 = 0$.

Segment $[\bar{m}, m_{R3|R4}]$ Expected utility in period $t = 0$ is

$EU =$
$\frac{1}{2} \frac{1}{\sqrt{c_0 v_0 (v_0 - m_0 + \frac{1}{4}(v_0 c_0 - m_0 c_0 + v_0 \gamma m_0 - m_0^2 \gamma))}} \left(c_0 v_0 \left(-1 - \frac{1}{4} c_0 + \frac{1}{4} v_0 \gamma - \frac{1}{4} 2 m_0 \gamma\right)\right)$
$\frac{dEU}{dm} < 0 \iff -1 - \frac{1}{4} c_0 + \frac{1}{4} v_0 \gamma - \frac{1}{4} 2 m_0 \gamma < 0 \iff v_0\gamma < 4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right)$.

If $4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right)$ is higher than $\frac{8(2-c_0)}{8-3c_0}$, our condition to be in this scenario $\bar{m} < m_{R3|R4} \left(\iff v_0\gamma < \frac{8(2-c_0)}{8-3c_0}\right)$ is a suficient condition for EU in this segment to be decreasing. So, it is direct to show that $4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right) > \frac{8(2-c_0)}{8-3c_0}$ because the right hand side is decreasing in c_0, so its maximum is attained at $c_0 = 1$ and it is equal to $8/5$ which is smaller than any feasible value of the expression in the left hand side. In this segment, the utility is maximized at $m_0 = \bar{m}$, which is smaller than the utility at $m_0 = 0$ from the analysis in the case above.

$EU = \sqrt{c_0 (v_0 - m_0) v_0 S(m_0)} = \sqrt{4(v_0 - \bar{m})^2} = 2 (v_0 - \bar{m})$

Segment $[m_{R3|R4}, m_{R4|R1}]$ Expected utility in period $t = 0$ is

$EU = \sqrt{c_0 (v_0 - m_0) v_0 \left(2 - \frac{1}{c_0 + \gamma m_0}\right)}$

computing the first derivative with respecto m_0
$\frac{dEU}{dm_0} = \frac{\sqrt{c_0 v_0}}{2} \frac{1}{\sqrt{(v_0 - m_0)\left(2 - \frac{1}{c_0 + \gamma m_0}\right)}} \left(\frac{v_0 \gamma}{(c_0 + \gamma m_0)^2} - 2 + \left(\frac{1}{c_0 + \gamma m_0} - \frac{m_0 \gamma}{(c_0 + \gamma m_0)^2}\right)\right)$ this quantity is less than zero, so the optimum is at $m_0 = m_{R3|R4}$, and this is smaller than the value of EU at $m_0 = 0$. To see why note that $\left(\frac{v_0 \gamma}{(c_0 + \gamma m_0)^2} - 2 + \left(\frac{1}{c_0 + \gamma m_0} - \frac{m_0 \gamma}{(c_0 + \gamma m_0)^2}\right)\right) < 0$ is equivalent to $v_0\gamma < 2(c_0 + \gamma m_0)^2 + m_0\gamma - (c_0 + \gamma m_0) \equiv W$ after re-arranging terms. The right-hand side of this expression, W, is higher than $\tau_L = \frac{8(2-c_0)}{8-3c_0}$: It is increasing in m_0 so the smallest possible value is at $m_0 = m_{R3|R4}$. At this value the expression W is $8 - c_0$, comparing

$$\frac{8(2 - c_0)}{8 - 3c_0} < 8 - c_0 \iff 0 < 48 - 24c_0 + 3c_0^2$$

which always hold in our case, because $c_0 < 2$.

Segment $[m_{R4|R1}, \infty]$ $EU = \sqrt{c_0 (v_0 - m_0) v_0 \frac{(c_0 + \gamma m_0)(1 + \beta)}{c_0 + \gamma m_0 + \beta}}$

$\frac{dEU}{dm_0} =$

$\sqrt{c_0 v_0}\frac{1}{2}\frac{1}{\sqrt{(v_0-m_0)\frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta}}}\left(\frac{-(c_0+\gamma m_0)^2(1+\beta)-(c_0+\gamma m_0)\beta(1+\beta)-\gamma m_0\beta(1+\beta)+\gamma v_0\beta(1+\beta)}{(c_0+\gamma m_0+\beta)^2}\right)$

This marginal EU is decreasing in m_0. To see why this is true, note that

$$-(c_0+\gamma m_0)^2(1+\beta)-(c_0+\gamma m_0)\beta(1+\beta)-\gamma m_0\beta(1+\beta)+\gamma v_0\beta(1+\beta)<0$$

is equivalent to

$$v_0\gamma < \frac{1}{\beta}(c_0+\gamma m_0)^2+(c_0+\gamma m_0)+\gamma m_0.$$

The right hand side of this expression is increasing in m_0, so the minimum is attained at $m_0 = m_{R4|R1}$ and it equals $\frac{\beta}{(\beta-1)^2}+2\frac{\beta}{(\beta-1)}-c_0$. The highest possible value of γv_0, $\tau_L = \frac{8(2-c_0)}{8-3c_0}$ is smaller than $\frac{8}{5}$ which, in turn, is always smaller than $\frac{\beta}{(\beta-1)^2}+2\frac{\beta}{(\beta-1)}-c_0$.

Therefore the maximum of EU in this segment is attained at $m_0 = m_{R4|R1}$.

In sum, the global maximum in this case is $m_0 = 0$. This follows from the fact that EU is continuous: $S()$ is continuous for all m_0 and EU in period $t = 0$ is also continuous at $\bar{m}$: in $t = 0$ in segment $[0, \bar{m}]$, EU evaluated at $\bar{m}$ is $2(v_0 - \bar{m})$ which is equal to the EU in segment $[\bar{m}, m_{R3|R4}]$ evaluated at $\bar{m}$. This can be shown noticing that $\frac{c_0 v_0 S(\bar{m})}{4} = v_0 - \bar{m}$, and reeplacing in EU in segment $[0, \bar{m}]$. Thus, the polity will stay at R3 in period 1.

For the proof of part 2. and 3. of Proposition 2 (the existence of cutoffs such that the polity will move away from the trap represented by region R3 in period 1), we consider the case in which $\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}} < v_0\gamma$. In this case $m_{R4|R1} < \bar{m}$ by Lemma 2. We proceed by analyzing the optimal decision of m_0 under different segments:

Segment $[0, m_{R3|R4}]$ Expected utility in period $t = 0$ is

$EU = v_0 - m_0 + \frac{c_0}{4}v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4}v_0\left(1+\frac{c_0+\gamma m_0}{4}\right) = v_0(1+\frac{c_0}{4}+\left(\frac{c_0}{4}\right)^2)+ m_0\left(\frac{c_0 v_0\gamma}{16}-1\right)$

In this case, it is not always the case that $\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}} > \frac{16}{c_0}$ under Assumption 1. In general, if γv_0 is higher than $\tau_M \equiv \max\left\{\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}}, \frac{16}{c_0}\right\}$ then the polity will move away from the trap region R3. Next, we show that it may either stay in R4 (peace) or move to R1 (peace and prosperity).

Segment $[m_{R3|R4}, m_{R4|R1}]$ Expected utility in period $t = 0$ is $EU = v_0 - m_0 + \frac{c_0}{4}v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4}v_0\left(2-\frac{1}{c_0+\gamma m_0}\right)$

The marginal utility of m_0 is: $-1+\frac{c_0 v_0}{4}\frac{\gamma}{(c_0+\gamma m_0)^2}$. The value of m_0 that maximizes the EU is $m_0 = \frac{1}{\gamma}\left(\sqrt{\frac{c_0 v_0\gamma}{4}}-c_0\right)$. For the optimum to be interior we need to compare it with the boundaries of this region $m_{R3|R4}$ and $m_{R4|R1}$. This is the case when $16/c_0 <$

$\gamma v_0 < \left(\frac{\beta}{\beta-1}\right)^2 \frac{4}{c_0}$ then the optimum is $m_0 = \frac{1}{\gamma}\left(\sqrt{\frac{c_0 v_0 \gamma}{4}} - c_0\right)$.[1] Thus, if $\tau_M < \gamma v_0 < \tau_H \equiv \max\left\{\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}}, \left(\frac{\beta}{\beta-1}\right)^2 \frac{4}{c_0}\right\}$ then the polity reaches R4. If $\tau_H < \gamma v_0$ then the polity conquers peace and prosperity in R1. Now we explore if it is possible to move to R1.

Segment $[m_{R4|R1}, \bar{m}]$ $EU = v_0 - m_0 + \frac{c_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4} v_0 \left(\frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta}\right)$ $\frac{dEU}{dm_0} = \frac{c_0 v_0}{4} \frac{(c_0+\gamma m_0)\gamma(1+\beta)+\gamma\beta(1+\beta)-(c_0+\gamma m_0)\gamma(1+\beta)}{(c_0+\gamma m_0+\beta)^2} - 1 = \frac{c_0 v_0}{4} \frac{\gamma\beta(1+\beta)}{(c_0+\gamma m_0+\beta)^2} - 1$ this implies the optimum $m_0 = \frac{1}{\gamma}\left(\sqrt{\frac{c_0 v_0 \gamma}{4}\beta(1+\beta)} - c_0 - \beta\right)$. It is streightforward to show that if $\tau_H = \max\left\{\frac{\frac{4}{c_0}\left(\frac{\beta}{\beta-1}-c_0\right)}{\frac{4}{c_0}-\frac{(1+\beta)}{\beta}}, \left(\frac{\beta}{\beta-1}\right)^2 \frac{4}{c_0}\right\} < \gamma v_0$ then the optimum is interior:

$$m_0 = \frac{1}{\gamma}\left(\sqrt{\frac{c_0 v_0 \gamma}{4}\beta(1+\beta)} - c_0 - \beta\right) > m_{R4|R1}.$$

This implies that for $\gamma v_0 > \tau_H$ the polity will be in the interior of R1 in period 1, reaching growth and prosperity.■

Proof of Proposition 3: As in the proof of Proposition 2, let us call $\bar{m}$ the value of m_0 that satisfies the equation: $\frac{v_0 S(\bar{m})}{c_0(v_0-\bar{m})} = \frac{4}{c_0^2}$. $\bar{m}$ is the value of m_0 at which regimes change in period $t = 0$.

Let us call $m_{R3|R2} = \frac{1}{\gamma}\left(\frac{4}{\beta} - c_0\right)$ and $m_{R2|R1} = \frac{1}{\gamma}(\beta - c_0)$ the values in which regimes change in period $t = 1$. The following Lemma shows the conditions on the parameters such that for any given m_0 we can fully describe the EU in period $t = 0$.

Lemma 3. *Under Assumption 1,*
i) If $0 < \gamma v_0 < \frac{16-4c_0\beta}{4\beta-(1+\beta)c_0}$ *then* $\bar{m} < m_{R3|R2}$
ii) If $\frac{16-4c_0\beta}{4\beta-(1+\beta)c_0} < \gamma v_0 < \frac{8(\beta-c_0)}{8-(1+\beta)c_0}$ *then* $m_{R3|R2} < \bar{m} < m_{R2|R1}$
iii) If $\frac{8(\beta-c_0)}{8-(1+\beta)c_0} < \gamma v_0$ *then* $m_{R2|R1} < \bar{m}$

Proof. It follows from replacing the definitions of $\bar{m}, m_{R3|R2}, m_{R2|R1}$ and following steps analogous to Lemma B.■

For the proof of part 1 in Proposition **3** we only need to find a cutoff such that the polity stays in $R3$. We propose $\sigma_L \equiv \frac{16-4c_0\beta}{4\beta-(1+\beta)c_0}$.

In this case, $v_0\gamma < \sigma_L = \frac{16-4c_0\beta}{4\beta-(1+\beta)c_0}$ is equivalent to a regime in which $\bar{m} < m_{R3|R2}$. Notice that $\frac{16-4c_0\beta}{4\beta-(1+\beta)c_0}$ is decreasing in both β and c_0. Thus, the highest feasible value of this expression is attained at $\beta = 2$ and $c_0 = 1$, $\frac{16-4\times1\times2}{4\times2-(1+2)\times1} = \frac{8}{5}$. Let us analyze the expected utility by segments:

[1] The slope in segment $[0, m_{R3|R4}]$ is positive, this plus continuity makes $\frac{1}{\gamma}\left(\sqrt{\frac{c_0 v_0 \gamma}{4}} - c_0\right)$ the optimum.

Segment $[0, \bar{m}]$ Expected utility in period $t = 0$ is

$EU = v_0 - m_0 + \frac{c_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4} v_0 \left(1 + \frac{c_0 + \gamma m_0}{4}\right) = v_0(1 + \frac{c_0}{4} + \left(\frac{c_0}{4}\right)^2) + m_0 \left(\frac{c_0 v_0 \gamma}{16} - 1\right)$

Since $\bar{m} < m_{R3|R4} \iff v_0\gamma < \sigma_L$ then $\frac{c_0 v_0 \gamma}{16} - 1 < 0$ so the optimal choice is $m_0 = 0$.[2]

Segment $[\bar{m}, m_{R3|R2}]$ Expected utility in period $t = 0$ is

$EU = \sqrt{c_0 v_0 (v_0 - m_0 + \frac{1}{4}(v_0 - m_0)(c_0 + \gamma m_0))}$

$\frac{dEU}{dm_0} = \frac{1}{2} \frac{1}{\sqrt{c_0 v_0 (v_0 - m_0 + \frac{1}{4}(v_0 c_0 - m_0 c_0 + v_0 \gamma m_0 - m_0^2 \gamma))}} \left(c_0 v_0 \left(-1 - \frac{1}{4} c_0 + \frac{1}{4} v_0 \gamma - \frac{1}{4} 2 m_0 \gamma\right)\right)$

$\frac{dEU}{dm} < 0 \iff -1 - \frac{1}{4} c_0 + \frac{1}{4} v_0 \gamma - \frac{1}{4} 2 m_0 \gamma < 0 \iff v_0 \gamma < 4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right)$.

$4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right)$ is higher than $\frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}$. Thus, EU in this segment is decreasing.

It is straightforward to show that $4 \left(\frac{1}{4} 2 m_0 \gamma + 1 + \frac{1}{4} c_0\right) > \frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}$: The lowest possible value of the left-hand side is at $m_0 = 0$, $4 + c_0$. If we compare

$$4 + c_0 > \frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}$$

is equivalent to

$$16 + 4 c_0 + c_0^2 + c_0 \beta c_0 < 16\beta + 4 c_0 \beta$$

the left-hand side of this expression can be at most 36, while the right-hand side one, at least 40 under Assumption 1. As a result, the maximum is attained at $m_0 = 0$ in the interval $[0, m_{R3|R2}]$.

Segment $[m_{R3|R2}, m_{R2|R1}]$ Expected utility in period $t = 0$ is

$EU = \sqrt{c_0 (v_0 - m_0) v_0 \left(\sqrt{\frac{c_0 + \gamma m_0}{\beta}} \frac{(1+\beta)}{2}\right)}$computing the first derivative with respect to m_0

$\frac{dEU}{dm_0} = \frac{\sqrt{c_0 v_0 \frac{(1+\beta)}{2}}}{2} \frac{1}{\sqrt{(v_0 - m_0)\left(\sqrt{\frac{c_0 + \gamma m_0}{\beta}}\right)}} \left(-\sqrt{\frac{c_0 + \gamma m_0}{\beta}} + (v_0 - m_0) \frac{1}{2} \left(\frac{c_0 + \gamma m_0}{\beta}\right)^{-1/2} \frac{\gamma}{\beta}\right)$ this quantity is less than zero, so the optimum is at $m_0 = m_{R3|R4}$, and this is smaller than the value of EU at $m_0 = 0$. To see why, $\left(-\sqrt{\frac{c_0 + \gamma m_0}{\beta}} + (v_0 - m_0) \frac{1}{2} \left(\frac{c_0 + \gamma m_0}{\beta}\right)^{-1/2} \frac{\gamma}{\beta}\right) < 0$ is equivalent to $v_0 \gamma < 2(c_0 + \gamma m_0) + m_0 \gamma$ after re-arranging terms. The right-hand side of this expression is higher than $\frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}$ for all m_0: It is increasing in m_0 so the smallest possible value is at $m_0 = 0$. At this value the minimal value of the expression expression is $2c_0$, comparing

$$\frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0} < 2 c_0$$

[2]Just replace $v_0 \gamma = \frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}$ in $\frac{c_0 v_0 \gamma}{16}$ so $\frac{c_0 v_0 \gamma}{16} = \frac{c_0 \left(\frac{16 - 4 c_0 \beta}{4\beta - (1+\beta) c_0}\right)}{16}$ which is smaller than $\frac{c_0 \left(\frac{8}{5}\right)}{16}$which in turn is less than 1.

which holds in our case since $\frac{16-4c_0\beta}{4\beta-(1+\beta)c_0} < \frac{8}{5} < 2$, because $1 \leq c_0 < 2$.

Segment $[m_{R2|R1}, \infty]$ $EU = \sqrt{c_0\left(v_0 - m_0\right) v_0 \frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta}}$

$\frac{dEU}{dm_0} = \sqrt{c_0 v_0}\frac{1}{2}\frac{1}{\sqrt{(v_0-m_0)\frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta}}}\left(\frac{-(c_0+\gamma m_0)^2(1+\beta)-(c_0+\gamma m_0)\beta(1+\beta)-\gamma m_0\beta(1+\beta)+\gamma v_0\beta(1+\beta)}{(c_0+\gamma m_0+\beta)^2}\right)$

This marginal EU is decreasing in m_0. To see why,

$$-\left(c_0+\gamma m_0\right)^2(1+\beta)-\left(c_0+\gamma m_0\right)\beta\left(1+\beta\right)-\gamma m_0\beta(1+\beta)+\gamma v_0\beta(1+\beta)<0$$

is equivalent to

$$v_0\gamma < \frac{1}{\beta}\left(c_0+\gamma m_0\right)^2+\left(c_0+\gamma m_0\right)+\gamma m_0.$$

The right hand side of this expression is increasing in m_0, so the minimum of this expression is attained at $m_0 = m_{R2|R1}$ and it equals $3\beta - c_0$. Comparing the highest possible value of γv_0, $\frac{16-4c_0\beta}{4\beta-(1+\beta)c_0}$ with the smallest value $3\beta - c_0$ the result follows.

In sum, the global maximum when $\bar{m} < m_{R3|R2}\left(\iff v_0\gamma < \frac{16-4c_0\beta}{4\beta-(1+\beta)c_0} = \sigma_L\right)$ **is** $m_0 = 0$. This follows from the fact that EU is continuous: $S()$ is continuous for all m_0 and EU in period $t = 0$ is also continuous at $\bar{m}$: in $t = 0$ in segment $[0, \bar{m}]$, EU evaluated at $\bar{m}$ is $2(v_0 - \bar{m})$ which is equal to the EU in segment $[\bar{m}, m_{R3|R4}]$ evaluated at $\bar{m}$. This can be shown noticing that $\frac{c_0 v_0 S(\bar{m})}{4} = v_0 - \bar{m}$, and reeplacing in EU in segment $[0, \bar{m}]$. Thus, the polity will stay at R3 in period 1.

For the proof of part 2. and 3. of Proposition 3 (the existence of cutoffs such that the polity will move away from the trap represented by region R3 in period 1), we consider the case in which $\frac{8(\beta-c_0)}{8-(1+\beta)c_0} < v_0\gamma$. In this case $m_{R2|R1} < \bar{m}$ by Lemma 3. We proceed by analyzing the optimal decision of m_0 under different segments:

Segment $[0, m_{R3|R2}]$ Expected utility in period $t = 0$ is

$EU = v_0 - m_0 + \frac{c_0}{4}v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4}v_0\left(1+\frac{c_0+\gamma m_0}{4}\right) = v_0(1+\frac{c_0}{4}+\left(\frac{c_0}{4}\right)^2) + m_0\left(\frac{c_0 v_0\gamma}{16}-1\right)$

In this case, $\frac{8(\beta-c_0)}{8-(1+\beta)c_0} < \frac{16}{c_0}$ under Assumption 1. Thus, there exists a threshold $\frac{16}{c_0}$ such that $\frac{8(\beta-c_0)}{8-(1+\beta)c_0} < v_0\gamma < \frac{16}{c_0}$, $m_0 = 0$ in this segment. If $\gamma v_0 > \frac{16}{c_0}$ then $m_0 = m_{R3|R2}$.

At $m_0 = 0$ the EU in this segment is$= v_0 - 0 + \frac{c_0}{4}v_0\left(1+\frac{c_0+0}{4}\right) = v_0 + \frac{c_0 v_0}{4} + \frac{c_0^2}{16}v_0$

Segment $[m_{R3|R2}, m_{R2|R1}]$ $EU = v_0 - m_0 + \frac{c_0}{4}v_0\left(\sqrt{\frac{c_0+\gamma m_0}{\beta}}\frac{(1+\beta)}{2}\right)$

The marginal utility of m_0 is: $-1 + \frac{c_0 v_0}{4}\frac{(1+\beta)}{2}\frac{1}{2}\sqrt{\frac{\beta}{c_0+\gamma m_0}}\frac{\gamma}{\beta}$. This marginal utility may be either negative, zero or positive for m_0 in the segment $[m_{R3|R2}, m_{R2|R1}]$. It is negative

in the segment $[m_{R3|R2}, m_{R2|R1}]$ if and only if the value at which the marginal utility is zero, $m_0 = \frac{\left(\frac{c_0 v_0 \gamma}{16}\frac{(1+\beta)\sqrt{\beta}}{\beta}\right)^2 - c_0}{\gamma}$, is smaller than $m_{R3|R2}$:

$$m_0 = \frac{\left(\frac{c_0 v_0 \gamma}{16}\frac{(1+\beta)\sqrt{\beta}}{\beta}\right)^2 - c_0}{\gamma} < m_{R3|R2} = \frac{(\frac{4}{\beta} - c_0)}{\gamma}$$

$$\iff v_0\gamma < \left(\frac{16}{c_0}\right)\frac{2}{1+\beta}$$

The marginal utility is zero–so the solution is interior in $R2$–if $m_0 = \frac{\left(\frac{c_0 v_0 \gamma}{16}\frac{(1+\beta)\sqrt{\beta}}{\beta}\right)^2 - c_0}{\gamma}$, is smaller than $m_{R2|R1}$ (and higher than $m_{R3|R2}$) or if

$$m_{R3|R2} < m_0 = \frac{\left(\frac{c_0 v_0 \gamma}{16}\frac{(1+\beta)\sqrt{\beta}}{\beta}\right)^2 - c_0}{\gamma} < m_{R3|R2} = \frac{(\beta - c_0)}{\gamma}$$

$$\iff \left(\frac{16}{c_0}\right)\frac{2}{1+\beta} < v_0\gamma < \left(\frac{16}{c_0}\right)\frac{\beta}{1+\beta}$$

Finally, the marginal utility is positive in this segment $[m_{R3|R2}, m_{R2|R1}]$ if $\left(\frac{16}{c_0}\right)\frac{\beta}{1+\beta} < v_0\gamma$.

Notice that a sufficient condition to have an interior maximum in $R2$ is that the EU evaluated at $m_{R2|R1}, v_0 - \frac{\beta - c_0}{\gamma} + \frac{c_0}{4}v_0\frac{(1+\beta)}{2}$, is higher than the EU evaluated at $m_0 = 0,\ v_0 + \frac{c_0 v_0}{4} + \frac{c_0^2}{16}v_0$, when $\left(\frac{16}{c_0}\right)\frac{2}{1+\beta} < v_0\gamma < \left(\frac{16}{c_0}\right)\frac{\beta}{1+\beta}$. This sufficient condition is

$$v_0 - \frac{\beta - c_0}{\gamma} + \frac{c_0}{4}v_0\frac{(1+\beta)}{2} > v_0 + \frac{c_0 v_0}{4} + \frac{c_0^2}{16}v_0$$

$$\frac{\gamma v_0 c_0}{16}(\beta - c_0 + \beta - 2) > \beta - c_0$$

$$\gamma v_0 > \left(\frac{16}{c_0}\right)\frac{\beta - c_0}{\beta - c_0 + \beta - 2}.$$

Notice that $\left(\frac{16}{c_0}\right)\frac{\beta - c_0}{\beta - c_0 + \beta - 2}$ could be either below $\left(\frac{16}{c_0}\right)\frac{2}{1+\beta}$, above $\left(\frac{16}{c_0}\right)\frac{\beta}{1+\beta}$ or in between them. Let us then define $\sigma_M \equiv \max\left\{\left(\frac{16}{c_0}\right)\frac{\beta - c_0}{\beta - c_0 + \beta - 2}, \left(\frac{16}{c_0}\right)\frac{2}{1+\beta}\right\}$ and

$$\sigma_H = \max\left\{\left(\frac{16}{c_0}\right)\frac{\beta}{1+\beta}, \left(\frac{16}{c_0}\right)\frac{\beta - c_0}{\beta - c_0 + \beta - 2}\right\}.$$

Hence, if $\sigma_M < v_0\gamma < \sigma_H$ then the polity stays in $R2$. If $\sigma_H < v_0\gamma$ then the polity moves to $R1$. The following shows the polity actually moves to the interior of $R1$ in this case.

Segment $[m_{R4|R1}, \bar{m}]$ $EU = v_0 - m_0 + \frac{c_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{c_0}{4} v_0 \left(\frac{(c_0+\gamma m_0)(1+\beta)}{c_0+\gamma m_0+\beta} \right)$

In this case, the marginal utility is $\frac{dEU}{dm_0} = \frac{c_0 v_0}{4} \frac{(c_0+\gamma m_0)\gamma(1+\beta)+\gamma\beta(1+\beta)-(c_0+\gamma m_0)\gamma(1+\beta)}{(c_0+\gamma m_0+\beta)^2} - 1 = \frac{c_0 v_0}{4} \frac{\gamma\beta(1+\beta)}{(c_0+\gamma m_0+\beta)^2} - 1$, so the interior optimum is $m_0 = \frac{1}{\gamma} \left(\sqrt{\frac{c_0 v_0 \gamma}{4} \beta(1+\beta)} - c_0 - \beta \right)$. It is stratightforward (using arguments analogous to the case in segment $[m_{R3|R2}, m_{R2|R1}]$) to show that if $\frac{16}{c_0} \frac{\beta}{(1+\beta)} < \gamma v_0$ then $m_0 = \frac{1}{\gamma} \left(\sqrt{\frac{c_0 v_0 \gamma}{4} \beta(1+\beta)} - c_0 - \beta \right) > m_{R2|R1}$. This implies that for $\gamma v_0 > \max \left\{ \frac{16}{c_0} \frac{\beta}{(1+\beta)}, \left(\frac{16}{c_0} \right) \frac{\beta - c_0}{\beta - c_0 + \beta - 2} \right\} = \sigma_H$ the polity moves to the interior of R1 in period 1, and the it conquers growth and order. ■