The Evolution of Reciprocity, Trust, and the Separation of Powers

Essays on Strategic Interactions under Incomplete Contracting

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2004

vorgelegt von
Florian Herold

Referent: Prof. Dr. Klaus M. Schmidt
Korreferent: Prof. Sven Rady, Ph.D.
Promotionsabschlussberatung: 9. Februar 2005

Acknowledgements

First and foremost I would like to thank my supervisor Klaus Schmidt for his excellent guidance, continuing encouragement and the great research environment at his chair. His superb support and advice were of invaluable help. Sven Rady and Ray Rees complete my thesis committee. I am much indebted for their support and helpful comments. The joint work with Kira Börner, co-author of chapter 2, was and continues to be a great pleasure and source of motivation.

I am very grateful to Björn Bartling, Jan Bender, Brigitte Gebhard, Georg Gebhardt, Hannah Hörisch and Susanne Kremhelmer for the excellent atmosphere at the Seminar for Economic Theory. They all helped the progress of my thesis in many ways. I am especially indebted for Georg's unfailing support and advise in many academic and practical questions, particularly, during the critical periods of starting and finishing this work. My thesis profited also from many other colleagues at the economics department at the University of Munich. In particular, I would like to thank Tobias Böhm, Florian Englmaier, Simone Kohnz, Ingrid Königbauer, Markus Reisinger, Katharina Sailer, Astrid Selders Ferdinand von Siemens, Daniel Sturm, and Hans Zenger.

Furthermore, I would like to thank the Faculty of Economics and Politics at Cambridge University, Darwin College and in particular Bob Evans, Christoph Kuzmics, Karsten Neuhoff, Andreas Park, and Andreas Pick for an inspiring academic year 2001/02 in Cambridge. Financial support from the European Commission, Marie Curie Fellowship HPMT-CT-2000-00056, and the EDGE-Programm are gratefully acknowledged.

This work profited immensely from comments and suggestions by several people. I am especially grateful for discussions with and support by Dirk Bergemann, Ted Bergstrom, Armin Falk, Ernst Fehr, Sten Nyberg, Monika Schnitzer, and Ran Spiegler.

Last but certainly not least, I thank Ulrike and my parents for their permanent backing and encouragement.

Contents

Introduction

This dissertation is composed of three self-contained essays on strategic interactions under incomplete contracting. Chapter 1 considers the evolution of reciprocal preferences in a setting where individuals live in separate groups and where there exist no higher level institutions that could enforce socially beneficial norms by offering rewards to cooperators and/or by punishing free-riders. Chapter 2 analyzes the costs and benefits of a separation of powers in an incomplete contracts framework. Chapter 3 finally shows that, even when important parts of a relationship could be arranged perfectly by a complete contract, contractual incompleteness arises endogenously if the proposal of a complete contract is perceived as a signal of distrust.

Economists typically analyze incentive problems under asymmetric information in the framework of contract theory: two or more parties can commit to a binding contract and there is an independent institution - the court - that enforces this agreement if the contract conditions only on verifiable contingencies. This framework is a powerful instrument to analyze optimal (second-best) incentives. In fact, most modern societies have institutions designed to enforce contracts that were deliberately signed by all relevant parties - at least if one party appeals to court.

The existence of a properly functioning judiciary system, however, requires a highly developed social system. A contract is nothing but ink written on paper. By itself, this does not force anybody to behave in a certain way. Nor does the sentence of a court by itself enforce the decision. A contract is worth the paper it is written on only if at least some individuals feel committed to enforce it.[1] Someone has to be

[1] A similar point is made by Mailath-Morris-Postlewaite [60] in the context of laws: Laws are nothing but cheap talk. They can only offer a focal point - selecting one equilibrium out of many and thereby changing the behavior of individuals.

willing to carry out the punishment, although this is costly. The purely self-centered agent of standard economic theory would only do so if someone offers him rewards or, alternatively, credibly threatens to punish him. But then, somebody else has to take the costs of giving this second order incentives, someone who will do so only if there is yet another person giving third order incentives and so forth. Either, there has to be an infinite chain of higher order punishments, or there must exist at least some individuals who are willing to enforce certain norms even if taking such a costly action is against their narrowly defined self-interest. The existence of individuals with social preferences is thus the basis for a developed social system and for institutions that are committed to enforce laws and contractual agreements.[2] In particular, players with reciprocal preference are committed to punish unfairly acting opponents or to reward friendly behavior. They can thus enforce norms if they consider it unfair that a norm is violated.

Recent economic experiments have shown that not all individuals always act selfishly and that some people are willing to give up monetary payoffs to reward friendly behavior and to punish hostile behavior even if interactions are anonymous and no future benefits can be expected. For a survey of these experimental findings see Fehr-Gächter [28] or Fehr-Schmidt [31]. From an evolutionary standpoint, however, these findings seem surprising. A purely self-interested agent always chooses an action that maximizes his material payoff. Thus, he should perform at least as good as any other type and finally dominate the population as a result of natural selection.

Chapter 1 offers an explanation for this puzzle. If individuals interact within separate groups, preferences for rewarding friendly behavior or preferences for punishing hostile behavior can survive evolution - even with randomly formed groups and even if individual preferences are unobservable. Intuitively, there are two evolutionary forces in our model: On the one hand, the material costs of rewarding or punishing favor the selfish relative to the reciprocal type. On the other hand, the preferences of each agent have a marginal influence on the distribution of preferences within each group. If the

[2] In return, once these institutions function properly, they may serve as a substitute for social preferences by offering a commitment device.

number of reciprocal players in a group is below a certain threshold cooperation breaks down and all members of such a group suffer a loss. In case an agent is pivotal for the enforcement of cooperation in his group he profits materially from having reciprocal preferences.

Preferences for rewarding can survive evolution as well as preferences for punishing. Yet, there exists an important structural difference between the evolution of these two types. Rewarders can always invade a selfish population of agents, but they can never drive out the selfish type completely. Punishers, in contrast, can not invade a purely selfish population, but once their fraction is above a certain threshold, they drive out all other preference types. This structural difference can be understood by the following observation: being a rewarder is particularly costly if many people cooperate - which happens frequently when there are many reciprocal players. In contrast, being a punisher is costly only if many players defect. Yet, being a punisher is cheap if most players cooperate - then no punishment is necessary. When there are many reciprocal players, cooperation is frequent.

This structural difference between the evolution of rewarders and punishers can lead to interesting co-evolutionary effects between both sides of reciprocity: punishers can not invade a purely selfish population directly, but rewarders can. The more rewarders invade, the higher the level of cooperation, and the easier it becomes for punishers to invade, too. Once they can invade successfully, they help to establish even more cooperation, and thus become even more successful. Eventually, punishers drive out all other preference types - rewarders as well as selfish agents.

In the context of social norm enforcement we can interpret cooperation as compliance with a social norm. The evolutionary process may for example be interpreted as a learning process by reinforcement - the more successful a player is with a certain behavior, the more likely he is to stick with it. Then the results of chapter 1 suggest that, firstly, rewards may be helpful for the development of a social norm. Yet, in the long run norms are more likely to be enforced by the threat of punishments. Secondly, to sustain a norm it is very crucial that there is a common agreement about a norm. Otherwise, those some people will violate it. This leads to punishments that are costly to both sides. Even worse - the norm may become unsustainable as individuals who punish norm-violators perform worse than individuals who do not enforce this norm.

This line of argument suggests that it is difficult to change a norm. More precisely, the attempt to switch from one norm to another can be very costly and may even fail - unless the change happens well coordinated and simultaneously for the entire population. This suggests a rationale for the institution of government: if some parameters of the economic environment change stochastically and if thereby the socially optimal norm changes over time - then proclamations of the government may serve as a focal point and coordinate the change from one norm to another.

This choice of a focal point gives decision-making power to the government which it can misuse led by its private interests. The population would like to have a control mechanism that enables it to change the focal point in case the government's decisions deviate too much from the social interest. Yet, again, an uncoordinated attempt to change the focal point would cause large social costs. Thus, clear rules are needed to control the government - a constitution is required that provides simple instructions when and how to change the focal point from the incumbent to another predetermined person or body. Such a constitutional rule can assign the focal point - the right to define a norm - for different functional tasks to separate entities.[3] The separation of powers is the most prominent example for such a rule - and the focus of chapter 2.

Chapter 2 analyzes the costs and benefits of a separation of powers in an incomplete contracts approach.[4] The assumption that only certain incomplete contracts can be written is present at two levels: Firstly, citizens only have the choice between constitutions: they choose either a constitution of a single ruler or they choose to separate the legislative from the executive power. Secondly, legislature can condition laws only on some exogenously given categories that classify public projects in general terms. However, the legislature cannot condition laws on the specific characteristics of every potentially upcoming project. Those can only be taken into account by the executive who makes case-to-case decisions.

[3] Similarly, such a constitutional rule could state "hold elections every 4 years and if the incumbent looses the majority of votes change the focal point to the candidate of the opposition".

[4] Chapter 2 is joint work with Kira Börner.

Thus, the legislature provides a decision-making framework by means of writing laws. The executive is left with the residual decision making rights. The legislature constrains the executive by law for some types of decisions and empowers it to decide in other policy areas. Chapter 2 considers this functional division of tasks as the key difference between legislature and executive. The executive has private interests which distort the policy choice away from the social optimum. Laws written by an independent legislation have the advantage of curbing this abuse of executive power. The legislature, however, can also have private interests which distort the law with respect to the socially optimal law. Yet, a law must be written in general terms and affects always a large number of potential policy-decisions. The legislature can only account for its expected private interests averaged over an entire category of potential projects. Extreme private interests of the legislature in some projects tend to cancel out with other projects of the same category, where the private bias is in the opposite direction. Whether a separation of powers or a single ruler is the better constitution depends on the relative intensity of private interests in the executive and in the legislature. Chapter 2 argues that a separation of powers is more important on the national or supra-national level, where decisions can have far-reaching consequences and private interests differ strongly. On a regional or local level, where governments have less scope for extreme decisions and the special characteristics of each project are particularly important, a separation of powers is less attractive and other control mechanisms tend to be more effective.

Chapter 3 now takes a social system with well functioning institutions and a court that enforces contractual agreements for granted. Individuals can write binding contracts, conditional on verifiable contingencies. The focus of chapter 3 is to demonstrate how the fear to signal distrust can lead a principal to refrain from writing a complete contract. Thus, asymmetric information about how much one party is trusting the other one can endogenously lead to contractual incompleteness for strategic reasons.

According to standard results in contract theory an optimal incentive contract should be conditional on all verifiable information containing statistical information about an agent's action or type.[5] Most real world contracts, however, condition only on few contingencies and often no explicit contract is signed at all. Chapter 3 offers an explanation for this stylized fact. If trust is an important element of a relationship, the fear to signal distrust to the other party can endogenously lead to contractual incompleteness. Designing a sophisticated complete contract with fines, punishments, and other explicit incentives signals distrust to the partner. A trustworthy partner would choose the desired action anyway. Insisting on explicit contractual incentives means therefore that the partner's trustworthiness is called into question. An atmosphere of trust, however, is of crucial importance for the functioning of most economic and noneconomic relationships. A principal may therefore prefer to leave a contract incomplete rather than to signal her distrust by proposing a complete contract.

More precisely, in our model an agent is one of two possible types, trustworthy or untrustworthy. The trustworthy agent is intrinsically motivated to work for a joint project with the principal - or to comply with a socially beneficial norm. The untrustworthy type is purely self-interested. In other words, the trustworthy agent works hard for a project - with or without a contractual enforcement. The untrustworthy agent, however, shirks - unless a contract forces him to work hard. The principal can have different beliefs about the agent's type: the stronger her belief that the agent is trustworthy the stronger the principal's *trust* in the agent. The relationship consists of two parts - in the first part high effort can be enforced by a contract. The second part is not contractible. Due to this second non-contractible part of the relationship it is important for the principal that the agent believes to be trusted.

If the belief of the principal were common knowledge, the principal - whatever her belief - would prefer to enforce high effort in the contractible part in order to avoid the risk that an agent shirks. However, a trusting principal expects the costs of forbearing from such a prescription to be lower. According to her belief the agent is likely to work hard anyway. Thus - under asymmetric information about the principal's belief, a trusting principal can separate herself from a distrusting one by proposing a less complete contract.

[5] See e.g. Holmstöm [47] or Laffont and Tirole [55].

The beginning of this introduction argued that the existence of some individuals with social preferences is the basis for a developed social and political system with institutions that enforce contracts signed between parties. Social preferences are thus the foundation for an environment in which binding contracts can be written. The very existence of heterogenous social preferences, however, can endogenously cause parties to refrain from the opportunity to write a complete contract - caused by the fear that proposing such a contractual agreement would signal distrust to the other party.

Chapter 1

Carrot or Stick?

The Evolution of Reciprocal Preferences in a Haystack Model

1.1 Introduction

This chapter addresses three questions concerning the evolution and co-evolution of the two characteristics of reciprocity - the willingness to reward friendly behavior and the willingness to punish hostile behavior. 1) How can preferences for rewarding, and preferences for punishing, survive the evolutionary competition with purely self-interested preferences? 2) What structural differences distinguish the evolution of the willingness to reward from the evolution of the willingness to punish? 3) How is the evolution of one side of reciprocity influenced by the evolution of the other side?

Self-interested preferences are a standard assumption in economic theory. From an evolutionary standpoint[1] this assumption seems justified due to the following argument: a rational self-interested individual can always mimic other behavior if by doing so he can maximize his expected material payoff. But, in the event of his action being

[1] The process of evolution may be interpreted in terms of both biological and cultural evolution or even as a process of learning. Under the weak assumptions described below, our results hold independently of which interpretation is used. Therefore, we postpone the discussion of the relevance of each interpretation to Section 1.3.

unobservable or punishment being impossible, he behaves selfishly and receives a higher material payoff. Apparently, self-interested individuals should always outperform other types and social preferences should vanish as a result of natural selection.

Several experimental studies however, offer substantial evidence that at least some people are not exclusively driven by self-interest. A significant number are willing to reward friendly and/or to punish hostile behavior of an opponent even if this is costly and does not maximize their own material payoffs[2]. In addition, a recent experimental study by Andreoni et al [4] finds interesting interactions between the possibilities of rewarding and punishing: when subjects have the option to reward as well as the option to punish the demand for rewards decreases significantly compared to a treatment where there is no option to punish. The demand for punishments in response to very bad offers[3] is however significantly higher compared to a treatment without the option to reward. How can we understand this pattern of behavior from an evolutionary standpoint, and how could reciprocal preferences survive natural selection in the first place?

This chapter shows that preferences for both rewarding and punishing can survive evolutionary competition with purely self-interested preferences if players interact within separate groups[4] and if they can condition their strategy on the distribution of preferences within their own group. This holds even if individual preferences are unobservable, groups are formed randomly and players interact anonymously in random pairings.

However, there are important structural differences between the evolution of preferences for rewarding and the evolution of preferences for punishing. Rewarders can successfully invade a population of self-interested players, but they cannot drive them out completely. Preferences for rewarding survive only in coexistence with self-interested preferences. Preferences for punishing on the other hand either drive out self-interested

[2] For a survey of the experimental literature see Fehr and Gächter [28] or Fehr and Schmidt[31].

[3] For medium range offers there is no significant change in the demand for punishments.

[4] The expression “Haystack Model” for settings where players interact only within randomly formed groups which are reshuffled after reproduction goes back to Maynard Smith’s [63] example of mice living and replicating over the summer within separate haystacks. At harvest time when the haystacks are cleared mice scramble out into the meadow and mix up completely before colonizing new haystacks in the next summer.

preferences or they die out themselves. The option to punish hostile behavior results either in a “culture of punishment” - where all players are willing to punish hostile behavior - or in a “culture of laissez faire” - where nobody is willing to incur the costs of punishing.

If there is both an option to reward friendly behavior and an option to punish hostile actions, further interesting effects arise from the co-evolution of the two aspects of reciprocity. Rewarders enhance the evolutionary success of preferences for punishing, but punishers tend to crowd out preferences for rewarding. In fact, rewarders may serve as a catalyst for the evolution of punishers. Rewarders can invade a population of self-interested types. Their existence can enable punishers to invade successfully, and finally to crowd out, both self-interested and rewarding types. Hence the option to reward friendly actions can crucially influence the equilibrium outcome even if in equilibrium nobody takes this option.

Our results are driven by the marginal effect a player has on the distribution of preferences within his group. This marginal effect is advantageous for a reciprocator and can outweigh the costs of rewarding or punishing. To see this, consider pairwise interactions of the following structure: a first moving player *(player 1)* may either cooperate or defect. Cooperation is costly for player 1 but profitable for a second moving player *(player 2)*. Player 2 observes this action and can then reward and/or punish player 1 (both are costly) or remain inactive. Player 1 cooperates only if he expects player 2 to reciprocate with a sufficiently high probability. Since individual preferences are unobservable, player 1 estimates the probability of meeting a certain type by the fraction of this type in his group. Hence players 1 cooperate if the number of reciprocal players in their group is above a certain threshold, otherwise they defect. Having reciprocal preferences leads to a material advantage for player 2 when he is pivotal in his group, i.e. his type is decisive for whether the number of reciprocal preferences in his group is just above or just below the threshold for cooperation.

Under what circumstances does this material advantage outweigh the losses incurred for rewarding or punishing? The intuition for our main result is derived from the following observation: the hope for reward as well as the fear of punishment can induce players 1 to cooperate. But when most players 1 cooperate it is relatively expensive for player 2 to reward cooperation whereas the willingness to punish is almost for free. On the other hand, when most players 1 defect, the willingness to reward is almost for free, whereas it is expensive to punish defection. A higher fraction of rewarders or punishers leads to a higher fraction of groups in which cooperation occurs. Therefore, rewarders are relatively successful when most players 2 are self-interested, whereas punishers become more successful the more players 2 have reciprocal preferences.

The existing evolutionary literature has paid little attention to the structural differences between the evolution of a willingness to reward and a willingness to punish and their mutual interaction. However, the question about how reciprocity or social preferences can survive evolution in sporadic interactions has been tackled by several authors from biology, psychology, economics and other social sciences.[5] The existing explanations[6] relate to three basic themes[7]: *commitment, assortation,* and *parochialism.*

Commitment: If preferences are observable, reciprocal preferences may serve as an advantageous commitment device. A reciprocal player is credibly committed to rewarding friendly or punishing unfriendly behavior. Therefore, he may induce friendly behavior of a first-moving player. This may enhance his evolutionary success. The results of Güth and Yaari [41], Güth [40], Bester and Güth [11] and partly Sethi [82], Höffler [46], and Guttman [42] are based on this argument.

Assortation: Efficiency enhancing behavior becomes evolutionarily more successful,

[5] On the separate issue of whether evolution selects the Pareto efficient equilibrium in coordination games see Robson [75], Kandori-Mailath-Rob [50], Robson-Vega-Redondo [76], and Kuzmics [52].

[6] For surveys of this literature see Sober and Wilson [86] and more recently Bergstrom [10] or Sethi and Somanathan [85].

[7] Eshel, Samuelson and Shaked [25] find a different explanation how altruism may survive based on the assumptions that people learn by imitation and interact and learn only locally on a 1 dimensional circle.
Huck and Oechssler [49] exploit a further mechanism to explain how preferences for punishing unfair behavior might survive evolution. They look at ultimatum games in which costs of punishing unfair behavior are very small compared to the punishment and the inverse group-size. In the role of a proposer punishers have the relative advantage over materialists in their group, that they are slightly less likely to be matched with a punisher. Therefore, unfair offers of materialists are more likely to be rejected. In the role of responders punishers have the disadvantage of incurring the costs of punishing. But if these costs are sufficiently small this disadvantage is more than compensated by the relative advantage when being in the role of a proposer.

if players are not matched randomly, but interact with higher probability with players of their own type. In particular, most of the literature on group selection focuses on this idea to explain the evolutionary survival of social preferences. Price [73] first offered a mathematical description. Bergstrom [9] investigated the relation between assortative matching and the evolution of cooperation. Notice that initially random groups may become assortative over time by the evolution of preferences inside groups if no reshuffling of groups occurs[8] (compare e.g. Cooper and Wallace [18]).

Parochialism: Types that act to enhance efficiency if they are in a group of mainly their own type and act to reduce efficiency if they are in a group of mainly self-interested players may survive in evolution even if the matching is non-assortative. Sethi and Somanathan [84] showed that conditional altruists who behave in a friendly way towards other altruists but spitefully towards materialists may be more successful than pure materialists. Similarly Gintis [36] looked at conditional punishers who punish defectors only if there are enough other punishers in their group. They also survive evolution[9].

The explanation of this chapter for the survival of reciprocal preferences is related to the idea of *commitment*. But we go beyond the existing literature by relaxing the assumption that individual preferences are observable[10]. Players observe only the overall distribution of preferences within their own group.[11] The marginal effect of a player on the distribution of preferences in his group drives our results[12]. Even more

[8] A different endogenous justification of assortative matching arises if preferences are partly observable, see e.g. Frank [32].

[9] Notice that punishers may enhance efficiency, because they can induce cooperation of a first-moving player.

[10] Bowles and Gintis [13] and Friedman and Singh [34] consider also the case when individual preferences are unobservable for the evolution of types who punish non-cooperative behavior. However, Friedman and Singh implicitly need the assumption that second-order punishment (i.e. the punishment of non-punishers) is costless. Bowles and Gintis consider a model where punishment takes the from of ostracism. Their model is too complex for an analytical solution but in simulations they also find that punishers can survive.

[11] Ely and Yilankaya [21] show in a model without group structure and with unobservable preferences over the outcome of the game that the distribution of aggregate play must be a Nash equilibrium. See also Ok and Vega-Redondo [68] for a justification of the evolution of self-interested preferences when preferences are unobservable. Dekel et al [20] analyze how the evolution of preferences changes with the degree of observability of individual preferences.

[12] A similar effect plays a role in the model by Höffler [46]. He considers a learning process of bounded rational workers in a stylized principal agent model. In equilibrium some agents play fair

importantly, our setting allows the analysis of both sides of reciprocity in a unified framework. We find crucial structural differences between the evolution of preferences for rewarding and the evolution of preferences for punishing. Finally, our framework enables us to demonstrate that the co-evolution of both sides of reciprocity influences the results decisively. Rewarders enhance the evolutionary success of punishers whereas punishers crowd out rewarders.

The remainder of the chapter is organized as follows: Section 1.2 presents the model and analyzes the three cases that arise naturally: 1) Player 2 might have only the costly option of rewarding cooperation. 2) Player 2 might have only the costly option of punishing defection. 3) Player 2 might have both the options - rewarding cooperation and punishing defection. Section 1.3 discusses our results and Section 1.4 concludes.

1.2 The Model

We use the indirect evolutionary approach[13] to describe the evolution of preferences: individuals may have different preferences. We only impose the restriction that preferences can be described by subjective utilities for each possible outcome. The subjective utility an individual assigns to an outcome may not coincide with the material payoff he receives. Individuals choose their strategies according to their own preferences and their knowledge about preferences of their opponents, i.e., they play perfect Bayesian equilibria. They receive material payoffs according to strategies played. A type who receives higher material payoffs has more offspring (or imitators) and his fraction grows. Subjective utilities of an individual are only important for determining his actions. The evolutionary success is only influenced by the resulting material payoffs.

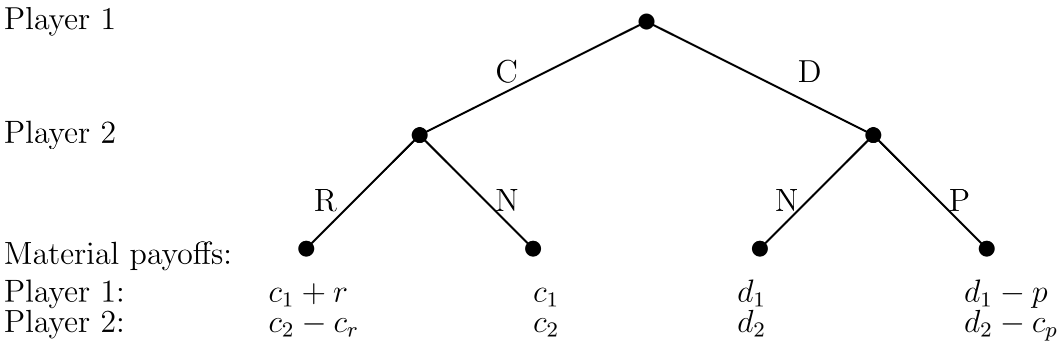
We consider pairwise sequential interactions of the following structure: *Player 1* moves first. He can either cooperate (C) or defect (D). Cooperation leads to a material gain for *player 2* ($c_2 > d_2$) but is costly for player 1 ($d_1 > c_1$). Player 2 observes the

and other don't - like the coexistence result in case 1 of our model.

[13] Compare Güth and Yaari [41].

action of player 1 and then chooses his reaction. Three cases are analyzed. In case 1, player 2 can reward cooperation of player 1 (by the amount of r) but this is costly (costs c_r) . In case 2, player 2 can punish defection of player 1 (by the amount p) which is also costly (costs c_p). In case 3, player 2 can do both and either reward or punish player 1. The interaction of case 3 is illustrated in figure 1.1. Case 1 and case 2 are obtained from this figure by removing the option to punish or the option to reward respectively.

Figure 1.1: Interaction


Player 1
C
D
Player 2
R
N
N
P
Material payoffs:
Player 1: $c_1 + r$ c_1 d_1 $d_1 - p$
Player 2: $c_2 - c_r$ c_2 d_2 $d_2 - c_p$

with $c_1 + r > d_1 > c_1 > d_1 - p$ and $c_2 > c_2 - c_r > d_2 > d_2 - c_p$.

If all players maximize only their own material payoff (and are known to do so) all three games are solved easily by backward induction. Player 2 never incurs any costs in the last stage. This is anticipated by player 1. Therefore, player 1 defects in the first stage. This outcome also tends to result in an evolutionary setting, if individual preferences are unobservable and if all players are matched randomly within the total population (i.e, no group structure is imposed)[14]. The reason is simple: someone who chooses a strategy which maximizes his material payoffs earns more than someone who doesn't. However, results change when the total population is divided up into separate groups and players interact only within their own group.

[14] Nöldeke and Samuelson [67] give an example to illustrate that in general the subgame-perfect Nash equilibrium is not the only evolutionary stable equilibrium. Similarly, the results by Sethi and Somanthan [83] rely on the fact that non-credible threats can survive in certain evolutionary settings. However, Hart [44] and Kuzmics [51] show that the subgame-perfect Nash equilibrium results if certain limits are taken in a suitable way. See also Ok and Vega-Redondo [68] for a justification of the evolution of self-interested preferences when preferences are unobservable.

Whether the fraction of players of a certain preferences type grows or shrinks depends on their individual material payoffs. With the law of large numbers in mind, we concentrate our analysis on deterministic approximations to the evolutionary dynamics[15]. The results of this chapter hold for any payoff-monotonic dynamics. By payoff monotonicity we mean that the fraction of a preference-type with higher (equal) average material payoff grows faster than (as fast as) the fraction of a preference-type with lower (equal) average material payoff. Furthermore, it is convenient to assume a continuous dynamics.

Assumption 1 *The evolutionary dynamics can be described by regular payoff monotonic growth rates.*[16]

Furthermore, we say that a population state forms a stable equilibrium if it is an Asymptotically Stable State - a standard concept in evolutionary game theory.[17]

What preference types are relevant for our analysis? In general, each player assigns a subjective von Neumann-Morgenstern-utility to each outcome. These subjective utilities depend on the actual position of the player and may differ completely from his material payoffs.[18] We are mainly interested in the evolution of preferences for the position of player 2 - reciprocal behavior is only possible in that position. However, the evolutionary success of preferences for position 2 depends on the behavior of players 1. In proposition 9 in the appendix we show that in position 1 no type of preferences can do better than self-interested ones and that in any stable equilibrium all players 1 must act consistently with payoff maximization. This result justifies simplifying the analysis by the slightly stronger

Assumption 2 *In position 1 all players maximize their expected material payoffs.*

[15] This is very common in evolutionary game theory even if not entirely innocuous. For some caveats for this approach see Boylan [14]. For a thorough discussion of a deterministic dynamics as a limit of a stochastic dynamics see Benaim and Weibull [7].

[16] The formal definition of a regular payoff monotonic growth rate can be found in Weibull [95] or in the appendix of this chapter.

[17] See Weibull [95] or the appendix of this chapter for the precise definition.

[18] In the most general case 3 there exist 4 possible outcomes. A preference type is therefore characterized by a tuple of 8 subjective utilities (modulo a linear transformation). The first 4 subjective utilities describe a players preferences if he happens to play in position of player 1, the remaining 4 subjective utilities describe his preferences in position of player 2.

But, when players happen to play in position 2, four classes of preferences may be relevant: we call someone a *"rewarder"* if he is willing to incur costs to reward a friendly action, a *"punisher"* if he is willing to incur costs to punish a hostile action,[19] a *"reciprocator"* if he is willing to do both, and *"self-interested"* if he is willing to do neither. We say someone has *"social preferences"* if he is either a rewarder, a punisher or a reciprocator. In case 1 and case 2 only two of these types matter.

An infinite population is divided up randomly into separate groups of $(2N)$ players. N players are drawn randomly to play in position 2, the remaining N players play in position 1.[20] By "randomly" we mean that a player's type does not influence the probabilities of the types of his group-members, i.e.:[21]

Assumption 3 *The probability that k_+ of the N players 2 in a group are rewarders, k_- are punishers, k_{rc} are reciprocal and k_s are self-interested (with $k_+ + k_- + k_{rc} + k_s = N$) is multinomial distributed:*[22]

$$M_{N,\gamma_-,\gamma_+,\gamma_{rc},\gamma_s}(k_+, k_-, k_{rc}, k_s) = \frac{N!}{k_+!k_-!k_{rc}!k_s!}\gamma_+^{k_+}\gamma_-^{k_-}\gamma_{rc}^{k_{rc}}\gamma_s^{k_s} \tag{1.1}$$

where γ_i is the fraction of the i-th type in the total population (hence $\gamma_+ + \gamma_- + \gamma_{rc} + \gamma_s = 1$).

In case 1 and case 2 only two types of preferences are relevant. Then, the multinomial distribution reduces to the binomial distribution:

$$B_{N,\gamma}(k) = \frac{N!}{k!(N-k)!}\gamma^k(1-\gamma)^{N-k}. \tag{1.2}$$

[19] The literature also calls preferences for rewarding "positively-reciprocal" and preferences for punishing "negatively-reciprocal".

[20] We might also reshuffle the positions of all players for each interaction. The main results would not change.

[21] Assumption 3 of regularly reshuffling may seem strong for real world applications. But precisely this assumption allows us to abstract from assortative group selection effects and to isolate the effects we are interested in. From a theoretical standpoint this strengthens our results: preferences for punishing or rewarding can survive evolution even without effects of assortative matching.

[22] An even more natural choice would be the multi-hyper-geometrical distribution (drawing without replacement). For simplicity we approximate it by the multinomial distribution. The qualitative results are not affected and the approximation is good for a large total population.

Individuals interact in random pairings within their group. Individuals do not know the type of their respective counterpart, but we assume them to know the frequency of each preference-type in their group:

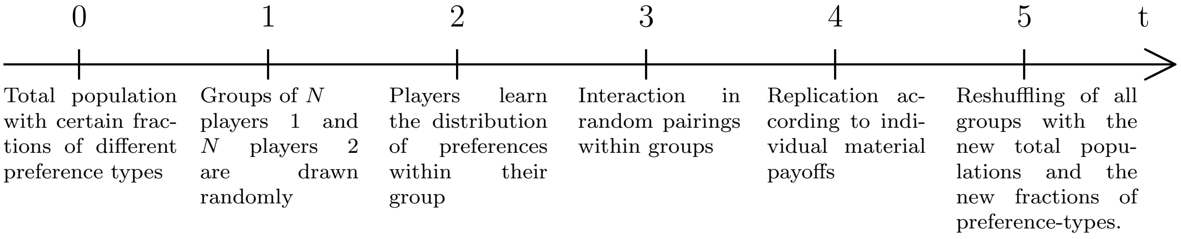
Assumption 4 *Individuals know the fractions of the different types within their own group (but they don't know the type of their randomly matched opponent).*

Most authors in this branch of literature make the stronger assumption that individual preferences are observable. We relax this assumption considerably by assuming only that the distribution of preferences within a group is observable. It may be impossible to guess your counterparts' individual preferences in a sporadic interaction. But most people will have a good estimate how likely they are to encounter one or the other type in their environment.[23]

In order to abstract from repeated games effects, we assume that individuals play anonymously and finitely often. Hence, player 2 need not fear any consequences in a later stage whatever action he takes.

After a finite number of interactions preferences are replicated according to received material payoffs and all groups are completely reshuffled. A new cycle starts with the new fractions γ_i of the different preference types in the total population. Timing of events in our model is illustrated graphically in Figure 1.2.

Figure 1.2: Timing of events

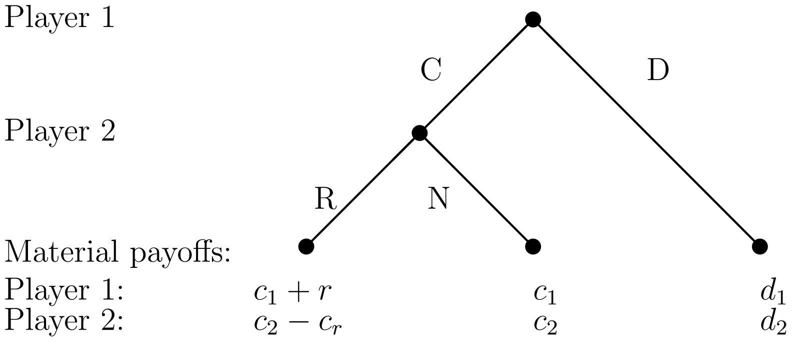

0
1
2
3
4
5
t
Total population with certain fractions of different preference types
Groups of N players 1 and N players 2 are drawn randomly
Players learn the distribution of preferences within their group
Interaction in random pairings within groups
Replication according to individual material payoffs
Reshuffling of all groups with the new total populations and the new fractions of preference-types.

[23] However, Assumption 4 is not entirely innocuous. If the fraction of self-interested types is not common knowledge - as assumed here - but has to be learned from other people's behavior in previous periods, then even a self-interested player 2 might have an incentive to reward cooperation because he anticipates his marginal influence on the learning of players 1. Hence self-interested players 2 might try to build a reputation - not individually, but of their group. Modelling the consequences of such a learning process seems an interesting but complicated task and is therefore left to future research.

1.2.1 Case 1: Costly Rewarding

Case 1 concentrates on the possibility that player 2 can reward friendly behavior of player 1. First, player 1 decides whether to cooperate or defect. Player 2 observes this action. In case player 1 cooperates, player 2 may either incur the costs to reward player 1 or refuse to do so.[24]. This game is also known as "trust game" and is illustrated in figure 1.3

Figure 1.3: Interaction in case 1



with $c_1 + r > d_1 > c_1$ and $c_2 > c_2 - c_r > d_2$.

Player 1 maximizes his expected material payoff. Player 2 either has preferences for rewarding and rewards cooperation or he has self-interested preferences and does not reward cooperation of player 1. The evolutionary process determines the fractions of each type in equilibrium.

We consider a group where k of the N players 2 have preferences for rewarding cooperation. Player 1 will base his decision whether to cooperate or defect on his expected material payoff. Player 1 does not know the type of his opponent, but he does know the fraction $\frac{k}{N}$ of players 2 in his group who would reward cooperation. Hence player 1 expects an average material payoff of $(c_1 + \frac{k}{N}r)$ for cooperation. If player 1 defects, he receives surely a payoff of d_1. Therefore, player 1 will cooperate if

[24] We could give player 2 an additional option to reward player 1 after defection. But it is straightforward to show that preferences for rewarding defection cannot be part of any stable equilibrium. Therefore, we ignore this possibility.

$c_1 + \frac{k}{N} r > d_1$ or equivalently if $k > N \frac{d_1 - c_1}{r}$.[25] Hence, cooperation occurs in a group only if the number of rewarding players 2 is above this threshold. We denote this threshold by k^*.

Definition 1 *k^* is the highest number of rewarding players 2 in a group which is still not sufficient to induce player 1 to cooperate. In other words*
$k \leq k^* \Rightarrow$ player 1 defects
$k > k^* \Rightarrow$ player 1 cooperates.

Calculation of k^* is straightforward:

$$k^* = \left\lfloor N \frac{d_1 - c_1}{r} \right\rfloor, \tag{1.3}$$

where $\lfloor x \rfloor$ denotes the largest natural number smaller or equal to the real number x. k^* is an integer with $0 \leq k^* \leq N - 1$.

In groups with k^* or fewer rewarding players 2 no cooperation occurs. Players 1 defect and players 2 receive a material payoff of d_2 - independently of their types. In groups with more than k^* rewarding players 2 players 1 cooperate. A rewarding player 2 receives a material payoff of $(c_2 - c_r)$. A self-interested player 2 exploits cooperation of player 1 and receives a material payoff of c_2. These payoffs are summarized in table 1.1.

Table 1.1: Material payoffs of player 2

	Payoffs in groups with $k \leq k^*$	Payoffs in groups with $k > k^*$
Rewarder	d_2	$c_2 - c_r$
Self-interested	d_2	c_2

For a player 2, the probability that exactly k of the other $(N-1)$ players 2 in his group have preferences for rewarding is $B_{N-1,\gamma}(k) = \frac{(N-1)!}{k!(N-1-k)!} \gamma^k (1-\gamma)^{N-1-k}$, where γ is the fraction of rewarders in the total population. If this player has preferences for rewarding the total number of rewarders in his group is $(k+1)$, otherwise it remains

[25] For notational simplicity we define the tie breaking rule that player 1 defects if his expected payoff for defecting equals that for cooperating.

k. Hence, a self-interested player 2 receives an expected material payoff of[26]

$$\bar{u}_s(\gamma) = d_2 \sum_{k=0}^{k^*} B_{N-1,\gamma}(k) + c_2 \sum_{k=k^*+1}^{N-1} B_{N-1,\gamma}(k) \tag{1.4}$$

and a rewarding player 2 receives an expected material payoff of

$$\bar{u}_+(\gamma) = d_2 \sum_{k=0}^{k^*-1} B_{N-1,\gamma}(k) + (c_2 - c_r) \sum_{k=k^*}^{N-1} B_{N-1,\gamma}(k). \tag{1.5}$$

Due to the assumption 1 of payoff monotonicity, the fraction of rewarding players grows (falls) if they receive a higher (lower) average payoff than the self-interested type. Hence, we can see from the sign of the difference

$$\bar{u}_+(\gamma) - \bar{u}_s(\gamma) = (c_2 - c_r - d_2) B_{N-1,\gamma}(k^*) - c_r \sum_{k=k^*+1}^{N-1} B_{N-1,\gamma}(k) \tag{1.6}$$

$$= \quad (c_2 - c_r - d_2)\binom{N-1}{k^*}\gamma^{k^*}(1-\gamma)^{N-1-k^*} - c_r \sum_{k=k^*+1}^{N-1} \binom{N-1}{k}\gamma^k(1-\gamma)^{N-1-k}$$

when the fraction of rewarding players increases, decreases or remains stable. First, we consider the case $c_2 - d_2 - c_r \leq 0$, i.e. gains of cooperation for player 2 are smaller than costs of rewarding. Then, all terms on the right hand side of equation 1.6 are negative (or zero) and a self-interested player 2 earns always more than a rewarding player 2.

Proposition 1 *If $c_2 - d_2 - c_r \leq 0$, i.e. the cost for rewarding exceed player 2's gains from player 1's cooperation, then only an entirely self-interested population is stable*[27].

But mainly we are interested in the case $c_2 - d_2 - c_r > 0$, i.e. gains from cooperation for player 2 exceed his costs of rewarding. Then, there is a chance for the survival of

[26] A different way to calculate this is to multiply the payoff of a rewarding (self-interested) player 2 in a group of k rewarding players 2, multiply it by k $(N-k)$ and weight it by the probability that a group has k rewarding players 2 (i.e. the binomial coefficient). If we sum this up over all $0 \leq k \leq N$ and divide it by the total number of players 2 of that type we get the average payoff. Of course, the results remain unchanged.

[27] A population is called stable if it is an asymptotical stable state of the dynamics. For details see e.g. Weibull [95] or the appendix of the working paper version.

preferences for rewarding . In fact, for $k^* < N - 1$ preferences for rewarding and self-interested preferences coexist in any stable equilibrium[28]:

Proposition 2 (Coexistence) *Let $c_2 - d_2 - c_r > 0$ and $k^* < N - 1$. Then, the monomorphic population states (i.e. states with a fraction $\gamma = 0$ or $\gamma = 1$ of rewarders) are unstable for any payoff monotonic dynamics. Preferences for rewarding can invade a self-interested population and self-interested preferences can invade a population of rewarders.*

Remark 1 *If $c_2 - d_2 - c_r > 0$ and $k^* = N - 1$ then only a monomorphic population of preferences for rewarding forms a stable equilibrium.*

All proofs are relegated to the appendix.

For an intuitive understanding of Proposition 2 first consider a population consisting almost entirely of rewarders. Then, almost all groups consist almost entirely of rewarding players 2. Therefore, players 1 cooperate in almost all groups. A rewarding player 2 receives a payoff of $(c_2 - c_r)$ only, whereas a self-interested player 2 saves the costs of rewarding and earns the higher payoff of c_2. Therefore, the fraction of self-interested players grows.

The intuition for why preferences for rewarding can invade a self-interested population is slightly more involved. Consider a population consisting almost entirely of self-interested players. Then the vast majority of groups contain too few rewarding players 2 to induce cooperation of players 1. In these groups self-interested and rewarding players 2 receive the same payoff d_2. But in a small number of groups the fraction of rewarding players 2 is above the threshold k^* and players 1 are willing to make the advanced concession of cooperation. Every player in these groups receives a higher payoff than most players in groups without cooperation. But the fraction of rewarding players 2 in these groups is at least $\frac{k^*}{N}$ and therefore far above the fraction of

[28] We could generalize this result slightly: Take any trait that (a) when the fraction of a group possessing the trait is less than $1 < k^* < N$, those with and without the trait do equally well; (b) when the fraction is above $k*$, all agents in the groups do better, but those with the trait do worse than those without; (c) agents are randomly assigned to groups. Then there is a positive fraction of agents with the trait in equilibrium. I would like to thank Herb Gintis and Bob Evans for pointing this out.

rewarders in the total population (which is close to zero). Therefore, rewarding players profit relatively more from these successful groups and can invade a self-interested population.

If $k^* = N - 1$ the result changes for the following reason: in this case self-interested preferences cannot invade a population of rewarders. Even if a self-interested player 2 is the only invader in his group he destroys cooperative behavior of players 1. Hence, if $k^* = N - 1$ rewarding players 2 always do at least as well as self-interested players 2.

According to proposition 2 only mixed populations are candidates for stable preference distributions. In fact, there exists a unique stable equilibrium.

Theorem 1 (Unique mixed equilibrium) *Let $c_2 - d_2 - c_r > 0$ and $k^* < N - 1$. Then there exists a unique stable equilibrium. Self-interested preferences and preferences for rewarding coexist in this equilibrium.*

Figure 1.4 illustrates the dynamics of the evolutionary process for an example.

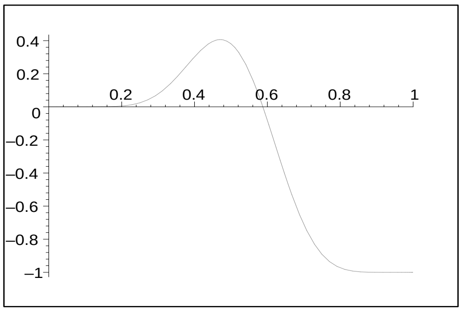


Figure 1.4: The difference in average material payoffs between rewarders and self-centered individuals ($\bar{u}_+ - \bar{u}_s$) plotted as function of γ for $N = 20, d_1 = 1, c_1 = 0, r = 2, d_2 = 5, c_2 = 0, c_r = 1$.The fraction of rewarders in the stable equilibrium of this example is $\gamma^{eq} \approx 0.5876$. If the fraction γ of rewarding individuals is below γ^{eq} then they earn a higher average material payoff and their fraction γ increases. If $\gamma > \gamma^{eq}$ rewarding players earn less and γ decreases. Due to the assumed continuity of the evolutionary dynamics, γ converges to γ^{eq}.

Efficiency: Player 1 cooperates only if his expected material payoff under cooperation is higher than under defection. On the other hand, preferences for rewarding can only survive if player 2 receives a higher material payoff after cooperation and rewarding than after defection. Hence the existence of preferences for rewarding can only lead to a Pareto-improvement (in material payoffs) relative to a purely self-interested population. But for $k^* < N-1$ non-rewarding self-interested players survive, too. Hence, inefficient defection occurs in some groups and the outcome is still inefficient.

Comparative Statics for Case 1

The fraction γ^{eq} of rewarders in the unique stable equilibrium is characterized by the equation $\bar{u}_+(\gamma^{eq}) - \bar{u}_s(\gamma^{eq}) = 0$. Inserting equation 1.6 and rearranging leads to

$$c_2 - c_r - d_2 = c_r \sum_{k=1}^{N-1-k^*} \frac{(N-1-k^*)!k^*!}{(N-1-k^*-k)!(k^*+k)!} \left(\frac{\gamma^{eq}}{1-\gamma^{eq}}\right)^k , \tag{1.7}$$

for $0 < \gamma^{eq} < 1$. The comparative statics is easily derived from this condition.

First, we consider the dependence of the equilibrium fraction γ^{eq} of rewarding players on the group-size N. N enters into equation 1.7 not only directly but also via $k^* = \left[N\frac{c_2-d_2}{r}\right]$. Due to the truncation, k^* is only almost proportional to N. In general, a higher group size N tends to decrease γ^{eq}. But, for some values, the truncation can invert this effect slightly. To avoid such problems, in the following proposition we concentrate on sequences of N for which $\frac{k^*}{N} \equiv c$ is kept constant.

Proposition 3 *An increase in the group size N, keeping $\frac{k^*}{N}$ constant, lowers the fraction γ^{eq} of preferences for rewarding in equilibrium.*

Intuitively, larger groups reduce the probability of being pivotal. Therefore, the advantage of being a rewarder is reduced. Hence, the fraction of rewarding players decreases in equilibrium.[29] This result is consistent with the common feeling that in large anonymous groups the level of cooperation is lower. The influence of a single player on the

[29] The last argument is not entirely complete. The probability of having to bear the costs for rewarding may also decrease and therefore a counterbalancing effect may arise. We can show that the equilibrium fraction of rewarders decreases with the group size, but so far we have not be able to show whether the equilibrium fraction does, or does not, converge to zero if the group size goes to infinity. Numerical results suggests that γ^{eq} decreases only slowly and may not converge to zero.

reputation of a large group is small. In larger groups, a smaller number of rewarding players survive in equilibrium.

Now we consider the dependence of γ^{eq} on the parameters of the game. We start with the influence of the costs c_r player 2 has to incur if he rewards cooperation.

Proposition 4 *Higher costs c_r of rewarding lead to a lower fraction γ^{eq} of the preferences for rewarding in equilibrium. Furthermore,* $\lim_{c_r \to 0} \gamma^{eq} = 1$ *and* $\lim_{c_r \to (c_2 - d_2)} \gamma^{eq} = 0$.

Intuitively, higher costs of rewarding do not influence the incentives of player 1, but reduce the fitness of rewarding players 2. Therefore, their fraction is reduced in equilibrium.

Proposition 5 *Higher gains of cooperation $(c_2 - d_2)$ lead to a higher fraction γ^{eq} of rewarding players 2 in equilibrium. Furthermore,* $\lim_{(c_2 - d_2) \to c_r} \gamma^{eq} = 0$ *and* $\lim_{(c_2 - d_2) \to \infty} \gamma^{eq} = 1$.

The intuition is again straightforward. If gains of cooperation increase, then gains from being pivotal increase for a rewarding player 2. The costs are not affected. Therefore, the fraction of rewarding players 2 increases.

Lemma 1 *If the threshold k^* of rewarding players 2 in a group (above which players 1 in that group start to cooperate) increases, then the fraction of rewarding players 2 in equilibrium increases.*

An increase in k^* means that there have to be more rewarding players 2 in a group in order to induce cooperation of player 1. Hence a smaller number of self-interested players 2 can free-ride without putting cooperation in danger. Therefore, the total number of self-interested players 2 decreases.

From lemma 1 we can easily derive two further results. The costs of cooperation for player 1 $(d_1 - c_1)$ and the amount of the possible reward r enter in equation 1.7 only through k^*. Hence,we obtain

Corollary 1 *The equilibrium fraction γ^{eq} of rewarding players increases (weakly) if the costs $(d_1 - c_1)$ of cooperation for player 1 increase.*

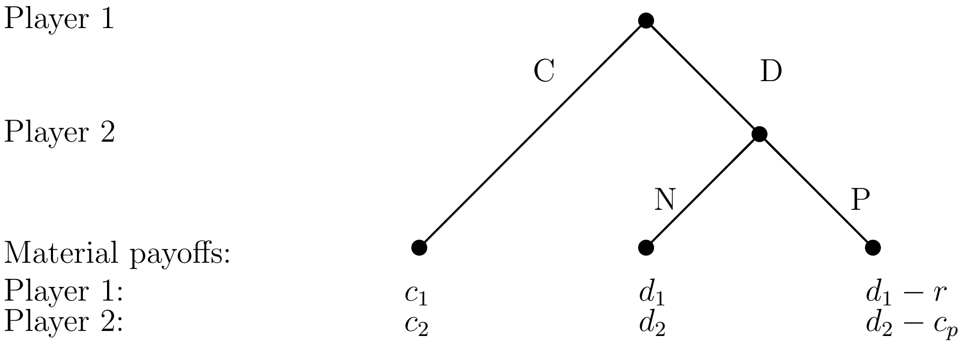
Corollary 2 *The equilibrium fraction γ^{eq} of rewarding players decreases (weakly) if the amount r by which a player 1 can be rewarded cooperation increases.*

Both corollaries might seem counterintuitive at first glance. But the intuition is similar to that of lemma 1. Increasing costs of cooperation or a decreasing rewards make it more difficult to induce player 1 to cooperate. Therefore, free-riding by a self-interested player 2 becomes more likely to destroy cooperation. Hence, the fraction of self-interested players has to decrease in equilibrium.

1.2.2 Case 2: Costly Punishment

In case 1, player 2 had only the possibility of reciprocating positively, i.e. rewarding a friendly action. In case 2, we analyze the evolution of preferences if it is only possible for player 2 to punish hostile behavior (i.e. defection) of player 1. This punishment is costly[30]. The interaction is illustrated in Figure 1.5.

Figure 1.5: Interaction in case 2



with $c_1 + p > d_1 > c_1; c_2 > d_2; c_p > 0$.

Player 2 has either preferences for punishing or self-interested preferences. A punishing player 2 is willing to incur the costs for punishing player 1 in case of defection. But if player 2 is self-interested, he avoids these costs and does not punish defection

[30] We might allow for this punishment after cooperation as well as after defection of the first player. But - similar to case 1 - preferences which lead to punishment after cooperation (e.g. spiteful preferences) vanish in our model due to natural selection. Again, we simplify the analysis by looking at the possibility of punishment only if player 1 defects.

of player 1. Player 1 maximizes his expected material payoff. In a group where k of the N players 2 have preferences for punishing player 1 expects an material payoff of $(d_1 - \frac{k}{N}p)$ after defection. After cooperation he receives a material payoff of c_1. Therefore, player 1 cooperates if and only if[31] $c_1 > d_1 - \frac{k}{N}p$ or equivalently if $k > N\frac{d_1-c_1}{p}$. Analogously to case 1 we denote the threshold by k^{**}.

Definition 2 *k^{**} is the largest number of punishing players 2 in a group that is still insufficient to induce a self-interested player 1 to cooperate. In other words*
$k \leq k^{**} \Rightarrow$ player 1 defects
$k > k^{**} \Rightarrow$ player 1 cooperates.

The calculation of k^{**} is straightforward:

$$k^{**} = \left[N\frac{d_1 - c_1}{p} \right]. \tag{1.8}$$

k^{**} is an integer with $0 \leq k^{**} \leq N - 1$.

In groups with k^{**} or fewer punishing players 2 no cooperation occurs. Players 1 defect. In response, punishing players 2 receive material payoffs of $(d_2 - c_p)$. Self-interested players 2 avoid costs of punishing and receive higher material payoffs of d_2. In groups with more than k^{**} punishing players 2, players 1 cooperate. Therefore, players 2 receive - independently of their types - material payoffs of c_2. The payoff structure is summarized in table 1.2.

Table 1.2: Material payoffs of player 2

	Payoffs in groups with $k \leq k^{**}$	Payoffs in groups with $k > k^{**}$
punisher	$d_2 - c_p$	c_2
self-interested	d_2	c_2

Now let γ be the fraction of punishers in the total population. Analogously to case 1 self-interested players 2 receive an expected material payoff of

$$\bar{u}_s(\gamma) = d_2 \sum_{k=0}^{k^{**}} B_{N-1,\gamma}(k) + c_2 \sum_{k=k^{**}+1}^{N-1} B_{N-1,\gamma}(k) \tag{1.9}$$

[31] Again, we assume the tie breaking rule that player 1 defects if he is indifferent.

and punishing players 2 the expected material payoff of

$$\bar{u}_{-}(\gamma) = (d_2 - c_p) \sum_{k=0}^{k^{**}-1} B_{N-1,\gamma}(k) + c_2 \sum_{k=k^{**}}^{N-1} B_{N-1,\gamma}(k). \tag{1.10}$$

Due to the assumption of payoff monotonicity, the fraction of punishers grows (falls) if punishing players 2 receive a higher (lower) average payoff than self-interested players 2. Hence we are interested in the sign of the difference

$$\bar{u}_{-}(\gamma) - \bar{u}_s(\gamma) = (c_2 - d_2) B_{N-1,\gamma}(k^{**}) - c_p \sum_{k=0}^{k^{**}-1} B_{N-1,\gamma}(k) \tag{1.11}$$

$$= \quad (c_2 - d_2) \frac{(N-1)!}{(k^{**})!(N-1-k^{**})!} \gamma^{k^{**}} (1-\gamma)^{N-1-k^{**}} - c_p \sum_{k=0}^{k^{**}-1} \frac{(N-1)!}{k!(N-1-k)!} \gamma^k (1-\gamma)^{N-1-k}.$$

For $0 < \gamma < 1$ follows

$$\bar{u}_{-}(\gamma) - \bar{u}_s(\gamma) \quad \gtreqless 0 \tag{1.12}$$

$$\Leftrightarrow \quad c_2 - d_2 \quad \gtreqless c_p \sum_{k=0}^{k^{**}-1} \frac{(k^{**})!(N-1-k^{**})!}{k!(N-1-k)!} \left(\frac{1-\gamma}{\gamma} \right)^{k^{**}-k}. \tag{1.13}$$

The right hand side of equation 1.13 is strictly decreasing and continuous in γ, tends to 0 if γ tends to 1 and to infinity if γ tends to zero. The left hand side of equation 1.13 has a fixed positive value. Hence, there exists only one equilibrium of mixed types that is unstable. We denote the fraction of punishers in this unstable equilibrium by γ^{cut}. The only stable equilibria are the corner solutions[32].

Theorem 2 *Let $k^{**} > 0$. Then, the two monomorphic equilibria - in which either all players have preferences for punishing or all players have self-interested preferences - are stable.*

The unique mixed equilibrium is not stable.

In contrast to case 1, the option for punishing defection drives the population to

[32] Like in Proposition 2 we could generalize this result. Take any trait such that (a) when the fraction of a group possessing the trait is above s^* (with $\frac{1}{N} < s^* < \frac{N-1}{N}$) then all agents do equally well; (b) when the fraction is less then s^* then all agents in the group do worse, but those without the trait do better; (c) agents are randomly assigned to groups. Then the two monomorphic equilibria - in which either all agents do have the trait or all agents do not have the trait are stable.

a monomorphic state. Either a "culture of punishment" develops, where all players are willing to punish, or a "culture of laissez faire", where nobody bothers to punish defectors. Figure 1.6 illustrates the evolutionary dynamics in Case 2.

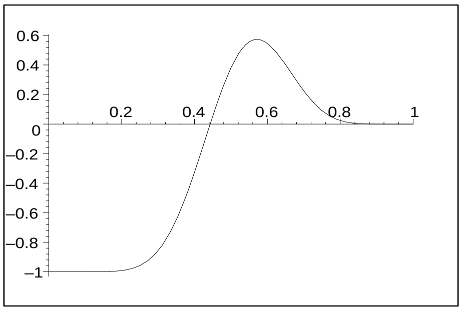


Figure 1.6: The difference in average material payoffs between punishers and self-centered individuals ($\bar{u}_- - \bar{u}_s$) is plotted as a function of γ for $N = 20, d_1 = 1, c_1 = 0, p = 2, d_2 = 5, c_2 = 0, c_p = 1$. The mixed equilibrium at $\gamma^{sep} \approx 0.443$ is unstable and separates the basins of attraction of both stable monomorphic equilibria. If $\gamma < \gamma^{sep}$ punishers perform worse and γ decreases to 0. If $\gamma > \gamma^{sep}$punishers perform better and γ increases to 1.

Theorem 2 is very intuitive. If virtually no player 2 is willing to punish defection, a single punisher is very unfit. In almost any group he is the only punisher and is unable to enforce cooperation of player 1. Player 1 defects and the punishing player 2 has to pay the costs c_p of punishing. Therefore, he is less fit than a self-interested player 2 who does not punish. On the other hand, if virtually all players 2 are willing to punish, they seldom have to prove this. Players 1 in almost all groups cooperate in order to avoid punishment. Only in a few groups in which the number of punishing players 2 is below the threshold k^{**}, the self-interested and punishing players 2 receive different payoffs. But most of these groups are just one punisher below the threshold. In these groups, a punishing player 2 is pivotal in inducing cooperation. Therefore, he benefits from his preferences.

In the equilibrium of a population of punishers, players 1 always cooperate and no player 2 has to prove his willingness to punish. How would the results change if players 1 make mistakes and fail to cooperate sometimes? Appendix 1.5.3 demonstrates that results change only slightly if the probabilities of mistakes are sufficiently small. There remain two stable equilibria. The equilibrium consisting only of self-interested preferences remains stable. However, a population consisting only of punishers is no longer stable. A small fraction of self-interested players can invade. But the fraction of self-interested invaders remains arbitrarily small if probabilities of mistakes are sufficiently small[33]. Hence, there might still develop a culture of punishment with a high fraction of punishers and a small fraction of self-interested players.

The results of Case 2 also differ from Case 1 in terms of efficiency (in material payoffs). In order to be able to rank the outcomes we take the point of view of a player who does not know yet whether he plays in player-position 1 or 2 and might play in either position with equal probability. Then, cooperation is efficient if $d_1 - c_1 < c_2 - d_2$, i.e. if player 2 profits more from cooperation than player 1 loses. However, defection (and no punishment) is efficient if $d_1 - c_1 > c_2 - d_2$. The option for punishing defection can enforce complete cooperation (in a world without mistakes and in the right equilibrium). If $d_1 - c_1 < c_2 - d_2$, this is efficient. But cooperation can also be enforced by the threat of punishment in cases where cooperation is inefficient. Hence, the possibility of punishing defection can be both efficiency enhancing or efficiency reducing.

The unstable mixed equilibrium separates the basins of attraction of both stable equilibria. If the initial fraction of punishers is below γ^{cut} then only self-interested players survive, otherwise only punishers. The lower the value of γ^{cut} the more initial population states evolve to a population of punishers. We relegate the comparative statics of γ^{cut} to appendix 1.5.2.

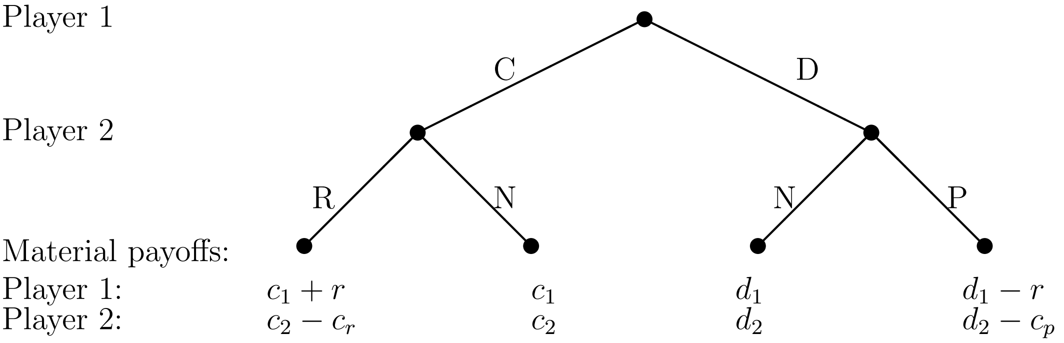
In case 2 there are two equilibria and we don't know whether a "culture of punishment" or a "culture of laissez faire" develops. However, case 3 suggests that the survival of preferences for punishing becomes more likely if player 2 has both options - punishing and rewarding. In fact, under suitable conditions only an entirely punishing population forms an evolutionary stable equilibrium in case 3.

[33] However, a moderate probability of mistakes may result in a significant shift of the punisher equilibrium. See Appendix 1.5.3 for details.

1.2.3 Case 3: Costly Rewarding or Costly Punishment

In case 3 player 2 has both options - costly punishing after defection and costly rewarding after cooperation. This allows us to analyze the co-evolution of preferences for rewarding and preferences for punishing, i.e. how the evolution of one side of reciprocity influences the evolution of the other side. The interaction is illustrated in Figure 1.7.

Figure 1.7: Interaction in case 3



with $c_1 + r > d_1 > c_1$ and $c_2 > c_2 - c_r > d_2 > d_2 - c_p$.

All players 1 maximize their expected material payoffs. There are four different types of players 2 [34]: **Self-interested** players neither reward cooperation nor punish defection. **Punishers** do not reward cooperation, but do punish defection. **Rewarders** reward cooperation, but do not punish defection. **Reciprocal** players both reward cooperation and punish defection. In order to reduce technical problems, we make the following[35]

Assumption 5 *The material loss p for player 1 after being punished equals his material gain r after being rewarded, i.e. $p = r$.*

Due to Assumption 5, punishers and rewarders have exactly the same influence on the behavior of players 1 in their group. Hence, material payoffs of all other players 2

[34] Again, we neglect generic cases of preferences which associate the same subjective utility with different outcomes.

[35] The general intuition for the results of this section holds without this assumption, but assumption 5 simplifies the analysis considerably.

are not affected if we replace a punisher by a rewarder or vice versa. We know from the analysis of case 2 that preferences for punishing are more successful if their own fraction grows. Hence, punishers also profit from a growing fraction of rewarders. Any kind of reciprocity helps to induce cooperation of players 1 and reduces the costs of being a punisher.

Remark 2 *Higher fractions of rewarders and higher fractions punishers enhance the evolutionary success of preferences for punishing.*

Conversely, we know from case 1 that the evolutionary success of preferences for rewarding relative to self-interested preferences decreases if their own fraction becomes too large. Hence the same must hold for too large a fraction of punishers. Furthermore, relative to preferences for punishing, the success of preferences for rewarding is reduced by an increase of the fraction of any type of reciprocity. The higher the fraction of rewarders or punishers, the more groups are above the threshold for cooperation. Therefore, costs of rewarding grow, whereas the costs of punishing fall.

Remark 3 *Higher fractions of rewarders and higher fractions of punishers reduce the evolutionary success of preferences for rewarding relative to the success of preferences for punishing.*

This interdependence between the evolution of both types of reciprocity has interesting consequences. Consider an entirely self-interested population. Preferences for punishing cannot invade such a population directly, as shown in Case 2. But preferences for rewarding can invade (see Case 1). If enough rewarders invade, they may serve as a “catalyst” and enable the invasion of punishers. The more punishers invade, the more successful they become and finally they drive out self-interested players as well as rewarders.

Remark 4 *Preferences for rewarding may serve as a catalyst for the evolution of preferences for punishing. Rewarders can invade an entirely self-interested population. Their existence enables punishers to invade, too. Finally, preferences for punishing become more and more successful and drive out self-interested preferences as well as preferences for rewarding.*

Now we look for stable equilibria in case 3. First, we check for stable monomorphic populations, i.e stable populations of only one preference-type.

Proposition 6 *The only monomorphic stable equilibrium consists entirely of punishers.*

Are there other stable equilibria consisting of several preference types? The answer depends on the parameters of the model. For certain parameters, this is the only stable equilibrium. For others, further stable equilibria exist. It is easier to capture the basic intuition if reciprocal preferences are neglected. Hence for the moment we restrict ourselves to the possibilities of self-interested preferences, preferences for rewarding and preferences for punishing. Consider a population consisting only of rewarders and self-interested and players. According to Case 1 this population evolves towards a unique equilibrium containing both preference types. Can a small fraction of punishers invade this equilibrium? The answer depends on the fraction γ^{eq} of rewarders in equilibrium. Since we assumed $p = r$, the effect of a rewarding player 2 on any other player 2 in his group is precisely the same as the effect of a punishing player 2 at the same place. Hence, preferences for punishing can invade this equilibrium if, and only if, the fraction γ^{eq} of rewarders in this equilibrium (determined by Equation 1.7) is higher than the threshold γ^{cut} (determined by Equation 1.13) above which preferences for punishing become more successful than self-interested ones. Preferences for punishing become relatively more successful, the higher their own fraction of the population. Therefore, once preferences for punishing can invade, they drive out all other preferences and the dynamics leads to the monomorphic equilibrium of preferences for punishing.

Proposition 7 *Let γ^{eq} be defined by equation 1.7 and γ^{cut} by equation 1.13.*
If reciprocal preferences are neglected, i.e. only the subspace of self-interested preferences, preferences for rewarding and preferences for punishing is considered, then

a) *if $\gamma^{eq} > \gamma^{cut}$, then the only stable equilibrium is a monomorphic population, where all players have preferences for punishing. The population converges to this equilibrium from any interior state.*

b) *If $\gamma^{eq} < \gamma^{cut}$, then there are precisely two stable equilibria. One stable equilibrium is the monomorphic population of preferences for punishing. In the other stable equilibrium preferences for rewarding and self-interested preferences coexist*[36]*. In this equilibrium the fraction of preferences for rewarding is γ^{eq}.*

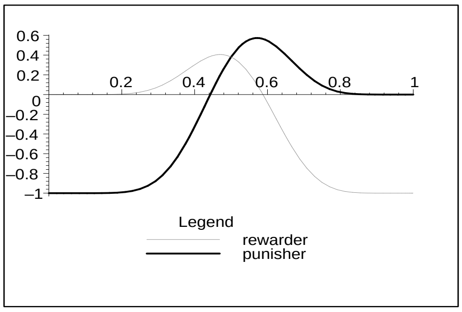


Figure 1.8: Case 3 with $\gamma^{eq} > \gamma^{cut}$: $(\bar{u}_+ - \bar{u}_s)$ and $(\bar{u}_- - \bar{u}_s)$ as functions of $\gamma \equiv \gamma_+ + \gamma_-$ for $N = 20, d_1 = 1, c_1 = 0, r = 2, d_2 = 5, c_2 = 0, c_r = 1$.

Figure 1.8 illustrates the dynamics of the evolutionary process in case 3 with the parameters of our previous examples. Here we have $\gamma^{eq} > \gamma^{cut}$ and the equilibrium with a fraction γ_{eq} of rewarders and a fraction $(1 - \gamma^{eq})$ of self-interested players is not stable: punishers earn a higher average payoff, invade successfully and drive out all other preferences.

Including reciprocal preferences does not change the basic intuition. Preferences for punishing still form a stable equilibrium and a mixture of a fraction of γ^{eq} with preferences for rewarding and $1 - \gamma^{eq}$ self-interested preferences remains a stable equilibrium under the slightly more restrictive condition $\gamma^{eq} < \min\{\gamma^{cut}; \gamma^h\}$, where γ^h is

[36] Notice that, even in the case of Prop.7b where we have still two stable equilibria, it is in Case 3 more likely to end up in the monomorphic equilibrium (compared to Case 2). This is meant in the spirit of the model by Kandori et al. [50]: imagine that each member of the entire (large but finite) population mutates with small probability to any other preference-type. Then, the minimum number of mutations necessary to move from the monomorphic equilibrium to the basin of attraction of the other equilibrium is exactly the same as in case 2. But the other way round fewer mutations are sufficient to move the population from the bi-morphic equilibrium to the basin of attraction of the monomorphic equilibrium. That is because the rewarders are advantageous for the invasion of the punishing type.

defined by the equation

$$c_2 - d_2 - c_r = c_p \sum_{k=0}^{k^{**}-2} \frac{(k^{**}-1)!}{k!} \frac{(N-k^{**})!}{(N-1-k)!} \left(\frac{1-\gamma^h}{\gamma^h} \right)^{k^{**}-1-k} . \tag{1.14}$$

The tightening of the condition is necessary to ensure that reciprocal preferences cannot invade the mixed equilibrium either. Furthermore, reciprocal preferences can be part of an equilibrium only under very special conditions. If most groups induce players 1 to cooperate, preferences for punishing tend to outperform reciprocal ones, since they do not have to bear the costs of rewarding. If, on the other hand, most groups are not able to induce cooperation, then preferences for rewarding tend to outperform reciprocal ones since they do not bear the costs of punishing in the frequent cases of defection. But for certain parameters there exist equilibria with a positive fraction of reciprocal preferences. These additional equilibria are not robust to small changes in parameters of the model and are not very plausible. Therefore, we relegate the discussion of these equilibria to the appendix 1.5.4 and focus attention on the discussed equilibria summarized in the following

Proposition 8 *Let γ^{eq} be defined by equation 1.7,γ^{cut} by equation 1.13 and γ^h by equation 1.14. Then, for any payoff-monotonic selection dynamics holds*

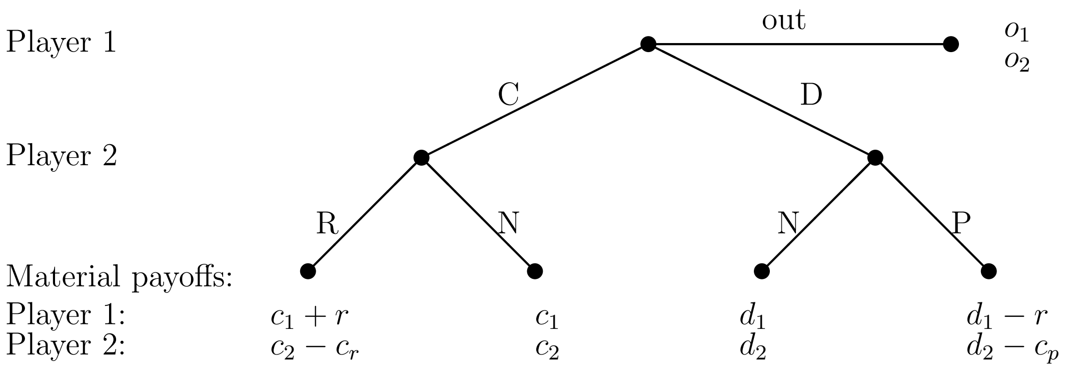
a) *a monomorphic population of punishers forms a stable equilibrium.*

b) *If $\gamma^{eq} < \min\{\gamma^{cut}, \gamma^h\}$, then also a population with a fraction γ^{eq} of rewarders and a fraction $(1-\gamma^{eq})$ of self-interested players forms a stable equilibrium.*

So far, preferences for punishing will - once they invade - drive out preferences for rewarding completely. In equilibrium either preferences for rewarding or preferences for punishing survive, but not both. However, both sides of reciprocity can survive in one equilibrium, if player 1 cannot be forced to participate in the interaction, i.e. player 1 has an additional outside option as illustrated in figure 1.9.

Now, a population consisting only of punishers is no longer stable. The threat of punishment alone can only deter player 1 from defecting. But player 1 opts out as long as he does not expect to be rewarded for cooperation. Analogous to case 1, reciprocal

Figure 1.9: Interaction in case 3 with outside option



with $c_1 + r > d_1 > o_1 > c_1$; $c_2 > c_2 - c_r > o_2 > d_2 > d_2 - c_p$.

players who reward and punish can invade the population of punishers. In some groups their willingness to reward induces players 1 to cooperate instead of opting out. This makes reciprocal preferences initially more successful until reciprocal preferences and preferences for punishing are in a mixed equilibrium[37]. Hence there is an equilibrium in which all or most players are willing to punish defection, some of them do reward cooperation and others don't[38].

A different explanation for the survival of both types of reciprocity arises if individuals engage in different types of interaction - sometimes similar to case 1, sometimes similar to case 2 or case 3. If players have general preferences and do not have different preferences for different types of interaction, then some rewarders, some punishers and some reciprocators can survive[39].

[37] This equilibrium is only Lyapunov stable but not asymptotically stable. That is because payoffs do not change if some reciprocal players are replaced by rewarders. But the set of population states with a fraction of $(1 - \gamma^{eq})$ of punishers, $\gamma \in [0; \gamma^{eq}]$ of reciprocators and $(\gamma^{eq} - \gamma)$ of rewarders forms an asymptotically stable set of equilibria.

[38] For certain parameters there exist further equilibria, but the detailed analysis is beyond the scope of this chapter.

[39] The question of how far preferences may depend on the respective interaction is a subtle one. On the one hand, preferences should not be expected to evolve independently for any type of interaction. On the other hand, people may well classify interactions by broad categories and their preferences may well depend on whether they assign a certain interaction to one category or another. Empirical evidence as well as theoretical approaches in the direction of Samuelson [79] could offer interesting insights to this question.

1.3 Discussion

Our results hold for any payoff-monotonic evolutionary dynamics. Hence, our selection dynamics can be interpreted as genetic evolution, cultural evolution or as a process of learning by success and failure. Furthermore, our results are robust to small mistakes: in appendix 1.5.3 we demonstrate that sufficiently small mistake-probabilities of players change the results only slightly.

How well do the findings of our evolutionary analysis fit empirically observed human behavior? A recent experimental paper by Andreoni et al. [4] studies human behavior in four treatments called Dictator, Carrot-Stick, Carrot and Stick. The last three treatments have the same structure as our three analyzed interactions, with the distinction that the choice variables in the experiment are not binary: first proposers can choose the fraction of their wealth they want to transfer to a responder. Then the responder can choose how much money he wants to invest in rewards or punishments[40]. The “Carrot treatment” and the “Stick treatment” replicate qualitatively the findings of several experimental studies[41]: the higher the transfer of the proposer, the higher the average reward and the lower the average punishment by the responders. In particular, virtually no punishments occur when offers are above the equal share. Furthermore, even after very generous proposals, some responders do not invest in rewards and, even after very small proposals, some responders do not spend money on punishments.

Most interestingly for our purpose, Andreoni et al. compare experimentally the demand for rewards in the Carrot versus the Carrot-Stick treatment and the demand for punishments in the Stick versus the Carrot-Stick treatment. If rewards **and** punishments are available, the demand for rewards decreases significantly (compared to the Carrot treatment) and the demand for punishments increases significantly for very low offers (compared to the Stick treatment). For medium and large offers, the demand for punishments does not change significantly.

These experimental findings have a natural interpretation in the light of our evo-

[40] For each cent invested by the responder five cents were added to (if it was a reward) or subtracted (if it was a punishment) from the proposer’s wealth.

[41] These studies typically analyze either the possibility of punishing or the possibility of rewarding, but not both.

lutionary model. Suppose the willingness to punish or reward differs for different contexts[42]. In a context where only rewards are available, in agreement with the experimental findings of the Carrot treatment, Case 1 of our model predicts that some subjects reward and some don't. In a context where only punishments are available, Case 2 of our model suggests that either a norm of punishment or a norm of no punishment prevails. We can interpret the observations for offers above the equal share as a norm of no punishment. Offers below the equal share are often punished. However, in contrast to the predictions of Case 2 of our model, not all responders punish low offers. Two simple extensions can naturally explain the experimental findings: first we show in Appendix 1.5.3 that already relatively small mistake-probabilities of the proposer shift the "only punishers equilibrium" to a mixed equilibrium with a non negligible fraction of non-punishers. Second, since there exist two equilibria in Case 2, different norms may have evolved for different real life contexts. In the artificial situation of the laboratory environment, some subjects may imagine themselves in a real life interaction where punishment is the norm, while others may compare it to an environment where non-punishment is the norm. Then the result is a mixture of punishers and non-punishers such as observed in the experiment.

Most interestingly, in an environment where rewards and punishments are both available, our evolutionary model predicts that the propensity to reward cooperation tends to be crowded out by a propensity to punish non-cooperation. This is consistent with the experimental results: the demand for rewards is drastically reduced in the Carrot-Stick treatment compared to the Carrot treatment, whereas the demand for punishments in the Carrot-Stick treatment compared to the Stick treatment increases at least in response to very low offers.

[42] In fact, if different types of interaction play an important role during the process of evolution, then natural selection will favor such context dependent preferences. Also, the well documented fact that framing effects can influence experimental findings significantly points to context dependent preferences. Admittedly, context dependent preferences cause serious and subtle problems for our modelling strategies as economists. Therefore, it may help to think instead of general preferences that refer to a context dependent norm.

1.4 Conclusions

This non-assortative group selection model offers an explanation for the evolutionary survival of both sides of reciprocal preferences. Despite the fact that individual behavior and preferences are unobservable, individuals continue to have a marginal effect on the "reputation" of their group, and this influences the behavior of the other players in their group. This effect is sufficient to enable preferences for rewarding and preferences for punishing to survive in the evolutionary competition with self-interested preferences. Both preferences for rewarding and preferences for punishing can induce cooperative behavior. But there is an intrinsic difference between the two preference types: preferences for rewarding tend to coexist with self-interested preferences, whereas preferences for punishing tend either to dominate the population completely or to vanish entirely. Furthermore, rewarders enhance the evolution of preferences for punishing. Preferences for rewarding are able to invade a self-interested population and may then, as a "catalyst", enable the invasion of preferences for punishing. Punishers, on the other hand, crowd out rewarders and may even drive them out completely.

1.5 Appendix

1.5.1 Proofs

Proof of Proposition 2

Equation 1.6 describes the difference between expected material payoffs as a function of the fraction γ of rewarders in the entire population. N is kept constant. The first term of the right hand side has a positive sign, the remaining terms are negative. First we consider small γ. The positive first term is of the order k^* in γ whereas the remaining negative terms are at least of the order $(k^* + 1)$ in γ. Hence the right hand side of equation 1.6 is positive for sufficient small γ. Therefore, γ grows if the fraction of the rewarders is sufficiently small. In other words, an entirely self-interested population is not stable.

In a similar way we prove that the fraction of self-interested players increases if most players have preferences for rewarding, i.e. if γ is close to 1. If γ converges to 1, then the first term of equation 1.6 converges to zero whereas the remaining sum of negative terms converges to $(-c_r)$ (in fact, the last term converges to $-c_r$ and all the remaining terms to zero). Hence, the fraction γ of rewarders decreases if their fraction of the total population is sufficiently large. In other words, a population consisting entirely of rewarders is not stable either, q.e.d.

Proof of Theorem 1

From proposition 2 we know that the difference between the average material payoffs $\bar{u}_{pos}(\gamma) - \bar{u}_s(\gamma)$ is above zero for small γ and below zero for γ close to 1. Since $\bar{u}_{pos}(\gamma) - \bar{u}_s(\gamma)$ is continuous in γ, there must exist an interior γ^{eq} with $\bar{u}_{pos}(\gamma^{eq}) - \bar{u}_s(\gamma^{eq}) = 0$. We will see in the next step, that γ^{eq} is the unique value strictly between 0 and 1 satisfying this equation. Hence, for all values below γ^{eq}, the difference is above 0 and for all values above γ^{eq} the difference is below 0. Hence this equilibrium is stable.

Uniqueness follows directly from the necessary condition for an interior equilibrium, i.e. equation 1.7:

$$c_2 - c_r - d_2 = c_r \sum_{k=1}^{N-1-k^*} \frac{(N-1-k^*)!k^*!}{(N-1-k^*-k)!(k^*+k)!} \left(\frac{\gamma}{1-\gamma}\right)^k . \tag{1.15}$$

The right hand side is strictly increasing in γ. The left hand side is constant. Therefore, equation 1.15 is satisfied at most for one γ^{eq}, q.e.d.

Proof of Proposition 3

We assumed $\frac{k^*}{N} \equiv q$ constant, i.e. $k^* = qN$ with $0 < q < 1$. We can rearrange the equilibrium condition 1.7 into

$$\begin{aligned} c_2 - c_r - d_2 &= c_r \sum_{k=1}^{N-1-k^*} \left(\left(\prod_{l=1}^{k} \frac{N-k^*-l}{k^*+l} \right) \left(\frac{\gamma}{1-\gamma}\right)^k \right) & (1.16) \\ &= c_r \sum_{k=1}^{N(1-q)-1} \left(\left(\prod_{l=1}^{k} \frac{(1-q)N-l}{qN+l} \right) \left(\frac{\gamma}{1-\gamma}\right)^k \right) . & (1.17) \end{aligned}$$

Now we prove that for constant γ the right hand side is strictly increasing in N. Since the left hand side is constant, γ^{eq} has to fall in order to equilibrate the two sides again. The number of terms increases with N. Since all terms in equation 1.16 are positive it is sufficient to prove that each term increases in N. By extending N to real numbers we find

$$\frac{\partial}{\partial N}\left(\frac{(1-q)N-l}{qN+l}\right) = \frac{l}{(qN+l)^2} > 0, \tag{1.18}$$

q.e.d.

Proof of Proposition 4

The right hand side of equation 1.7 is strictly increasing in γ. For any value of c_r equation 1.7 must hold in equilibrium. If we now choose a new $c_r^{new} > c_r$, the left hand side becomes smaller whereas, if we keep γ fixed, the right hand side would increase. Therefore, γ has to decrease in order to decrease the right side and satisfy equation 1.7 again. Hence the new equilibrium fraction of rewarders is lower. Furthermore, if c_r

tends to 0, then the left hand side tends to the positive value $c_2 - d_2$, whereas the right hand side would tend to zero if $\frac{\gamma}{1-\gamma}$ remained bounded from above. Therefore, γ must tend to 1 if c_r tends to zero. Finally, if c_r tends to $(c_2 - d_2)$, then the left hand side tends to zero, but the right hand side can only tend towards zero if γ tends to zero, too, q.e.d.

Proof of Proposition 5

The proof of proposition 5 is completely analogous to the proof of proposition 4.

Proof of Lemma 1

We consider the equilibrium condition in form of equation 1.16. The left hand side is not affected by a change in k^*. The right hand side is affected in two ways if k^* increases. First, the number of terms is reduced and second, each of the remaining terms becomes smaller. Both effects diminish the value of the right hand side. Therefore, γ^{eq} has to increase in order to equilibrate both sides again, q.e.d.

Proof of Corollary 1 and Corollary 2

$k^* = [N\frac{d_1-c_1}{r}]$ is weakly increasing in $(d_1 - c_1)$ and weakly decreasing in r. Hence, corollary 1 and 2 follow directly from lemma 1, q.e.d.

Proof of Proposition 6

Neither an entirely self-interested population nor a population consisting entirely of rewarders can be stable, since both are not even stable in the subspace of rewarders and self-interested players (case 1). Furthermore, an entirely reciprocal population can not be stable, since punishers perform strictly better in the subspace of punishers and reciprocators. (In this subspace cooperation occurs in all groups, but preferences for punishing save the costs of rewarding.) The only stable monomorphic population consists entirely of punishers. Take any state γ in a sufficiently small neighborhood of $(\gamma_+ = 0, \gamma_- = 1, \gamma_{rc} = 0, \gamma_s = 0)$. Rewarders, reciprocators and self-interested players all earn strictly less than punishers in such a neighborhood (rewarders and reciprocators because they have to pay the costs of rewarding in most interactions; self-interested players by reasons completely analog to case 2). Therefore the fraction of punishers grows faster than all other types and the state must converge to the monomorphic punisher population.

Proof of Proposition 7

Step 1 proves that the equilibria in proposition 7 are stable. Step 2 shows that no other equilibrium can be stable (in the subspace without reciprocal preferences). Step 3 proves that in case a the population converges from any interior state to a monomorphic population of punishers.

Step 1: The monomorphic population of punishers forms a stable equilibrium by prop. 6. Theorem 1 in case 1 states that there exists a unique stable equilibrium with a fraction of γ^{eq} rewarders in the subspace of self-interested preferences and preferences for rewarding. It remains to be shown that, for $\gamma^{eq} < \gamma^{cut}$ (i.e. in case b)), this equilibrium is also stable in the subspace which includes preferences for punishing. Let γ_+ be the fraction of rewarders, γ_- the fraction of punishers in the total population and define $\tilde{\gamma} \equiv \gamma_+ + \gamma_-$. Due to $p = r$ (assumption 5) the action of player 1 depends only on the total fraction $\tilde{\gamma}$ of rewarders and punishers. Hence, expected material payoffs of any type of player 2 depend only on $\tilde{\gamma}$, too:

$$\bar{u}_s(\tilde{\gamma}) = d_2 \sum_{k=0}^{k^*-1} B_{N-1,\tilde{\gamma}}(k) + d_2 B_{N-1,\tilde{\gamma}}(k^*) + c_2 \sum_{k=k^*+1}^{N-1} B_{N-1,\tilde{\gamma}}(k) \tag{1.19}$$

$$\bar{u}_+(\tilde{\gamma}) = d_2 \sum_{k=0}^{k^*-1} B_{N-1,\tilde{\gamma}}(k) + (c_2 - c_r)\, B_{N-1,\tilde{\gamma}}(k^*) + (c_2 - c_r) \sum_{k=k^*+1}^{N-1} B_{N-1,\tilde{\gamma}}(k) \tag{1.20}$$

$$\bar{u}_-(\tilde{\gamma}) = (d_2 - c_p) \sum_{k=0}^{k^*-1} B_{N-1,\tilde{\gamma}}(k) + c_2 B_{N-1,\tilde{\gamma}}(k^*) + c_2 \sum_{k=k^*+1}^{N-1} B_{N-1,\tilde{\gamma}}(k) \tag{1.21}$$

Hence we obtain for the differences in average material payoffs

$$\bar{u}_+(\tilde{\gamma}) - \bar{u}_s(\tilde{\gamma}) = (c_2 - d_2 - c_r)\, B_{N-1,(\tilde{\gamma})}(k^*) + (-c_r) \sum_{k=k^*+1}^{N-1} B_{N-1,(\tilde{\gamma})}(k) \quad (1.22)$$

$$\bar{u}_-(\tilde{\gamma}) - \bar{u}_s(\tilde{\gamma}) = -c_p \sum_{k=0}^{k^*-1} B_{N-1,(\tilde{\gamma})}(k) + (c_2 - d_2)\, B_{N-1,(\tilde{\gamma})}(k^*) \quad (1.23)$$

$$\bar{u}_-(\tilde{\gamma}) - \bar{u}_+(\tilde{\gamma}) = -c_p \sum_{k=0}^{k^*-1} B_{N-1,(\tilde{\gamma})}(k) + c_r \sum_{k=k^*}^{N-1} B_{N-1,(\tilde{\gamma})}(k) \quad (1.24)$$

The first equation corresponds to equation 1.7 of case 1 - only γ is now replaced by $(\tilde{\gamma})$. In particular, the fraction of rewarders will increase relative to the fraction of self-interested preferences if $\tilde{\gamma} < \gamma^{eq}$ (and decrease if $\tilde{\gamma} > \gamma^{eq}$). Similarly, the second equation corresponds to equation 1.13 of case 2. In particular, this means that the fraction of punishers decreases relative to the fraction of self-interested players as long as $\tilde{\gamma} < \gamma^{cut}$. By putting these two observations together we see that, for $\gamma^{eq} < \tilde{\gamma} < \gamma^{cut}$, both the fraction of preferences for rewarding γ_+ and the fraction of preferences for punishing γ_- decrease relative to the fraction $(1-\gamma_+-\gamma_-)$ of self-interested preferences. Hence, $(\tilde{\gamma})$ decreases in absolute terms. The dynamic is continuous in γ_+ and γ_- and therefore also in $(\tilde{\gamma})$. Hence, if initially $\gamma_+^0 + \gamma_-^0 < \gamma^{cut}$, then $(\tilde{\gamma})$ remains below (or equal to) $\min\{\gamma_+^0 + \gamma_-^0, \gamma^{eq}\}$. In particular, γ_- decreases relative to the fraction of self-interested preferences with a rate strictly above a constant strictly positive rate. Hence γ_- also converges absolutely to zero. Now it is straightforward to prove asymptotic-stability of the population-state $(\gamma_+, \gamma_-, \gamma_s) = (\gamma^{eq}, 0, 1-\gamma^{eq})$. Let $\epsilon < \frac{\gamma^{cut}-\gamma^{eq}}{3}$. For any initial population state in the ϵ-neighborhood of $(\gamma^{eq}, 0, 1-\gamma^{eq})$, $\gamma_+^0 + \gamma_-^0 < \gamma^{cut}$ holds. Therefore, γ_- converges to zero. Due to the continuity of all average payoff functions in γ_+ and γ_- the convergence of γ_- to 0 implies convergence of γ_+ to γ^{eq}.

Step 2: Now we prove that there are no other equilibria in the subspace of preferences for rewarding, preferences for punishing and self-interested preferences. First preferences for punishing cannot coexist with self-interested preferences in a stable equilibrium: replacing any small fraction of the self-interested players by punishers enhances material payoffs of punishers relative to self-interested players. Hence an equilibrium containing both types cannot be stable. Furthermore, preferences for rewarding and preferences for punishing cannot form an equilibrium: cooperation would occur in all groups and rewarders earn less because they have to bear the costs of rewarding. A completely rewarding or completely self-interested population is not stable as shown in case 1. Hence the only remaining candidates for stable equilibria are either a population of only punishers (in fact, this equilibrium is stable by prop. 6) or a heterogenous population of self-interested preferences and preferences for rewarding. Case 1 showed that in equilibrium the fraction of preferences for rewarding has to be γ^{eq} and the fraction of self-interested preferences $1-\gamma^{eq}$. In step 1 we have shown that this equilibrium is stable for $\gamma^{eq} < \gamma^{cut}$. It remains to be shown that this equilibrium is not stable for $\gamma^{eq} > \gamma^{cut}$. At $(\gamma_+ = \gamma_{eq}, \gamma_- = 0, \gamma_s = 1-\gamma_{eq})$ we have $\tilde{\gamma} = \gamma_{eq} > \gamma_{cut}$ and therefore punishers earn a strictly higher profit than self-interested players. This contradicts stability as a result of[43]

Lemma 2 *If a state γ is asymptotically stable in some payoff-monotonic selection dynamics, then all types in the support $C(\gamma)$ earn at γ an expected payoff at least as high as any other type.*

Step 3 It remains to be shown that for $\gamma^{eq} > \gamma^{cut}$ the population converges from any interior state to the equilibrium of a monomorphic population of punishers. From any interior state, and for any regular selection dynamics, the population state does not reach the boundaries in finite time[44], i.e. no preference-type vanishes completely in finite time. If initially $\tilde{\gamma} < \gamma^{eq}$, then γ_+ grows with positive rate relative to self-interested preferences as long as $\tilde{\gamma} < \gamma^{eq}$ and, in particular, the point where $(\tilde{\gamma}) = \gamma^{cut} + \frac{\gamma^{eq}-\gamma^{cut}}{2}$ is reached in finite time. In the area where $\gamma^{cut} < (\tilde{\gamma}) < \gamma^{eq}$ preferences for rewarding and preferences for punishing are both more successful than self-interested preferences. Therefore, once $(\tilde{\gamma}) \geq \gamma^{cut} + \frac{\gamma^{eq}-\gamma^{cut}}{2}$ holds, the dynamic process never changes this. Hence, after a finite time, γ_- increases with a strictly positive rate compared to γ_s. In particular, this implies that γ_s converges to 0. Hence $(\tilde{\gamma})$ converges

[43] The proof of lemma 2 is analogous to the proof of prop. 4.8 in Weibull [95] and written upon request.

[44] Compare e.g. Weibull [95] page. 141

to 1 and, in particular, $(\tilde{\gamma}) > \gamma^{eq}$ after some finite time. Then rewarding players 2 become less successful than self-interested players 2 (and therefore less successful than punishing players 2) and converge to 0, too. In the end, only preferences for punishing survive and the fractions of other preferences converge to zero.

Proof of Proposition 8

Part a): See prop 6.

Part b): If $\gamma^{eq} < min\{\gamma^{cut}, \gamma^{h}\}$, then we can show that in the state $(\gamma_+ = \gamma^{eq}, \gamma_- = 0, \gamma_{rec} = 0, \gamma_s = 1 - \gamma^{eq})$ punishers and reciprocators earn strictly less than rewarders and self-interested players. Since average material payoffs of all types are continuous in γ_+, γ_-, γ_{rc} and γ_s, this means that punishers and reciprocators earn also strictly less in a sufficiently small neighborhood of this state. Hence, from any sufficiently close state the population will converge to the state $(\gamma_+ = \gamma^{eq}, \gamma_- = 0, \gamma_{rec} = 0, \gamma_s = 1 - \gamma^{eq})$. That punishers earn a strictly lower material payoff in this equilibrium was show in the proof of proposition 6. To see that reciprocators earn a strictly lower expected material payoff consider the average material payoffs of a rewarder and of a single reciprocal invader in this equilibrium:

$$\bar{u}_+ = d_2 \sum_{k=0}^{k^*-1} B_{N-1,\gamma^{eq}}(k) + (c_2 - c_r) \sum_{k=k^*}^{N-1} B_{N-1,\gamma^{eq}}(k) \tag{1.25}$$

$$\bar{u}_{rc} = (d_2 - c_p) \sum_{k=0}^{k^*-2} B_{N-1,\gamma^{eq}}(k) + (c_2 - c_r) \sum_{k=k^*-1}^{N-1} B_{N-1,\gamma^{eq}}(k) \tag{1.26}$$

Hence

$$\bar{u}_{rc} - \bar{u}_+ = -c_p \sum_{k=0}^{k^*-2} B_{N-1,\gamma^{eq}}(k) + (c_2 - d_2 - c_r) B_{N-1,\gamma^{eq}}(k^* - 1). \tag{1.27}$$

For $0 < \gamma^{eq} < 1$ we obtain by dividing through $B_{N-1,\gamma^{eq}}(k^* - 1)$ the equivalence

$$\bar{u}_{rc} - \bar{u}_+ \gtreqless 0 \tag{1.28}$$

$$\Leftrightarrow c_2 - d_2 - c_r \gtreqless c_p \sum_{k=0}^{k^*-2} \frac{(k^*-1)!}{k!} \frac{(N-k^*)!}{(N-1-k)!} \left(\frac{1-\gamma^{eq}}{\gamma^{eq}}\right)^{k^*-1-k} . \tag{1.29}$$

The right hand side of equation 1.29 is strictly decreasing in γ^{eq}. Furthermore, the right hand side would be equal to the left hand side if $\gamma^{eq} = \gamma^h$ (this was precisely the definition of γ^h). Hence, for $\gamma^{eq} < \gamma^h$, the right hand side is strictly larger than the left hand side and therefore $\bar{u}_{rc} - \bar{u}_+ < 0$, q.e.d.

Preferences in Position 1

The following proposition helps to justify the Assumption 2 that all players 1 maximize their expected material payoff.

Proposition 9 *Let $\mathcal{M}'$ be the set of a finite number of any possible subjective preferences about outcomes and let $\mathcal{M} \supset \mathcal{M}'$ include in addition to each preference type m' in $\mathcal{M}'$ a corresponding type m that has identical preferences in position 2 but purely self-interested preferences in position 1. Let $M = |\mathcal{M}|$ and let $\Delta = \{(\gamma_1, \ldots, \gamma_M) | \sum_{i=1}^{M} \gamma_i = 1\}$ be the state space of all probability distributions over the preference types in $\mathcal{M}$. Furthermore, take Assumption 1 and Assumption 3 that now refers to the Multinomial distribution $M_{N,\gamma_1,\ldots,\gamma_M}(k_1, \ldots, k_M) = \frac{N!}{\prod_{i=1}^{M} k_i!} \prod_{i=1}^{M} \gamma_i$.*

a) *No preference type can earn a higher expected material payoff than the corresponding type that has identical preferences in position 2 but purely self-interested preferences in position 1.*

b) *In any stable state γ all preference types $m_i \in \mathcal{M}$ with a fraction $\gamma_i > 0$ must act consistently with payoff maximization in player-position 1 in all groups that occur with positive probability.*

Proof of Proposition 9 Player 2 conditions his choice of action only on the action chosen by player 1. In particular, he does not condition his behavior on the distribution of preference-types in his group, even if he could do so. This fact is due to our assumption that players have preferences only about outcomes and not about other players' preferences or their distribution. Since player 2 can observe the action of player 1, he can guess the outcomes resulting from his own action directly. Hence player 2 chooses his strategy independently of his beliefs about the other players' preferences and therefore independently of the distribution of preference-types in his group. Now consider the pair of preferences described in part a): In position 2 both types are identical and earn therefore the same expected material payoff. In position 1 expected material payoffs depend only on the choice of whether to cooperate or to defect. Since the self-interested type chooses by definition the option that maximizes his expected material payoff, no other type can do better. We prove part b) by contradiction. Assume that there exists a stable state γ_{st} with a positive fraction of players 1 acting with positive probability in a way that earns them an expected material payoff strictly below the optimum. Then the corresponding type from part a) earns a strictly higher expected material payoff. This contradicts stability as a result of lemma 2, q.e.d.

1.5.2 Comparative Statics for Case 2

Let γ^{cut} be the fraction of punishers in the unstable mixed equilibrium. This fraction separates the basins of attraction of the stable equilibria. If the initial fraction of punishing players is below the cutoff γ^{cut} then this fraction decreases until the entire population has self-interested preferences and nobody punishes defection. If, on the other hand, the initial fraction of punishing players is above the cutoff γ^{cut}, then this fraction increases until the entire population has preferences for punishing. One might therefore interpret the value of γ^{cut} as an indicator for how likely it is to end up in one or the other equilibrium[45]. The comparative statics of γ^{cut} is analogous to case 1 and can be derived directly from equation 1.13.

Higher costs of punishing diminish the basin of attraction of the punisher equilibrium:

[45] Again, this interpretation is in the spirit of the model by Kandori et. al [50], where the size of the basins of attraction determines the long run equilibrium

Proposition 10 *If the costs c_p - which a player 2 has to bear in order to punish - increase, then γ^{cut} increases, i.e. there have to be initially more punishers in order to end up in the punishing equilibrium. Furthermore, $\lim_{c_p \to 0} \gamma^{cut} = 0$ and $\lim_{c_p \to \infty} \gamma^{cut} = 1$.*

The intuition is straightforward: the higher the number of punishing players, the cheaper it is to be a punisher. If the costs of punishing increase, punishers become less fit. Hence punishing players need a higher fraction of punishers in order to be at least as successful as non-punishers.

Higher gains from cooperation for player 2 are good for punishers. Hence the basin of attraction for their equilibrium becomes larger:

Proposition 11 *If the gains of cooperation for the players 2 $(c_2 - d_2)$ increase, then γ^{cut} decreases, i.e. a lower initial fraction of punishing players is necessary in order to end up in the punishing equilibrium. Furthermore, $\lim_{(c_2-d_2) \to 0} \gamma^{cut} = 1$ and $\lim_{(c_2-d_2) \to \infty} \gamma^{cut} = 0$.*

Again, the intuition is straightforward: the higher the gains of cooperation for a player 2, the higher his profit from being pivotal in inducing cooperation of players 1. Therefore a lower fraction of punishers is necessary in order to make punishing more successful than non-punishing.

Lemma 3 *If the threshold k^{**} of punishing players 2 in a group above which the players 1 start to cooperate increases then γ^{cut} increases, i.e. there are more punishing players necessary in order to end up in the punishing equilibrium.*

Intuitively, a higher threshold k^{**} makes it more probable for an individual to be in a group in which the number of punishers is too low to induce cooperation. In these groups being a punisher is costly. Therefore, fitness of punishers is lower and a higher initial fraction of punishers is necessary to make punishing more successful than non-punishing.

Corollary 3 *If player 1's costs for cooperation $(d_1 - c_1)$ increase, then γ^{cut} increases weakly, i.e. a higher or equal fraction of punishers is necessary in order to end up in the punishing equilibrium.*

Corollary 4 *If player 2's losses due to a punishment p increase, then γ^{cut} decreases weakly.*

1.5.3 Extension: Small Mistakes

This appendix considers the possibility that some players 1 fail to play optimally. Assume that player 1 makes a mistake with small probability ϵ. In that case, he defects even so he should cooperate and vice versa. If ϵ is sufficiently small, results of case 1 and case 2 change only slightly:

Proposition 12 *If players 1 make a mistake with sufficiently small probability ϵ then*

in Case 1 *there exist two stable equilibria: in the first equilibrium the fraction γ_ϵ^{eq} of preferences for rewarding is close to the equilibrium fraction without mistakes γ^{eq}. In the second equilibrium only self-interested preferences survive (i.e. $\gamma = 0$). If ϵ tends to zero then γ_ϵ^{eq} tends to γ^{eq}. Moreover, the basin of attraction of the self-centered equilibrium tends to zero.*

in Case 2 *there remain two stable equilibria. The monomorphic equilibrium where all players have self-interested preferences is still stable. But a monomorphic population of punishers is no longer stable. Instead, there is a second stable equilibrium with a high fraction of punishers and a low fraction of self-interested players. If ϵ tends to zero, the fraction of preferences for punishing in this equilibrium is arbitrarily close to 1.*

We discuss and prove only Case 2. The proof for Case 1 is analogous and written upon request.

The intuition for Case 2 of proposition 12 is straightforward. In a world of no mistakes and in the equilibrium where all players 2 are willing to punish, this threat is costless: player 1 cooperates and no punishment is necessary. But, if players 1 make sometimes mistakes, being a punisher is costly. If almost everybody else is a punisher, the probability of being pivotal tends to zero. But because of mistakes the costs of punishing do not vanish. A monomorphic population of punishers is therefore no longer stable.

For a more formal proof consider average payoffs of both types. A self-interested player 2 receives an average material payoff of

$$\bar{u}_s(\gamma) = (d_2 + \epsilon (c_2 - d_2)) \sum_{k=0}^{k^{**}} B_{N-1,\gamma}(k) + (c_2 - \epsilon (c_2 - d_2)) \sum_{k=k^{**}+1}^{N-1} B_{N-1,\gamma}(k) \quad (1.30)$$

and the punishing type receives

$$\bar{u}_-(\gamma) = (d_2 - c_p + \epsilon (c_2 - d_2 + c_p)) \sum_{k=0}^{k^{**}-1} B_{N-1,\gamma}(k) + (c_2 - \epsilon (c_2 - d_2 + c_p)) \sum_{k=k^{**}}^{N-1} B_{N-1,\gamma}(k). \quad (1.31)$$

Hence the difference in average payoffs between the two types is

$$\begin{aligned} & \bar{u}_-(\gamma) - \bar{u}_s(\gamma) \\ = \; & -(1-\epsilon) c_p \sum_{k=0}^{k^{**}-1} B_{N-1,\gamma}(k) + ((1-2\epsilon)(c_2 - d_2) - \epsilon c_p) B_{N-1,\gamma}(k^{**}) \\ & -\epsilon c_p \sum_{k=k^{**}+1}^{N-1} B_{N-1,\gamma}(k) \\ = \; & (1-2\epsilon) \left(-c_p \sum_{k=0}^{k^{**}-1} B_{N-1,\gamma}(k) + (c_2 - d_2) B_{N-1,\gamma}(k^{**}) \right) - \epsilon c_p. \quad (1.32) \end{aligned}$$

This difference is continuous in γ and ϵ. For $\gamma = 0$, the difference is negative. Hence the monomorphic equilibrium of self-interested preferences remains stable. For $\gamma = 1$ the difference is also negative. Therefore, self-interested preferences can invade a population of punishers. However, for ϵ sufficiently small, there still exists a second stable equilibrium in addition to $\gamma = 0$. In this second equilibrium punishers and self-interested types coexist. The fraction of punishing types in this equilibrium converges to 1 if ϵ tend to 0.

Proof: Existence: for $\epsilon = 0$, there exits a γ_0 where the difference is positive. Due to continuity in ϵ, the difference at this γ_0 is still positive for sufficiently small ϵ. Since the difference is negative at $\gamma = 1$, there must exist a stable equilibrium between γ_0 and 1 due to continuity in γ.

Exactly one more stable equilibrium: the term in large brackets in equation 1.32 is a polynomial of finite order. Hence there are only a finite number of local minima.

Let Δ be the minimum value of all local minima above zero. For $\epsilon < \frac{\Delta}{2\Delta+c_p}$ we obtain $(1-2\epsilon)\Delta - \epsilon c_p > 0$ and therefore all local minima with positive value remain positive. Hence, for sufficiently small ϵ, there are still only two γ for which $\bar{u}_-(\gamma) - \bar{u}_s(\gamma) = 0$ - one (still unstable) equilibrium close to the old unstable equilibrium and one stable equilibrium close to $\gamma = 1$, q.e.d.

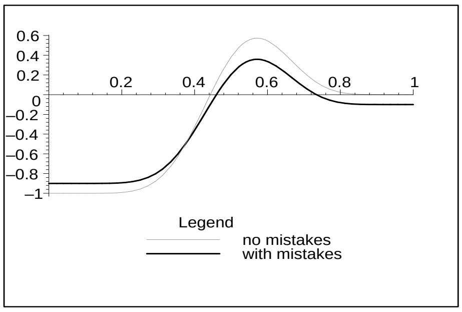


Figure 1.10: Case 2 with mistakes of probability $\epsilon = 0.1$ and with $\gamma^{eq} > \gamma^{cut}$: $(\bar{u}_+ - \bar{u}_s)$ and $(\bar{u}_- - \bar{u}_s)$ as functions of $\gamma \equiv \gamma_+ + \gamma_-$ for $N = 20, d_1 = 1, c_1 = 0, r = 2, d_2 = 5, c_2 = 0, c_r = 1$.

Equation 1.32 is also helpful for analyzing the case of mistake probabilities that are not arbitrarily small but are of moderate size. We can adjust the payoff difference $(\bar{u}_- - \bar{u}_s)$ by re-scaling it slightly with $(1-2\epsilon)$ and then shifting it downwards by ϵc_p. Figure 1.10 demonstrates this for our example of case 2 with a mistake probability of $\epsilon = 0.1$. Here, in the stable "punisher equilibrium" with mistakes, a fraction of $\gamma_- \approx 0.73$ has preferences for punishing, but a fraction of $(1-\gamma_-) \approx 0.26$ has self-interested preferences.

1.5.4 Further Equilibria in Case 3

First, we derive a sufficient condition under which there are no other stable equilibria than those of proposition 8. Second, we analyze the conditions under which there exist stable equilibria consisting only of reciprocal and self-interested preferences and third we give the intuition for why no further stable equilibria exist.

The following lemma limits the possible candidates for stable equilibria.

Lemma 4 *An equilibrium with a positive fraction γ_{rc} of reciprocal players can only be stable if the fraction of γ_s self-interested players is also positive.*

Proof: If $\gamma_s = 0$ then players 1 cooperate in all groups. Therefore, preferences for punishing earn c_1 in all groups, whereas reciprocal preferences earn only $c_1 - c_r$. Hence the equilibrium is not stable, q.e.d.

In case 3 players 1 cooperate in their group if, and only if, $c_1 + \frac{k_+ + k_{rc}}{N} r > d_1 - \frac{k_- + k_{rc}}{N} r$, i.e. $k_+ + k_- + 2k_{rc} > N \frac{d_1 - c_1}{r}$. We define $k_{ef} \equiv k_+ + k_- + 2k_{rc}$, $\tilde{\gamma} \equiv \gamma_+ + \gamma_-$ and $W(k_{ef}) \equiv W_{N-1,\gamma_- + \gamma_+,\gamma_{rc}}(k_{ef})$ as the probability that a group of $N-1$ players has the characteristic k_{ef}, i.e.

$$W(k_{ef}) = \sum_{\substack{\tilde{k},k_{rc}=0 \\ \tilde{k}+2k_{rc}=k_{ef}}}^{N-1} \binom{N-1}{\tilde{k}, k_{rc}} \tilde{\gamma}^{\tilde{k}} \gamma_{rc}^{k_{rc}} \left(1 - \tilde{\gamma} - \gamma_{rc}\right)^{N-1-\tilde{k}-k_{rc}}. \tag{1.33}$$

The probabilities of a group having any characteristic k_{ef} must be 1, i.e.

$$\sum_{k_{ef}=0}^{2N-2} W(k_{ef}) = 1. \tag{1.34}$$

We can write average material payoffs in new notation

$$\bar{u}_{rc} = (d_2 - c_p) \sum_{k_{ef}=0}^{k^*-2} W(k_{ef}) + (c_2 - c_r) \sum_{k_{ef}=k^*-1}^{2N-2} W(k_{ef}) \tag{1.35}$$

$$\bar{u}_+ = d_2 \sum_{k_{ef}=0}^{k^*-1} W(k_{ef}) + (c_2 - c_r) \sum_{k_{ef}=k^*}^{2N-2} W(k_{ef}) \tag{1.36}$$

$$\bar{u}_- = (d_2 - c_p) \sum_{k_{ef}=0}^{k^*-1} W(k_{ef}) + c_2 \sum_{k_{ef}=k^*}^{2N-2} W(k_{ef}) \tag{1.37}$$

$$\bar{u}_s = d_2 \sum_{k_{ef}=0}^{k^*} W(k_{ef}) + c_2 \sum_{k_{ef}=k^*+1}^{2N-2} W(k_{ef}) \tag{1.38}$$

An equilibrium with $\gamma_{rc} > 0$ can only be stable if $\bar{u}_{rc} = \bar{u}_s$ (by lemma 4) and if $\bar{u}_+ \leq \bar{u}_{rc}$ and $\bar{u}_- \leq \bar{u}_{rc}$:

Lemma 5 *The following conditions are all necessary for a stable equilibrium with a* *fraction* $\gamma_{rc} > 0$*:*

1.

$$c_p \sum_{k_{ef}=0}^{k^*-2} W(k_{ef}) + c_r \sum_{k_{ef}=k^*-1}^{2N-2} W(k_{ef}) = (c_2 - d_2)\left(W\left(k^* - 1\right) + W\left(k^*\right)\right) \tag{1.39}$$

2.

$$c_p \sum_{k_{ef}=0}^{k^*-2} W(k_{ef}) \leq (c_2 - d_2 - c_r) W(k^* - 1) \tag{1.40}$$

3.

$$c_r \sum_{k_{ef}=k^*} W(k_{ef}) \leq (c_2 - d_2 - c_r + c_p) W(k^* - 1) \tag{1.41}$$

The next corollary follows directly from condition 1.39:

Corollary 5 *If*

$$\sup_{\substack{\gamma_{rc}\in[0,1]\\(\gamma_+ + \gamma_-)\in[0,1-\gamma_{rc}]}} \left(W(k^* - 1) + W(k^*)\right) < \frac{\min\{c_p, c_r\}}{c_2 - d_2} \tag{1.42}$$

then there is no stable equilibrium with $\gamma_{rc} > 0$*, i.e. the stable equilibria of proposition 8 are the only ones.*

We now analyze the conditions under which there exists a stable equilibrium consisting only of self-interested and reciprocal preferences, i.e. $\gamma_+ = \gamma_- = 0 = \gamma_+ + \gamma_-$. Notice that

$$W_{N-1,0,\gamma_{rc}}(k_{ef}) = \begin{cases} 0 & \textit{if } k_{ef} \textit{ odd} \\ \frac{(N-1)!}{\left(\frac{k_{ef}}{2}\right)!\left(N-1-\frac{k_{ef}}{2}\right)!}\gamma_{rc}^{\left(\frac{k_{ef}}{2}\right)}\left(1-\gamma_{rc}\right)^{N-1-\frac{k_{ef}}{2}} & \textit{if } k_{ef} \textit{ even.} \end{cases} \tag{1.43}$$

It follows directly that, for even k^*, condition 1.40 and condition 1.41 cannot be both fulfilled. Hence for even k^* there is no stable equilibrium consisting only of self-interested and reciprocal preferences.

For odd k^*, on the other hand, things are different. If an equilibrium with a positive fraction of reciprocal and self-interested preferences is stable in the subspace of this two types, it is stable also for invasion of preferences for rewarding or preferences for punishing. This can be seen directly from average payoffs. In such an equilibrium, the probability $W_{N-1,0,\gamma_{rc}}(k_{ef} = k^*)$ of a group with characteristic $k_{ef} = k^*$ is 0, but these are the only groups where preferences for rewarding or preferences for punishing perform better than self-interested ones. To check existence of a stable equilibrium in the subspace of self-interested and reciprocal preferences we just have to consider differences in average material payoffs of this two types, i.e.

$$\bar{u}_{rc}(\gamma_{rc}) - \bar{u}_s(\gamma_{rc}) = \tag{1.44}$$
$$c_p \sum_{k=0}^{k^*-2} B_{N-1,\gamma_{rc}}(k) + (c_2 - d_2 - c_r)\left(B_{N-1,\gamma_{rc}}(k^*-1) + B_{N-1,\gamma_{rc}}(k^*)\right) + c_r \sum_{k=k^*+1}^{N-1} B_{N-1,\gamma_{rc}}(k)$$

For $k^* \geq 2$, this difference is negative for γ_{rc} sufficiently close to zero or one. Hence there exists a stable equilibrium if, and only if,

$$\sup_{\gamma_{rc}\in[0,1]} \left(\bar{u}_{rc}(\gamma_{rc}) - \bar{u}_s(\gamma_{rc})\right) > 0. \tag{1.45}$$

This is summarized in the following

Corollary 6 *Consider case 3:*

a) *For even k^* there is no stable equilibrium consisting of only reciprocal and self-interested preferences.*

b) *For odd k^* and $k^* \geq 2$ there exists a stable equilibrium of only self-interested and reciprocal preferences if, and only if,*

$$\sup_{\gamma_{rc}\in[0,1]} \left(\bar{u}_{rc}(\gamma_{rc}) - \bar{u}_s(\gamma_{rc})\right) > 0. \tag{1.46}$$

Finally, we give the intuition for why there are no stable equilibria with a mixture of three or all four different preference-types in case 3. An equilibrium with positive fractions of self-interested preferences and preferences for punishing is not stable because preferences for punishing become more successful than self-interested ones if a small deviation in their favor occurs. Similarly, an equilibrium with positive fractions of preferences for rewarding and reciprocal preferences is not stable because a small deviation in favor of reciprocal preferences makes reciprocal preferences more successful than preferences for rewarding. Hence, in case 3 maximally two preference types coexist in equilibrium.

1.5.5 Some Definitions from Evolutionary Game Theory

Most standard concepts in evolutionary game theory are formulated for the evolution of strategies. Here we look at the evolution of preferences. The main reason is that this term captures better the aim of this chapter - we want to understand why people might have reciprocal preferences. The following definitions are analogous to their counterparts in evolutionary game theory.

Let $t_1, t_2, \ldots, t_n$ be a finite number of possible preference types and $\gamma_1, \gamma_2, \ldots, \gamma_n$ their fractions of the total population, i.e. $\gamma_1, \gamma_2, \ldots, \gamma_n \geq 0$ and $\sum_{i=1}^{n} \gamma_i = 1$. We call the vector $\gamma \equiv (\gamma_1, \gamma_2, \ldots, \gamma_n)$ a population state. The set of all possible population states is therefore a $n-1$ dimensional simplex in R^n. We call this set Δ. The following definitions concerning selection dynamics is analogous to the ones commonly used in evolutionary game theory (see e.g. Weibull [95]). We focus on continuous selection dynamics defined on the simplex Δ in terms of growth rates $g_i(\gamma)$ for the population shares associated with each preference type $i \in n$ as follows

$$\dot{\gamma_i} = g_i(\gamma)\gamma_i \tag{1.47}$$

where g is a function with open domain X containing Δ.

Definition 3 *A* **regular growth rate function** *is a Lipschitz continuous function $g : X \rightarrow R^n$ with open domain X containing Δ, such that $g(\gamma) \cdot \gamma = 0$ for all $\gamma \in \Delta$*

For any regular growth rate function there exists a unique solution $\xi(t, \gamma^0)$ to equation 1.47 through any initial value γ^0. Moreover ξ is continuous in $t \in T$ and $\gamma^0 \in \Delta$ (Picard-Lindelöf theorem).

Definition 4 *A regular growth rate is called* **payoff monotonic** *if for all $\gamma \in \Delta$*

$$u(t_i, \gamma) < u(t_j, \gamma) \Leftrightarrow g_i(\gamma) < g_j(\gamma), \tag{1.48}$$

where $u(t_i, \gamma)$ stands for the average material payoff of type t_i when the state of the total population is γ.

Hence payoff monotonicity means that the fraction of types receiving higher average material payoffs grow with a higher rate.
To check the stability of a population state, we look at asymptotic stability. We will refer in all proofs to the metric induced by the maximum-norm. The proofs would extend straightforward to other metrics (e.g. the Euclidian-metric).

Definition 5 *A population-state γ is called* **Lyapunov stable** *if every neighborhood B of γ contains a neighborhood B^0 of γ such that $\xi(t, \gamma^0) \in B$ for all $\gamma^0 \in B^0 \cap C$and $t \geq 0$.*
A state $\gamma \in C$ is called **asymptotically stable** *if it is Lyapunov stable and there exists a neighborhood B^* such that $\lim_{t\to\infty} \xi(t, \gamma^0) = \gamma$ holds for all $\gamma^0 \in B^* \cap C$.*

Definition 6 *A* **closed set** *$A \subset C$ is* **Lyaponnov stable** *if every neighborhood B of A contains a neighborhood B^0 of A such that $\gamma_+(B^0 \cap C) \subset B$. A closed set $A \subset C$ is* **asymptotically stable** *if it is Lyapunov stable and if there exists a neighborhood B^* of A such that $\xi(t, x^0)_{t\to\infty} \to A$ for all $x^0 \in B^* \cap C$.*

Chapter 2

The Costs and Benefits of the Separation of Powers*

- an incomplete contracts approach

2.1 Introduction

> *"There can be no liberty where the executive, legislative, and judicial branches are under one person or body of persons because the result is arbitrary despotism (tyranny)"*
>
> Charles-Louis de Secondat, Baron de Montesquieu (1748) "L'Esprit des Lois"

This chapter analyzes the costs and benefits of separating legislature and executive in an incomplete contracts framework: The legislature sets up a decision-making framework that leaves the executive with the residual control rights on the implementation of public projects.

At least since the famous work of Montesquieu [65], the separation of political power into executive, legislature, and judiciary is the most prominent mechanism to protect democratic systems against the abuse of power. The constitutions of most

[*] This chapter is joint work with Kira Börner, University of Munich.

democracies today prescribe such a separation of the political bodies. Yet, while the principle of a separation of powers is unquestioned as an essential ingredient for modern democratic constitutions, the nature and the degree of a separation between executive and legislature remains a contested issue. Is a strict separation of powers always better than a single political body that fulfills both executive and legislative tasks? What are the costs and benefits of separating the task of writing laws from the task of taking a concrete decision in every single case? Is the trade-off different for the local, the national, or the supra-national level?

In order to assess the advantages and the disadvantages of a separation of the executive and the legislature, we propose an incomplete contracts approach. The executive can choose to implement policies. Each policy entails a public good project that may have positive or negative welfare consequences. Under a system of separation of powers, the legislature provides a decision-making framework by writing laws. The executive is left with the residual decision-making rights. Thus, the legislature can constrain the executive by law or empower it to decide in some circumstances. The executive has private interests which may distort the policy choice with respect to the social optimum. Laws written by an independent legislature have the advantage of curbing this abuse of executive power.

It is prohibitively costly, however, to write different laws for every single possibly upcoming project in advance. Laws must be written in general terms and cannot condition on every particular characteristic of each policy project. It is therefore not possible to write a complete contract for the executive's decisions. Laws can be contingent only on some general features which categorize potential projects in general terms. A law prescribing an action affects therefore the entire category and may be suboptimal for some of the affected projects. Prescribing no action, on the other hand, leaves the decision to the executive that is influenced by its private interests. Thus, the legislature faces a trade-off between granting unchecked political power to the better-informed executive and avoiding extreme policy outcomes by a law prescribing the decision.

This definition and formalization of the separation of powers in an incomplete contracts framework proposes an explicit division of tasks of the political bodies. Such a definition of the separation of powers can be found in the political science literature, e.g., Schultz (pp. 191) [81]. Our approach is complementary to the existing formal political economy literature on the separation of powers that does not clearly distinguish between a division of tasks and mechanisms of mutual control of otherwise equivalent political bodies. There are some theoretical models that see constitutions as incomplete contracts: Aghion, Alesina, and Trebbi [1], Aghion and Bolton [2], and Persson, Roland, and Tabellini [70]. Yet, these papers do not focus on the incompleteness of the contract between legislature and executive, but see the whole constitution as an incomplete contract between the sovereign, the citizens, and the political decision-makers. We extensively review the literature in section 2.3.

Why is it precisely the task of writing laws that should be separated from the tasks of implementing policies and making day-to-day, case-dependent decisions? We want to identify the trade-off that is implicit in a separation of the executive and legislative bodies in a political system. For this purpose, we compare the system of separation of powers to a system with a single, unconstrained executive. Whenever the legislature does not maximize social welfare but pursues private interests, also the separation of powers leads to distortions: The laws that control the behavior of the executive are biased in the direction of the legislature's private interests. Depending on the intensity of private interests in executive and legislature, there are cases where a separation of powers is dominated by a political system with a single ruler.

The differences in how strictly executive and legislative bodies are separated becomes particularly clear if we consider different levels of government. On the one hand, we find the strictest separation of powers on the national or supranational levels of government. Examples are the United States, but also the European Union, where we have a clear distinction between the executive tasks of the European Commission as opposed to the legislative powers of the Council (even with the new European Constitution, the European Parliament can hardly be considered as the main legislative body). The European Union is an extreme case also in the sense as it does not directly legitimize its two most important institutions via elections.[1] On the other hand, the farther we go

[1] Tabellini [90] discusses the issue of political accountability of the EU institutions.

down to lower levels of government, we find that a separation of powers receives much less emphasis. Local governments usually put more weight on their executive and on the implementation of specific policy projects.

A separation of powers seems to be especially attractive for national or even supranational levels of government. It serves to avoid extreme decisions by the better informed executive. To achieve this is particularly important when large damage can result from extreme decisions. This is the case on the national level, where far-reaching decisions, as, for example, about peace or war, are made. Even more, decisions on the supranational level are taken unchecked by any higher sovereign power and affect several countries and their constituencies at once. On the local level, in turn, governments are restricted in their action space by the provisions made by the higher levels of government. Thus, extreme decisions have a much smaller scope and affect smaller constituencies. A separation of powers thus seems to be most important where political power is not checked by a higher instance, where large damages can be effected, and where private interests differ strongly.

In most modern democratic constitutions, the separation of powers is complemented by other mechanisms of checks and balances, such as elections. As our focus lies on the separation of powers, we do not include elections in our model. Intuitively, elections provide an incentive scheme that will draw the political decisions towards the social interest. However, they are unable to give the right incentives in case of extreme private interests of the executive. Such extreme decisions can only be prevented by a law, enacted by a separate legislature - at least if the private bias of the legislature is not too large.

The remainder of this chapter is organized as follows: Section 2.2 sets up the incomplete contracts model of a separation of powers and derives our main results. In section 2.3, we review the literature that is related to our work. In section 2.4 we discuss our results and finally conclude.

2.2 The Model

First, we present the basic model in which the political bodies are not restricted by elections. Politicians are randomly chosen from the total population and are assumed to maximize their personal benefits if the constitution gives them the right to decide.

In order to keep the model tractable we model policy choices as simply as possible: Opportunities for public projects arise, which may have a positive or negative expected net social benefit. A policy choice is simply the decision whether to implement such a project or not.[2]

The legislature can only set a general decision-making framework by writing laws that constrain the executive, whereas the executive can condition its decision on the particular characteristics of each single project. We consider this to be the key difference of the tasks of legislature and executive. At the legislation stage, when the laws are written, there are many potentially upcoming projects. Laws can only categorize these projects in very general terms. Thus, each potential project falls into one category. This category can be described by one general feature, that distinguishes all projects of this category from other projects. The expected social benefit of a project conditional on its general category is denoted by $g \in [-G; G]$. For simplicity, g is assumed to be uniformly and independently distributed on $[-G; G]$.

When the executive takes a decision it knows the particular project in question. Thus, the executive can condition its decision on the specific characteristics of each single project - in addition to the general feature g of the project's category. The difference of the social value of the specific project and the expected value for projects of this category is denoted by $s \in [-S; S]$. We assume, again for simplicity, that s is uniformly and independently distributed on $[-S; S]$. The net social benefit of a public project with general feature g and specific characteristic s is thus $(g + s)$.

For example, the broader categories of general features g might be “building streets”,

[2] This way of modelling policy choices captures only the temptation of a politician to distort decisions in his own interest. We do not consider the problem of how far a constitution protects itself against its abolition or a circumvention. Implicitly we assume that there is an independent properly functioning judiciary and/or a civil society that is committed to enforce the constitution as well as constitutional laws.

"providing electricity" or "use of torture". Laws can condition on this broad categories. A law could state: "electricity must be provided", "nobody must be tortured" or "the decision whether to build a street is left to the responsible official of the executive". But laws cannot be conditioned on each particular case. It is, for example, prohibitively costly to write a law for the building of every single street. Therefore, complete contingent contracts among legislature and executive are not possible.

Citizens Preferences We assume an infinite number of citizens $i \in \{1, \ldots, \infty\}$. For each project j each citizen i has a personal individual interest $\lambda_{ij} \in [-\Lambda_j, \Lambda_j]$ which denominates the difference between his private utility from the project and the project's social value $g_j + s_j$. Ex ante, these λ_{ij} are assumed to be uniformly distributed over $[-\Lambda_j, \Lambda_j]$. The personal interest is not known to the citizen at the beginning, but will be revealed at a later stage of the game. Ex ante, the citizens have no information about the social benefit $(g_j + s_j)$ of an potentially upcoming project j. They do, however, know the distributions of all relevant parameters.

For notational simplicity, we neglect the index for the project j in the remainder of the paper. The net utility of a public project with general feature g and specific characteristic s for citizen i is thus

$$U_i = g + s + \lambda_i. \tag{2.1}$$

We normalize the citizen's utility when no project is implemented to $\underline{U}_i = 0$.

Citizen i would like a project to be implemented if and only if

$$g + s + \lambda_i > 0. \tag{2.2}$$

The citizens are the sovereign in the model. In the beginning, they choose whether they would like to set up a constitution with a separation of powers or a system with a single ruler. They then delegate their decision-making power to the government. The government, learning ex ante about the characteristics of the policy projects, has some additional information. This makes the government's choices better than direct decisions by the citizens, as, for example, in a direct democracy. We think of the government as agents who specialize in information gathering and are therefore granted the decision-making rights. Hence, the function of government is to collect the information necessary for the political decision-making process.

Due to the symmetry of the distribution of the λ_i and the infinite number of citizens[3] it is **socially optimal** to realize a public project if and only if

$$g + s \geq 0. \tag{2.3}$$

As the citizens' private interests are independently and uniformly distributed over $[-\Lambda, \Lambda]$, this coincides with the median voter's decision.

The finite size of the support of g, s and λ requires a number of case distinctions. We concentrate on the most interesting cases by making the following

Assumption 6 $\Lambda > 2G; \Lambda > 2S$

Assumption 6 states that for every project there exists somebody who wants to implement it (even when the social value is very low) and somebody who does not want to implement it (even if the social value is very high). Hence, there is always a risk that a politician, who is picked randomly from the distribution of citizens, might take the wrong decision.

The Legislature The legislature writes laws and thereby sets the decision-making framework for the executive. The legislature can therefore only take into account an expected private interest averaged over all potential projects of a general category. This private bias of the legislature at the stage of writing laws is denoted by $\gamma \in \mathbb{R}$.

If this bias of legislature were larger than the range of general features G, it would never be interesting for the constituency to have a separated legislature. We concentrate on the interesting case, by making

Assumption 7 $|\gamma| < G$.

[3] The law of Varadarajan ensures that the empirical distribution of citizens converges weakly to the distribution of the λ_i.

Since the legislature can condition laws only on the general category of a public project, it might be willing to leave certain decisions to the executive. The reason is that the executive can include the specific characteristics of the particular project in its decision. In order to keep the model simple, we assume that there are only three types of laws: For public projects of a certain category a law can

- either prohibit the implementation
- or prescribe the implementation
- or leave the decision to the executive (not write a law).

The Executive The executive has the residual decision-making rights. Whenever there is no law prescribing what to do, it may choose whether or not to implement a project. We call the private interests of the executive λ_{ex}. We assume that the executive is a single person who is chosen randomly from the population. If law does not prescribe a certain action, the executive will implement a project if and only if

$$\lambda_{ex} \geq -(g+s). \tag{2.4}$$

Timing of Events The timing of events is as follows:

- **Period 0:** Citizens choose their constitution behind a "veil of ignorance". They do not know their personal interests at that time. Their only information is the distribution of the g, s, λ_i, and the absolute value of γ. They decide on the basis of this information whether to separate legislature and executive or whether to have a single ruler. That is, in period 0, citizens choose the socially optimal constitution.
- **Period 1:** In this legislative period, the legislature is writing laws - if the constitution is one of separated powers. It may write laws conditioning on the general feature g of possible upcoming public projects and on its private bias γ for this general category. These laws can prescribe that possible upcoming projects of a

certain category have to be implemented and can forbid to implement projects of other categories. The legislature may also write no law for some categories of public projects. With this, it leaves the implementation decision to the executive.

- **Period 2:** Opportunities for public projects arise. The executive observes the general feature g, the specific characteristic s and its personal interest λ_{ex} for a project. If there is a law, the executive has to comply with it. If the decision is left to the executive, it decides whether or not it wants to implement the project following its private interests.

2.2.1 Benchmark Case: The Legislature as a Social Planner

First, we want to consider a benchmark case: What kind of laws would a social planner write in order to maximize social welfare? This is the case when the legislature does not take its personal interests into account at the time of writing the law. The legislature as a social planner then understands that the executive has more detailed information about the concrete projects as it can observe the special characteristics s. However, it also knows that the executive will take its decisions in order to maximize its personal interests.

The expected social benefits of a law on projects in a category with general feature g are:

1. If the law **prohibits** the implementation of projects in this category:

$$W_{proh}(g) = 0 \tag{2.5}$$

2. If the law **enforces** the implementation of projects in this category:

$$W_{enf}(g) = g \tag{2.6}$$

3. If **no law** is written for this category, then the executive will decide in his own interest. It will implement the project if and only if $\lambda_{ex} > -(g + s)$. For the

uniform distribution of λ_{ex}, the probability of implementation is therefore

$$prob\left(\lambda_{ex} > -(g+s)\right) = \frac{1}{2} + \frac{g+s}{2\Lambda} \tag{2.7}$$

The expected social welfare of a project with general feature g if no law is written is[4]

$$W_{noLaw}(g) = \int_{-S}^{S} (g+s) prob\left(\lambda_{ex} > -(g+s)\right) \frac{1}{2S} ds = \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda}, \tag{2.8}$$

for $g + S < \Lambda$. This is guaranteed by assumption 6.

The social planner chooses the contract that maximizes total welfare for each category of public projects.[5] We call these laws the socially optimal laws. This, however, does not mean that the first best is implemented. In our setting of incomplete contracts, the socially optimal laws constitute the welfare-maximizing choice. It is obvious that the first best could be reached if complete contingent contracts were possible. In the first best, a social planner would simply implement all projects with $(g+s) > 0$.

Proposition 13 (The Socially Optimal Law) *Total welfare is maximized by the legislature if projects of a category with general feature g are*

$$\begin{cases} \textit{prohibited} & \textit{if } -G \leq g \leq -\frac{\Lambda}{2} + \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \\ \textit{executive's decision} & \textit{if } -\frac{\Lambda}{2} + \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} < g < \frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \\ \textit{enforced} & \textit{if } \frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \leq g \leq G \end{cases}$$

The proof is in appendix 2.5.

The intuition for this result is simple: If the general feature g has a value close to zero, the social planner has little information whether the project should be implemented or not. The special characteristic s is then decisive. As the private interests of the executive are distributed around $g+s$, it takes the right decision with a high probability. Therefore, the decision should be left to the executive. If on the other

[4] For details of the calculation see appendix 2.5.

[5] The utility of the executive and legislature are negligible for total social welfare: as the population is infinite, a finite number of individuals has a weight of 0.

hand the general feature g has a high absolute value, the legislature has already a good idea whether the project should be implemented or not. If now the executive would like to decide differently, this would reflect more likely its private interests than its more detailed information. Hence, a law should prescribe the action.

Comparative Statics for the Optimal Law How does the optimal law change when the legislature is less well informed about the social value of a project? From proposition 13 we can see how a wider range $[-S, S]$ of the specific project characteristics influences the optimal law:

Corollary 7 *If the range S of specific characteristics increases, then the socially optimal law leaves the decision to the executive for a larger range of g.*

Corollary 7 and corollary 8, below, follow directly from proposition 13. The intuition is straightforward: If the specific characteristics of a project become more important, the social planner is less able to infer the total social benefit of a project. Then the executive, who can take this details into account, is more apt to decide.

Similarly, we can see from proposition 13 how a wider range $[-\Lambda, \Lambda]$ of the private interests of the executive influences the optimal law. Such a wider range of private interests means that there are more citizens with extreme private interests, from whom the executive could be picked.

Corollary 8 *If the range Λ of private interests increases, then the socially optimal law leaves the decision to the executive for a smaller range of g.*

The intuition is that a broader range of private interests makes it more likely that the executives private benefit is not in line with the public benefit of a project. From this, we can the already the benefits of a separation of powers: A separated legislature avoids extreme decisions by the executive. If the executive is more prone to take extreme decisions, the legislature can neutralize that by choosing a law that is more restrictive.

2.2.2 Legislature with Private Interests

Not only the executive, but also the legislature has private interests.[6] In this section, we consider how an independent legislature with private interests may distort laws in its own favor. The legislature can only write laws conditioning on the general project characteristics g. As there are many projects for each general category of projects, the legislature cannot condition the law directly on its private interest λ_{leg}. We thus assume that the legislature conditions the law on its expected value of its private interest, γ, which we call the private bias of the legislature. This bias γ quantifies how the private interests of the legislature differ from the public interest at the stage of writing laws. Consider a public project of a category with general feature g. Then, the expected private benefit of this project for the legislature is $g + \gamma$, as s and λ_{leg} for each project are not known at this stage.

Hence, the legislature's utility of an implemented public project with general feature g, expected private interest of the legislature γ, and special characteristic s is

$$U^{leg}(g, \gamma, s) = g + \gamma + s \tag{2.9}$$

Notice that laws can only be conditioned on g and γ. Now the legislature maximizes its private benefits. When writing a law, the legislature compares:

1. If the law **prohibits** the implementation of projects in this category, the expected utility of the law is:
$$EU^{leg}_{proh}(g, \gamma) = 0 \tag{2.10}$$

2. If the law **enforces** the implementation of projects in this category, the expected utility of the law is:
$$EU^{leg}_{enf}(g, \gamma) = g + \gamma \tag{2.11}$$

3. If **no law** is written for this category, the executive decides in its own interest

[6] These private interests of the legislature may arise, e.g., due to party ideologies or due to an imperfect representations of the citizens in the legislative body. Most constitutions take measures, such as a large parliament of representatively chosen deputies to keep the bias of legislature moderate.

and implements the project if and only if $\lambda_{ex} > -(g+s)$. The probability of implementation remains $prob\left(\lambda_{ex} > -\left(g+s\right)\right) = \frac{1}{2} + \frac{g+s}{2\Lambda}$, as given in equation 2.7. The expected private benefit for legislature if no law is written is therefore (under assumptions 6 and 7)[7]

$$EU^{leg}_{noLaw}(g,\gamma) = \int_{-S}^{S} (g+s+\gamma) prob\left(\lambda_{ex} > -\left(g+s\right)\right) \frac{1}{2S} ds = \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda} + \left(1+\frac{g}{\Lambda}\right)\frac{\gamma}{2}. \tag{2.12}$$

For each project category, the legislature writes the law which maximizes its private benefits:

Proposition 14 (Law chosen by legislature) *A law maximizes the legislature's expected benefits if projects of a category with general feature g and expected private interest of the legislature γ are*

$$\begin{cases} \textit{prohibited} & \textit{if } -G \leq g \leq -\frac{\Lambda+\gamma}{2} + \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} \\ \textit{executive's decision} & \textit{if } -\frac{\Lambda+\gamma}{2} + \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} < g < \frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} \\ \textit{enforced} & \textit{if } \frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} \leq g \leq G. \end{cases}$$

The proof is in the appendix 2.5. In order to get a better intuition on how the bias γ of the legislature distorts the law, we use the Taylor approximation around $\gamma = 0$ for the lower threshold $g^l_{\Lambda,S}(\gamma) \equiv -\frac{\Lambda+\gamma}{2} + \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}$ and the upper threshold $g^u_{\Lambda,S}(\gamma) \equiv \frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}$:

$$\begin{aligned} g^l_{\Lambda,S}(\gamma) &\approx g^l_{\Lambda,S}(0) - \frac{1}{2}\left(1 + \frac{\frac{\Lambda}{2}}{\sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}}\right)\gamma - \frac{\frac{S^2}{3}}{8\left(\sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}\right)^3}\gamma^2 + \ldots \\ g^u_{\Lambda,S}(\gamma) &\approx g^u_{\Lambda,S}(0) - \frac{1}{2}\left(1 + \frac{\frac{\Lambda}{2}}{\sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}}\right)\gamma + \frac{\frac{S^2}{3}}{8\left(\sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}\right)^3}\gamma^2 + \ldots \end{aligned} \tag{2.13}$$

In general, a Taylor approximation is good only for sufficiently small γ. Yet, we show in appendix 2.5 that the intuition described below holds for all γ.

[7] For details of the calculation see appendix 2.5.

In first order both thresholds are shifted by the same amount. The shift is directed opposite to the sign of γ and the absolute value of the shift is larger than γ. For intuition consider a $\gamma > 0$: The legislature is more likely to prefer the implementation of a project than the social planner. Hence, it enforces projects of categories which the social planner would leave to the decision of the executive. It also leaves projects to the executive's decision that the social planner would prohibit. This means that both thresholds are shifted towards more negative values for positive γ. In fact, they are shifted by more than $-\gamma$. The reason is that the legislature knows that the executive has private interests which are on average more centered. The executive therefore has a tendency to counterbalance the bias of the legislature. The legislature shifts the project categories in order to neutralize this effect.

From proposition 14 we can directly derive the comparative statics for the law chosen by a biased legislature, analogously to corollary 7 and 8:

Corollary 9 *If the range Λ of private interests of the executive increases then the law chosen by legislature leaves the decision to the executive for a smaller range of g. Furthermore, if Λ goes to infinity then the range of g for which the executive decides becomes an arbitrarily small interval around the general feature $g = -\gamma$.*

The proof is in appendix 2.5. The intuition is similar to the one in corollary 8. Yet, the range for which the executive decides is now centered around $-\gamma$, due to the legislature's bias.

2.2.3 The Optimal Constitution

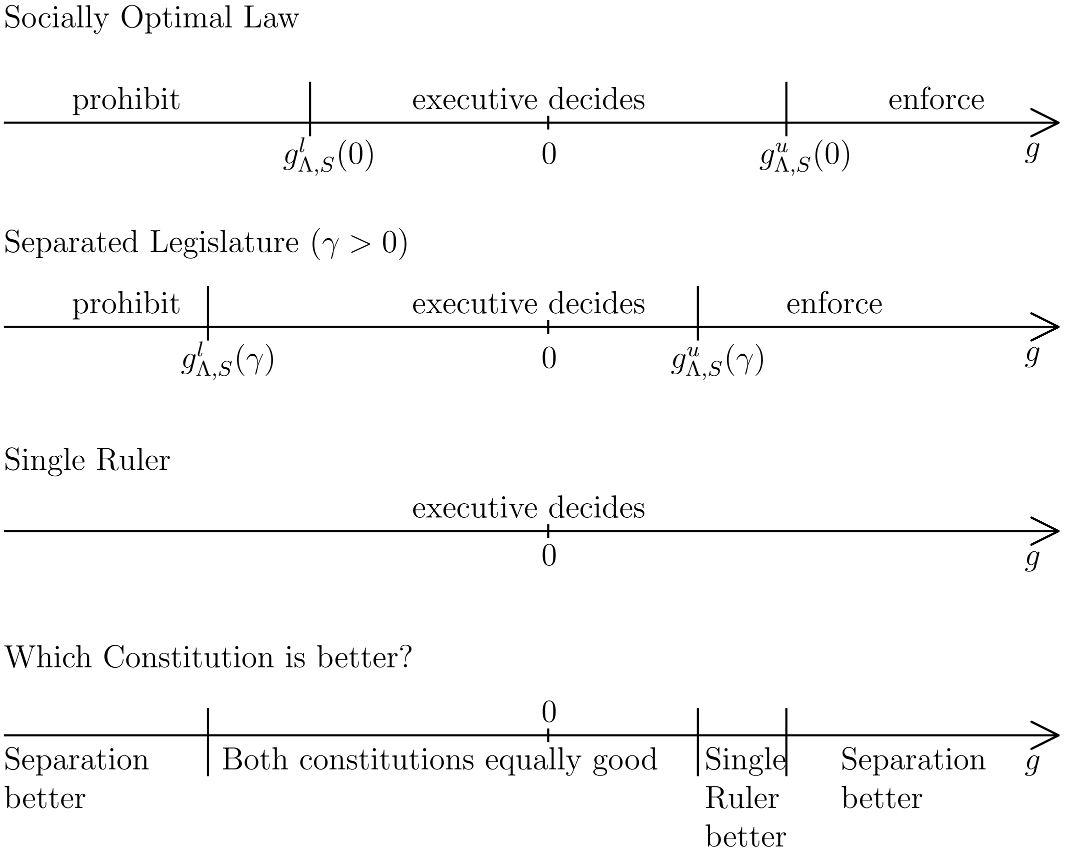
In period 0, citizens choose their constitution behind the "veil of ignorance". At this stage, they know neither their later position nor their private interests. They know only the structure of the game and the probability distributions of all relevant variables.[8] Citizens can separate the legislature from the executive (then the resulting laws are described by proposition 14) or install a "single ruler". Such a single ruler is

[8] This concept of a veil that takes away from citizens the possibility to decide on the basis of their private interests, and makes them decide in the social interest, goes back to Rawls [74] and is widely used in political philosophy and in formal political economy.

an executive that is not constrained and will not constrain himself by laws but decides each single case according to his private interests.[9]

Citizens choose to separate the legislature from the executive if and only if this maximizes their expected social welfare. For given distributions of s, λ_{ex}, and g and a given parameter γ, we can calculate expected social welfare under both constitutions and compare them. But, even for the uniform distribution, calculations become quite cumbersome. Therefore, we relegate the formal analysis to the appendix 2.5 and analyze the important features referring to figure 2.1. Notice that the regime of separation of powers as well as a system with a single ruler may be optimal under certain circumstances. Which constitution is better depends crucially on the range $[-G, G]$ of possible general project characteristics. We can state

Figure 2.1: Choice of the Constitution



Socially Optimal Law

Separated Legislature ($\gamma > 0$)

Single Ruler

Which Constitution is better?

[9] Even a single ruler might enact some laws in order to regulate inter-citizen relationships or in order to stimulate investment of citizens by committing himself not to expropriate their gains ex post. These kind of laws are not the focus of this paper.

Proposition 15 *Separation of powers is not always optimal. The range of general project features $[-G, G]$ influences the constitutional choice. Consider $\gamma > 0$. Then*

- *for $G < g^u_{\Lambda,S}(\gamma)$ both constitutions do equally well.*
- *for $g^u_{\Lambda,S}(\gamma) < G < g^u_{\Lambda,S}(0)$ the single ruler is more attractive.*
- *for $G > g^u_{\Lambda,S}(0)$ separation of powers becomes more attractive with an increasing G.*

This result follows directly from figure 2.1. Intuitively, an independent legislature, that is, a regime of separation of powers, is particularly valuable if there are a number of categories where the general feature gives already a clear signal whether it contains beneficial projects or not. Then, the legislature can get quite reliable information about the nature of a project. When the range of possible general project characteristics is smaller, this signal is more likely to be dominated by the special project characteristics. Therefore, the relative attractiveness of a system with a single ruler increases. For an intermediate range of G a constitution with a single ruler is socially optimal. This results from the bias of the legislature which distorts the law under a system of separation of powers. This distortion is felt at the thresholds of the law (for an intermediate range of G).

Which constitution is better also depends on the absolute value of the bias of the legislature γ. Separation of powers is optimal for instance when γ is close to zero (and if G is not too small). Then, the laws under the separation of powers are arbitrarily close to the socially optimal law and dominate the outcome under a single ruler.

Proposition 16 *The expected welfare under a single ruler increases relative to the expected welfare under separation of powers if the absolute value of the bias of the legislature γ becomes larger. For $\gamma = 0$ separation of powers is always optimal, and for $|\gamma| \to G$ a system with a single ruler yields a higher expected welfare.*

The proof is in appendix 2.5. For an intuitive explanation, please refer to proposition 14 and to figure 2.1. An increase in the absolute value of γ shifts the thresholds $g^l_{\Lambda,S}(\gamma)$ and

$g^u_{\Lambda,S}(\gamma)$ away from their socially optimal values. The larger the bias of the legislature, the more is the law distorted with respect to the social optimal law. That is, the range of general project characteristics g where, for $\gamma > 0$, projects are enforced which should be left for the executive to decide (or, for $\gamma < 0$, projects are prohibited which the executive should decide), becomes larger. This makes the regime of a single ruler more attractive.

The constitutional choice is also influenced by the range $[-\Lambda, \Lambda]$ of possible private interests of the executive and the range $[-S, S]$ of special project characteristics:

Proposition 17 *If the range $[-\Lambda, \Lambda]$ of private interests of the executive is sufficiently large and/or the range $[-S, S]$ of special project characteristics is sufficiently small separation of powers leads to higher expected welfare in comparison with a single ruler.*

The proof is in appendix 2.5. Intuitively, for large Λ, the private interests of the executive dominate its decision and its private information about s does hardly influence its choice. Similarly, a small s means that the executive has few additional information on the projects value. If the executive would decide differently from the legislature, the reason is most likely its private interests and not its better information on s.

Thus, a regime of separation of powers is not always optimal. In particular, a system with a single ruler becomes more attractive if the bias of the legislature is large, general project characteristics G are not too widely spread and special project characteristics S are spread in a large range. Also, the single ruler becomes a better constitutional choice if the range of possible private interests of the executive is small.

2.3 Related Literature

The literature on the principle of separation of powers and the optimal constitutional choice in political science and political philosophy, following the tradition of Montesquieu, is extensive. A classic are the Federalist Articles by Hamilton, Madison and Jay [43]. They prepared, and argued for, the constitution of the United States as a federal state with a strong central government.[10] A recent overview of the research investigating the implementation of the principle of separation of powers in the US political practice is Schramm and Wilson [80]. As our model presents a formal political economy approach to the question of separation of powers, we do not discuss this political science literature further.[11]

Without referring directly to the issue of the separation of powers, Rawls [74] created a theoretical framework to evaluate normatively the choice of political institutions. Individuals choose the basic principles and rules for their society behind a veil of ignorance, i.e., without being informed about their individual endowment and capabilities. This idea has become the basis for the more recent formal models of the optimal endogenous constitutional choice.

While there is a vast formal political economy literature that treats political institutions as exogenous constraints for policy-making, the strand of research that seeks to explain the choice or the emergence of political institutions is both more recent and much smaller. The paper that most explicitly focusses on the effects of a separation of powers on political accountability is Persson, Roland, and Tabellini [70]. The constitution is interpreted as an incomplete contract between the voters and the political decision-makers. Thus, the incentive schemes that voters offer to the politicians, e.g., by elections, can only be implicit. Due to informational asymmetries and their decision-making power, politicians are able to appropriate rents from holding office. The authors demonstrate that having two selfish decision-makers may be worse than having a single one due to a common pool problem.[12] With separation of powers, or rather, with the institutionalization of two political decision-makers, the right timing and a clear accountability of decisions to the two political bodies as well as a require-

[10] An interesting discussion of the relevance of the Federalists's ideas for the contemporary study of political institutions is given in Grofman and Wittman [39].

[11] A related literature is the one on the delegation of decision-making power from the legislature to the executive. In this context, Volden [94], building on Epstein and O'Halloran [23] models the effects of a separation of power on bureaucratic discretion, assuming that the executive receives the power to veto legislative decisions. Buchanan [16], Mueller [66], and Voigt [92] and the collection of essays by Voigt [93] give an overview over classic and recent papers on the more general question of constitutional design.

[12] In an earlier related paper, Brennan and Hamlin [15] argue that a separation of powers is harmful for the voters as it introduces an externality between the political bodies that is not internalized.

ment of consent, where no political body can independently claim the use of government resources, may solve the common pool problem and successfully curb rent extraction by the policy-makers. The constitution then specifies the rules of interaction between the two decision-makers or political bodies. Yet, while the authors discuss that, e.g., the budget authority may be granted to only one of the two political bodies, they do not model a truly functional separation of powers in legislature and executive. Rather, they see the separation of powers as a prerequisite for the institutionalization of checks and balances, that is, of mutual control mechanisms of the executive and the legislature. For this argument, it is not necessary that the two powers fulfill different tasks. The important feature instead is that both political bodies have decision-making rights and control each other.

Also Laffont (chapter 3) [53] interprets constitutions as contracts among the voters and the politicians. He focusses on the design of the incentive structure within governmental institutions. In his view, checks and balances improve efficiency and should be able to abolish politicians' incentives for collusion. Furthermore, he shows that information extraction may be easier if there is more then one informed agent. Yet, the focus in his work also lies on checks and balances. The separation of powers is seen only as a means of creating several bodies of government that can be used to control each other.

Most of the rest of the formal political economy literature on endogenous constitutional choice has a different focus. Rather than analyzing a separation of powers within government, they look at the relation of the sovereign and the political decision-makers. The constitution is seen as an incomplete contract between the sovereign, i.e., the people, and the policy makers. In most papers of this kind, the political institutions within the government are not modelled in detail. Building on the work of Romer and Rosenthal [78], and Laffont [54], Aghion and Bolton [2] analyze the endogenous choice of decision-making procedures. They argue that if complete social contracts, i.e., constitutions that specify policy outcomes for all possible states of the world, are impossible, majority rules are preferred over unanimity as the future decision-making rule. Then, the political decision-making process consists of agenda-setting by the majority coalition and an eventual compensation to make citizens who loose from the policy join the majority coalition.[13] The optimal majority rule is derived from a trade-off between minimizing the scope of expropriation by the majority and the costs of compensating vested interests. In an extended version of the model by Aghion and Bolton [2], Erlenmeier and Gersbach [24] introduce flexible majority rules that depend on the scope of a proposed policy, e.g., by requiring majorities both in the group of affected taxpayers and in the group of people that do not pay taxes under a new policy. This rules can in many cases implement the social optimum. Gersbach [35] further extends this approach.

A recent paper by Aghion, Alesina, and Trebbi [1] focusses on the optimal constitutional choice of a minimum blocking minority and derives a similar basic trade-off for the delegation of unchecked power to the political decision-makers.[14] The abuse of power is the more probable, the more unchecked power is transferred to a political leader. On the other hand, too many checks and balances make necessary reforms unlikely, as these can then easily be blocked by a minority of citizens. Also in this model, the constitution is an incomplete contract between the citizens and the political decision-makers. The political decision-maker is controlled directly by the constituency. There is no further separation of governing tasks within the government. Thus, the paper focusses exclusively on the majority needed to implement a policy as an instrument to control political power.

The approach of our paper clearly differs from this formal literature on checks and balances by focussing on the functional division of tasks between legislature and executive. This view is in line with the non-formal political science literature that acknowledges the functional division between the legislature and the executive and considers the separation of powers as qualitatively different from other institutions of checks and balances (see, e.g., Schulz [81]). In our model, the interaction between legislature and executive is governed by incomplete contracts where the executive retains the residual decision-making rights. The legislature in our model sets out a decision-

[13] Standard models of agenda-setting are, e.g., Romer and Rosenthal [77], or Baron and Ferejohn [5]

[14] In a paper by Messner and Polborn [64] voters choose also endogenously such a super-majority rule in the context of overlapping generations.

making framework which prescribes some actions to the executive but also empowers it to take own decisions. We argue that such a separation of tasks goes hand in hand with the informational endowment of the two political bodies. As laws are to be used in several instances, the legislature, at the stage of lawmaking, is not informed about the characteristics of all special cases to which the law might apply. By explicitly assuming different tasks for executive and legislative bodies, we can derive the trade-offs that are present in the institution of a separation of powers on the one hand and a system with an elected single ruler on the other hand. Compared to the recent formal literature, we thus concentrate on a specific, but important, question: Why is a functional separation of powers installed in most constitutions and what are its costs and benefits?

2.4 Conclusion

We analyze costs and benefits of separating legislature and executive in an incomplete contracts model. Laws - the task of the legislature - set only a framework for decision-making. The executive takes the residual decisions. Laws have the advantage of curbing the abuse of executive power due to private special interests. But laws can not condition on the specific characteristics of each single project. In particular when also the legislature pursues its private interests, laws may restrict the executive for the wrong project categories.

According to our model, it is not always optimal to have a separation of powers. However, a system with a legislature separated from the executive is important under three circumstances: First, if private interests of the executive can strongly deviate from the public interest. Secondly, if the bias of legislature at the stage of writing laws tends to be moderate. Finally, if for some general categories of projects the public interest is clearly visible ex ante and the details of each single project are of minor importance for the social value of these projects.

Our results can be used to explain the constitutional design of different levels of government: On the national level, where it is difficult for citizens to assess the social benefits of a project (for example national security), separation of powers is optimal.

On lower levels of government, where the welfare implications of a single project (such as a local road) are more clearly visible to the voters, a single ruler may be the optimal choice of constitution. This result can thus explain the observation that low levels of government often show a combination of executive and legislative powers in only one body of government.

When the distortions of laws by a biased legislature are kept low, this strengthens the case for a regime of separation of powers. A number of constitutional designs try to take account of this: The legislature tends to be a parliament, representing most groups of the population. Furthermore, the requirement of public debate in parliament could be seen to enhance the abilities of voters to monitor the process of lawmaking. Most importantly, the general character of laws makes it more difficult for the legislature to hide private interests from the voters. Of course, there are many ways by which a legislature can be influenced, a prominent one being lobbying. When lobbies are supporting very extreme interests, this might again reinforce the bias of the legislature. However, it seems plausible that a single ruler and the executive in general will at least be as sensitive to lobbying as the legislature.[15]

The paper can also contribute some thoughts to the discussion about a constitution for the European Union: As a supranational European government would be very hard to monitor for citizens, the model would argue for a strict separation of powers. An additional problem in the case of the EU is that the most important political bodies are not directly elected. Members of the Council are only indirectly elected, being the governments of the member states. Members of the Commission are hardly disciplined by any electoral process. For this constellation, our model suggests that a separation of powers is even more important, as, except for the European Parliament, the EU lacks direct legitimation by elections. Of course, this is more a way of organized thinking about the issue than a concrete recommendation. The model has to neglect many aspects of the new constitution, most importantly, that it represents a federal system.

[15] The relative sensitivity of different political powers to lobbying is an interesting issue for further research.

2.5 Appendix

Proof of Proposition 13

First, we calculate the probability that an executive decides to implement a project with general feature g and special feature s. Remember that λ_{ex} is uniformly distributed on $[-\Lambda, \Lambda]$ and that $-\Lambda < -(g+s) < \Lambda$ due to assumption 6. Hence,

$$\begin{aligned} prob\left(\lambda_{ex} > -\left(g+s\right)\right) &= \int_{-(g+s)}^{\Lambda} \frac{1}{2\Lambda}\, d\lambda_{ex} \\ &= \frac{1}{2} + \frac{g+s}{2\Lambda}. \end{aligned} \tag{2.14}$$

Second, we calculate the social welfare for a project with general feature g if the decision is left to the executive. The probability that this project has a special feature between s and $s+ds$ is $\frac{ds}{2S}$. Such a project is implemented by the executive with probability $prob\left(\lambda_{ex} > -\left(g+s\right)\right)$ and then yields a social benefit of $(g+s)$. With probability $1 - prob\left(\lambda_{ex} > -\left(g+s\right)\right)$ such a project is not implemented. Then the social benefit is 0. Integrating over all s we obtain for $g+S < \Lambda$:

$$\begin{aligned} W_{noLaw}(g) &= \int_{-S}^{S} (g+s) prob\left(\lambda_{ex} > -\left(g+s\right)\right) \frac{1}{2S} ds \\ &= \int_{-S}^{S} (g+s)\left(\frac{1}{2} + \frac{g+s}{2\Lambda}\right)\frac{1}{2S} ds \\ &= \frac{1}{2S}\int_{-S+g}^{S+g} (s')\left(\frac{1}{2} + \frac{s'}{2\Lambda}\right) ds' = \frac{1}{2S}\int_{-S+g}^{S+g} \left(\frac{s'}{2} + \frac{(s')^2}{2\Lambda}\right) ds' \\ &= \frac{1}{2S}\left(\left[\frac{s^2}{4}\right]_{-S+g}^{S+g} + \left[\frac{s^3}{6\Lambda}\right]_{-S+g}^{S+g}\right) \\ &= \frac{(S+g)^2 - (-S+g)^2}{8S} + \frac{(S+g)^3 - (-S+g)^3}{12\Lambda S} \\ &= \frac{4Sg}{8S} + \frac{2S^3 + 6Sg^2}{12S\Lambda} = \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda}. \end{aligned} \tag{2.15}$$

Third, the social planner chooses the law with the highest expected social welfare. For a general feature $g > 0$, enforcing a project yields a higher expected social benefit than prohibiting the project. Hence, the social planner compares ”enforcing” projects of this category with "writing no law". The decision is left to the executive if and only if

$$\begin{aligned} W_{noLaw}(g) &> W_{enf}(g) \\ \Leftrightarrow \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda} &> g \\ \Leftrightarrow g^2 - 2\frac{\Lambda}{2}g + \left(\frac{\Lambda}{2}\right)^2 - \left(\frac{\Lambda}{2}\right)^2 + \frac{S^2}{3} &> 0 \\ \Leftrightarrow \left(g - \frac{\Lambda}{2}\right)^2 &> \left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3} \\ \Leftrightarrow \left|g - \frac{\Lambda}{2}\right| &> \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \\ \Leftrightarrow g < \frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} & \quad or \quad g > \frac{\Lambda}{2} + \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}. \end{aligned} \tag{2.16}$$

Due to assumption 6 the second case is not relevant. Hence, the socially optimal law should leave the decision to the executive if $0 < g < \frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}$ and enforce the project if $\frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \leq g \leq G$.

Similarly, for a general feature $g \leq 0$ the social planner compares the benefits of prohibiting such projects with the expected benefits of leaving the decision to the executive. He will leave the decision to the executive if and only if

$$\begin{aligned} W_{noLaw}(g) &> W_{proh}(g) \\ \Leftrightarrow \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda} &> 0 \\ \Leftrightarrow \left|g - \left(-\frac{\Lambda}{2}\right)\right| &> \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} \\ \Leftrightarrow g < -\frac{\Lambda}{2} - \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}} & \quad or \quad g > -\frac{\Lambda}{2} + \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}. \end{aligned} \tag{2.17}$$

Due to assumption 6 the first case is not relevant. Hence, the socially optimal law should leave the decision to the executive if $0 \geq g > -\frac{\Lambda}{2} + \sqrt{\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}}$ and prohibit the project if $-\frac{\Lambda}{2}+\sqrt{\left(\frac{\Lambda}{2}\right)^2-\frac{S^2}{3}} \geq g \geq -G$,

q.e.d.

Proof of Proposition 14

We calculate the expected benefits for the legislature if the decision on a project of a category with general feature g and private interests for the legislature γ is left to the executive:

$$\begin{aligned}
EU^{leg}_{noLaw}(g,\gamma) &= \int_{-S}^{S}(g+s+\gamma)prob\left(\lambda_{ex}>-\left(g+s\right)\right)\frac{1}{2S}ds \\
&= \int_{-S}^{S}(g+s+\gamma)\left(\frac{1}{2}+\frac{g+s}{2\Lambda}\right)\frac{1}{2S}ds \\
&= \frac{1}{2S}\int_{-S+g}^{S+g}(s'+\gamma)\left(\frac{1}{2}+\frac{s'}{2\Lambda}\right)ds' \\
&= \frac{1}{4S}\int_{-S+g}^{S+g}\left(s'\left(1+\frac{\gamma}{\Lambda}\right)+\frac{(s')^2}{\Lambda}+\gamma\right)ds' \\
&= \frac{1}{4S}\left(\left(1+\frac{\gamma}{\Lambda}\right)\left[\frac{s^2}{2}\right]_{-S+g}^{S+g}+\left[\frac{s^3}{3\Lambda}\right]_{-S+g}^{S+g}+\gamma\left[s\right]_{-S+g}^{S+g}\right) \\
&= \frac{g^2}{2\Lambda}+\frac{g}{2}+\frac{S^2}{6\Lambda}+\frac{\gamma}{2}\left(1+\frac{g}{\Lambda}\right). \qquad (2.18)
\end{aligned}$$

The legislature chooses the law which yields it the highest expected private benefits. For $g+\gamma>0$, enforcing a project yields a higher expected social benefit than prohibiting the project. Hence, the legislature compares ”enforcing” projects of this category with ”writing no law”. It leaves the decision to the executive if and only if

$$\begin{aligned}
EU^{leg}_{noLaw}(g,\gamma) &> EU^{leg}_{enf}(g,\gamma) \\
\Leftrightarrow \frac{g^2}{2\Lambda}+\frac{g}{2}+\frac{S^2}{6\Lambda}+\frac{\gamma}{2}\left(1+\frac{g}{\Lambda}\right) &> g+\gamma \\
\Leftrightarrow g^2+2\frac{\gamma-\Lambda}{2}g+\left(\frac{\gamma-\Lambda}{2}\right)^2 &> \left(\frac{\gamma-\Lambda}{2}\right)^2+\Lambda\gamma-\frac{S^2}{3} \\
\Leftrightarrow \left(g-\frac{\Lambda-\gamma}{2}\right)^2 &> \left(\frac{\Lambda+\gamma}{2}\right)^2-\frac{S^2}{3}. \qquad (2.19)
\end{aligned}$$

For $\left(\frac{\Lambda+\gamma}{2}\right)^2 < \frac{S^2}{3} \Leftrightarrow -\Lambda - \frac{2}{\sqrt{3}}S < \gamma < -\Lambda + \frac{2}{\sqrt{3}}S$ this inequality would hold for all g. But due to assumptions 6 and 7 we have $\gamma > -\left(1 - \frac{1}{\sqrt{3}}\right)\Lambda > -\Lambda + \frac{2}{\sqrt{3}}S$, and thus this case is excluded. Hence, the legislature leaves the decision to the executive only if

$$\begin{aligned} \Leftrightarrow \left|g - \frac{\Lambda-\gamma}{2}\right| &> \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} \\ \Leftrightarrow g < \frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} \quad &or \quad g > \frac{\Lambda-\gamma}{2} + \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}. \end{aligned} \tag{2.20}$$

The second case is impossible. To show this consider the function $f(\Lambda,\gamma,S) \equiv \frac{\Lambda-\gamma}{2} + \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}$. This function is decreasing in S and increasing in γ (since, $\frac{\partial f}{\partial \gamma} = -\frac{1}{2} + \frac{1}{2}\frac{\left(\frac{\Lambda+\gamma}{2}\right)}{\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}} > 0$). Hence, $f(\Lambda,\gamma,S) > f(\Lambda,\gamma = -0.4\Lambda, S = 0.5\Lambda) > \frac{\Lambda-(-0.4\Lambda)}{2} > 0.5\Lambda > G$. This contradicts the second case. Hence, we obtain for $g+\gamma > 0$ that the legislature leaves the decision to the executive if

$$-\gamma < g < \frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}. \tag{2.21}$$

For $g+\gamma < 0$, enforcing a project yields a lower expected private benefit for the legislature than prohibiting the project. Hence, the legislature compares "prohibiting" projects of this category with "writing no law". It leaves the decision to the executive if and only if

$$\begin{aligned} EU^{leg}_{noLaw}(g,\gamma) &> EU^{leg}_{proh}(g,\gamma) \\ \Leftrightarrow \frac{g^2}{2\Lambda} + \frac{g}{2} + \frac{S^2}{6\Lambda} + \frac{\gamma}{2}\left(1 + \frac{g}{\Lambda}\right) &> 0 \\ \Leftrightarrow g^2 + 2\frac{\gamma+\Lambda}{2}g + \left(\frac{\gamma+\Lambda}{2}\right)^2 &> \left(\frac{\gamma+\Lambda}{2}\right)^2 - \Lambda\gamma - \frac{S^2}{3} \\ \Leftrightarrow \left(g + \frac{\Lambda+\gamma}{2}\right)^2 &> \left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}. \end{aligned} \tag{2.22}$$

For $\left(\frac{\Lambda-\gamma}{2}\right)^2 < \frac{S^2}{3} \Leftrightarrow \Lambda - \frac{2}{\sqrt{3}}S < \gamma < \Lambda + \frac{2}{\sqrt{3}}S$ this inequality would hold for all g. Yet, due to assumptions 6 and 7, we have $\gamma < \left(1 - \frac{1}{\sqrt{3}}\right)\Lambda < \Lambda - \frac{2}{\sqrt{3}}S$, and hence this case is excluded. Hence, the legislature leaves the decision to the executive only if

$$\begin{aligned} \Leftrightarrow \left| g - \left(-\frac{\Lambda + \gamma}{2} \right) \right| \quad &> \quad \sqrt{\left(\frac{\Lambda - \gamma}{2} \right)^2 - \frac{S^2}{3}} \\ \Leftrightarrow g < -\frac{\Lambda + \gamma}{2} - \sqrt{\left(\frac{\Lambda - \gamma}{2} \right)^2 - \frac{S^2}{3}} \quad &or \quad g > -\frac{\Lambda + \gamma}{2} + \sqrt{\left(\frac{\Lambda - \gamma}{2} \right)^2 - \frac{S^2}{3}}. \end{aligned} \tag{2.23}$$

Now the first case is impossible.The function $h(\Lambda, \gamma, S) \equiv -\left(\frac{\Lambda+\gamma}{2} - \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} \right)$ is monotonically increasing in S and γ (since $\frac{\partial h}{\partial \gamma} = -\frac{1}{2} + \frac{1}{2} \frac{\left(\frac{\Lambda-\gamma}{2}\right)}{\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}} > 0$). Hence, $h(\Lambda, \gamma, S) < h(\Lambda, \gamma = 0.4\Lambda, S) < -\frac{\Lambda+0.4\Lambda}{2} < -0.5\Lambda < -G$, what contradicts the first case. Hence, for $\gamma + g < 0$ only the second case remains and the legislature leaves the decision to the executive if

$$-\gamma > g > -\frac{\Lambda + \gamma}{2} + \sqrt{\left(\frac{\Lambda - \gamma}{2} \right)^2 - \frac{S^2}{3}}, \tag{2.24}$$

q.e.d.

Derivation of the Comparative Statics on $g^l_{\Lambda,S}(\gamma)$ and $g^u_{\Lambda,S}(\gamma)$

Proof of the Generality of the Intuition suggested by the Taylor Approximation

From the Taylor Approximation we derived intuitively that an legislative bias of γ shifts the law chosen by legislature compared to the optimal law by more than γ. More precise the lower as well as the upper threshold are shifted by more than $-\gamma$. In fact, this is true for all γ as

$$\begin{aligned} \left(g^l_{\Lambda,S}(\gamma)\right)' \quad &= \quad \left(-\frac{\Lambda + \gamma}{2} + \sqrt{\left(\frac{\Lambda - \gamma}{2} \right)^2 - \frac{S^2}{3}} \right)' = -\frac{1}{2} \left(1 + \frac{\left(\frac{\Lambda-\gamma}{2}\right)}{\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}} \right) < -1, \\ \left(g^u_{\Lambda,S}(\gamma)\right)' \quad &= \quad \left(\frac{\Lambda - \gamma}{2} - \sqrt{\left(\frac{\Lambda + \gamma}{2} \right)^2 - \frac{S^2}{3}} \right)' = -\frac{1}{2} \left(1 + \frac{\left(\frac{\Lambda+\gamma}{2}\right)}{\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}} \right) < -1, \end{aligned}$$

and the thresholds of the optimal law are identical to $\gamma = 0$. Furthermore, the distance between both thresholds increases with γ, since for all γ

$$\left(g^{u}_{\Lambda,S}\left(\gamma\right) - g^{l}_{\Lambda,S}\left(\gamma\right)\right)' = \frac{1}{2}\left(\frac{\left(\frac{\Lambda-\gamma}{2}\right)}{\sqrt{\left(\left(\frac{\Lambda-\gamma}{2}\right)\right)^{2} - \frac{S}{3}}} - \frac{\left(\frac{\Lambda+\gamma}{2}\right)}{\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^{2} - \frac{S^2}{3}}}\right) > 0. \tag{2.25}$$

The las inequality follows from the fact that $\left(\frac{\Lambda-\gamma}{2}\right) < \left(\frac{\Lambda+\gamma}{2}\right)$ and the observation that the function $\tilde{f}(x) \equiv \frac{x}{\sqrt{x^2-c}}$ is strictly monotonic increasing. Monotonicity follows e.g. from $\tilde{f}'(x) = -\frac{c}{\sqrt{x^2-c}} < 0$.

Dependence of $g^{l}_{\Lambda,S}\left(\gamma\right)$ and $g^{u}_{\Lambda,S}\left(\gamma\right)$ on Λ

First, we calculate

$$\begin{aligned}
\frac{\partial}{\partial\Lambda} g^{l}_{\gamma,S}\left(\Lambda\right) &= \frac{\partial}{\partial\Lambda}\left(-\frac{\Lambda+\gamma}{2} + \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^{2} - \frac{S^2}{3}}\right) = -\frac{1}{2} + \frac{\frac{\Lambda-\gamma}{2}}{2\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^{2} - \frac{S^2}{3}}} > 0, \\
\frac{\partial}{\partial\Lambda} g^{u}_{\gamma,S}\left(\Lambda\right) &= \frac{\partial}{\partial\Lambda}\left(\frac{\Lambda-\gamma}{2} - \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^{2} - \frac{S^2}{3}}\right) = \frac{1}{2} - \frac{\frac{\Lambda+\gamma}{2}}{2\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^{2} - \frac{S^2}{3}}} < 0.
\end{aligned}$$

Hence, $g^{l}_{\Lambda,S}\left(\gamma\right)$ increases in Λ, $g^{u}_{\Lambda,S}\left(\gamma\right)$ decreases in Λ, and the difference - the interval where the decision is left to the executive - shrinks with an increasing range Λ

of private interests of the executive. Furthermore,

$$
\begin{aligned}
\lim_{\Lambda\to\infty} g^{l}_{\gamma,S}(\Lambda) &= \lim_{\Lambda\to\infty}\left(-\frac{\Lambda+\gamma}{2}+\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^{2}-\frac{S^{2}}{3}}\right)\\
&= \lim_{\Lambda\to\infty}\left(-\gamma-\frac{\Lambda-\gamma}{2}+\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^{2}-\frac{S^{2}}{3}}\right)\\
&= -\gamma+\lim_{x\to\infty}\left(-x+\sqrt{x^{2}-\frac{S^{2}}{3}}\right) = -\gamma+\lim_{x\to\infty}\frac{-1+\sqrt{1-\frac{S^{2}}{3x^{2}}}}{\frac{1}{x}}\\
&= -\gamma+\lim_{x\to\infty}\frac{\frac{2S^{2}}{(3x^{3})2\sqrt{1-\frac{S^{2}}{3x^{2}}}}}{-\frac{1}{x^{2}}} = -\gamma+\lim_{x\to\infty}\frac{\frac{S^{3}}{3}}{x\sqrt{1-\frac{S^{2}}{3x^{2}}}} = -\gamma \qquad (2.26)\\
\lim_{\Lambda\to\infty} g^{u}_{\gamma,S}(\Lambda) &= \lim_{\Lambda\to\infty}\left(\frac{\Lambda-\gamma}{2}-\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^{2}-\frac{S^{2}}{3}}\right)\\
&= \lim_{\Lambda\to\infty}\left(-\gamma+\frac{\Lambda+\gamma}{2}-\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^{2}-\frac{S^{2}}{3}}\right)\\
&= -\gamma+\lim_{x\to\infty}\left(x-\sqrt{x^{2}-\frac{S^{2}}{3}}\right) = -\gamma, \qquad (2.27)
\end{aligned}
$$

where we applied L'Hôpital's rule. Hence, for $\Lambda \to \infty$ the range of g for which the decision is left to the executive becomes an arbitrarily small interval around $-\gamma$.

Auxiliary Calculations on the Optimal Constitution

First, we calculate the expected difference in social welfare between a single ruler and a constitution of separated powers, in which the legislation has a private bias of γ. To simplify notation let $g^{l} \equiv g^{l}_{\Lambda,S}(\gamma)$ and $g^{u} \equiv g^{u}_{\Lambda,S}(\gamma)$. Furthermore, let $-G \leq g^{l} \leq g^{u} \leq G$. Then

$$
\begin{aligned}
& \Delta\left(EW_{SgR}(\gamma)-EW_{SepPow}(\gamma)\right)\\
=\; & \frac{1}{2G}\left(\left[\frac{g^{3}}{6\Lambda}+\frac{g^{2}}{4}+\frac{S^{2}g}{6\Lambda}\right]_{-G}^{g^{l}}+\left[\frac{g^{3}}{6\Lambda}-\frac{g^{2}}{4}+\frac{S^{2}g}{6\Lambda}\right]_{g^{u}}^{G}\right)\\
=\; & \frac{2\left(\left(g^{l}\right)^{3}-\left(g^{u}\right)^{3}\right)+3\Lambda\left(\left(g^{l}\right)^{2}+\left(g^{u}\right)^{2}\right)+2S^{2}\left(g^{l}-g^{u}\right)}{24G\Lambda}+\frac{G^{2}}{6\Lambda}-\frac{G}{4}+\frac{S^{2}}{6\Lambda}
\end{aligned}
$$

Digressions:

$$\begin{aligned}
g^l - g^u &= -\Lambda + \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} + \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} \qquad (2.28)\\
\left(g^l\right)^2 + \left(g^u\right)^2 &= \Lambda^2 + \gamma^2 - \frac{2S^2}{3} - (\Lambda+\gamma)\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} - (\Lambda-\gamma)\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}\\
\left(g^l\right)^3 - \left(g^u\right)^3 &= -\Lambda^3 + \Lambda S^2 + \left(\Lambda^2 + \Lambda\gamma + \gamma^2 - \frac{S^2}{3}\right)\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}\\
&\quad + \left(\Lambda^2 - \Lambda\gamma + \gamma^2 - \frac{S^2}{3}\right)\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}. \qquad (2.29)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left(EW_{SgR}(\gamma) - EW_{SepPow}(\gamma)\right)\\
= &\frac{\Lambda^2 - 2S^2 + 3\gamma^2}{24G} + \frac{-\Lambda^2 + 2\gamma^2 + \frac{4}{3}S^2 - \gamma\Lambda}{24G\Lambda}\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}\\
&+\frac{G^2+S^2}{6\Lambda} - \frac{G}{4} + \frac{-\Lambda^2 + 2\gamma^2 + \frac{4}{3}S^2 + \gamma\Lambda}{24G\Lambda}\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}. \qquad (2.30)
\end{aligned}$$

Proof of Proposition 16

Notice in equation 2.30 that the expected difference in social welfare between both constitutions depends only on the absolute value of γ. We can therefore concentrate on the case $\gamma > 0$, without loss of generality. Then, we want to show that the constitution of a single ruler becomes more attractive relative to a separation of powers, when the bias of the legislature, γ, increases. We show

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\left(EW_{SgR}(\gamma) - EW_{SepPow}(\gamma)\right) > 0\\
\Leftrightarrow\quad & 6\Lambda\gamma + (4\gamma - \Lambda)\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} - \left(-\Lambda^2 + 2\gamma^2 + \frac{4}{3}S^2 - \gamma\Lambda\right)\frac{\frac{\Lambda-\gamma}{2}}{2\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}}\\
&+ (4\gamma + \Lambda)\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} + \left(-\Lambda^2 + 2\gamma^2 + \frac{4}{3}S^2 + \gamma\Lambda\right)\frac{\frac{\Lambda+\gamma}{2}}{2\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}} > 0
\end{aligned}$$

$$\begin{aligned}
\Leftrightarrow \quad & 6\Lambda\gamma + 4\gamma\left(\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}} + \sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}\right) \\
& + \Lambda\left(\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}} - \sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}\right) \\
& + 2\left(\left(\frac{\Lambda}{2}\right)^2 - \frac{S^2}{3}\right)\left(\frac{1}{\sqrt{\left(1-\frac{3S^2}{3(\Lambda-\gamma)^2}\right)}} - \frac{1}{\sqrt{1-\frac{4S^2}{3(\Lambda+\gamma)^2}}}\right) \\
& + \gamma\left(\frac{\Lambda}{2}+\gamma\right)\frac{\frac{\Lambda+\gamma}{2}}{\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2 - \frac{S^2}{3}}} + \gamma\left(\frac{\Lambda}{2}-\gamma\right)\frac{\frac{\Lambda-\gamma}{2}}{\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2 - \frac{S^2}{3}}} > 0. \qquad (2.31)
\end{aligned}$$

For $\gamma \geq 0$ all terms are positive. This proves that for an increasing absolute value of γ separation of powers becomes socially less attractive compared to a single ruler.

Proof of Proposition 17

We want to consider the effect of S and Λ on the optimal constitution.

Small S:

Consider first $S = 0$. Then

$$\begin{aligned}
& \left(EW_{SgR}(\gamma) - EW_{SepPow}(\gamma)\right) \\
= \; & \frac{\Lambda^2 + 3\gamma^2}{24G} + \frac{-\Lambda^2 + 2\gamma^2 - \gamma\Lambda}{24G\Lambda}\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2} + \frac{G^2}{6\Lambda} - \frac{G}{4} + \frac{-\Lambda^2 + 2\gamma^2 + \gamma\Lambda}{24G\Lambda}\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2} \\
= \; & \frac{\gamma^2 - G^2}{4G} + \frac{G^2}{6\Lambda} < 0, \qquad (2.32)
\end{aligned}$$

where the last inequality follows from the assumption that $\gamma < G$. Hence, separation of powers is better than a non-elected single ruler, if $s = 0$. As the expected difference in welfare between both constitutions is continuous in S, separation of powers is better for all sufficiently small S (given $\gamma < G$).

Large Λ: Consider

$$\begin{aligned}
&\lim_{\Lambda\to\infty}\left(EW_{SgR}\left(\Lambda\right)-EW_{SepPow}\left(\Lambda\right)\right)\\
= \quad &\frac{3\gamma^2-2S^2}{24G}-\frac{G}{4}\\
&+\frac{2\gamma^2+\frac{4}{3}S^2}{24G}\lim_{\Lambda\to\infty}\left(\sqrt{\left(\frac{1}{2}-\frac{\gamma}{2\Lambda}\right)^2-\frac{S^2}{3\Lambda^2}}+\sqrt{\left(\frac{1}{2}+\frac{\gamma}{2\Lambda}\right)^2-\frac{S^2}{3\Lambda^2}}\right)\\
&+\frac{1}{24G}\lim_{\Lambda\to\infty}\left(\Lambda^2-(\Lambda+\gamma)\sqrt{\left(\frac{\Lambda-\gamma}{2}\right)^2-\frac{S^2}{3}}-(\Lambda-\gamma)\sqrt{\left(\frac{\Lambda+\gamma}{2}\right)^2-\frac{S^2}{3}}\right)\\
= \quad &\frac{3\gamma^2-2S^2}{24G}-\frac{G}{4}+\frac{2\gamma^2+\frac{4}{3}S^2}{24G}\\
&+\frac{1}{24G}\lim_{\Lambda\to\infty}\left(\frac{1-\sqrt{\left(\frac{1-\left(\frac{\gamma}{\Lambda}\right)^2}{2}\right)^2-\frac{(\Lambda+\gamma)^2S^2}{3\Lambda^4}}-\sqrt{\left(\frac{1-\left(\frac{\gamma}{\Lambda}\right)^2}{2}\right)^2-\frac{(\Lambda-\gamma)^2S^2}{3\Lambda^4}}}{\frac{1}{\Lambda^2}}\right)\\
\overset{L'H\hat{o}pital}{=} \quad &\frac{5\gamma^2-\frac{2}{3}S^2}{24G}-\frac{G}{4}\\
&+\frac{1}{24G}\lim_{\Lambda\to\infty}\left(-\frac{\frac{\left(1-\left(\frac{\gamma}{\Lambda}\right)\right)\gamma^2+\frac{2}{3}S^2\left(1+\frac{\gamma}{\Lambda}\right)\left(1+\frac{2\gamma}{\Lambda}\right)}{2\Lambda^3\sqrt{\left(\frac{1-\left(\frac{\gamma}{\Lambda}\right)^2}{2}\right)^2-\frac{(\Lambda+\gamma)^2S^2}{3\Lambda^4}}}+\frac{\left(1-\left(\frac{\gamma}{\Lambda}\right)^2\right)\gamma^2+\frac{2}{3}S^2\left(1-\frac{\gamma}{\Lambda}\right)\left(1-\frac{2\gamma}{\Lambda}\right)}{2\Lambda^3\sqrt{\left(\frac{1-\left(\frac{\gamma}{\Lambda}\right)^2}{2}\right)^2-\frac{(\Lambda-\gamma)^2S^2}{3\Lambda^4}}}}{-2\frac{1}{\Lambda^3}}\right)
\end{aligned}$$

$$= \quad \frac{5\gamma^2-\frac{2}{3}S^2}{24G}-\frac{G}{4}+\frac{1}{4\cdot 24G}\left(\frac{\gamma^2+\frac{2}{3}S^2}{\sqrt{\left(\frac{1}{2}\right)^2}}+\frac{\gamma^2+\frac{2}{3}S^2}{\sqrt{\left(\frac{1}{2}\right)^2}}\right)=\frac{\gamma^2-G^2}{4G}. \qquad (2.33)$$

For $\gamma < G$ this term is clearly negative. Hence, separation of powers is better than a single ruler, if Λ is sufficiently large, q.e.d.

Chapter 3

Contractual Incompleteness as a Signal of Trust

3.1 Introduction

This paper demonstrates how the fear to signal distrust can lead endogenously to incomplete contractual agreements.

According to standard results in contract theory an optimal contract should be conditional on all verifiable information containing statistical information about an agent's action or type.[1] Most real world contracts, however, condition only on few contingencies and often no explicit contract is signed at all. The costs of writing a complete contract or the limited ability to foresee all relevant contingencies can only partially explain the observed contractual incompleteness. There remain many relationships in which a simple contract could help to avoid potentially severe incentive problems at relatively low costs. Nonetheless, many people abstain from writing a complete contract. Why?

This paper argues that designing a sophisticated complete contract with fines, punishments and other explicit incentives signals distrust to the other party.

Consider the example of a scientist hiring a research assistant. Some potential incentive problems could be avoided by simple contractual arrangements. For instance,

[1] See e.g. Holmstöm [47] or Laffont and Tirole [55].

if one part of the assistant's work consists of collecting some data, the scientist could give him the right incentives by announcing to spot-check his work and to fire the assistant in case she detects some faked data. The potential damage for the scientist in case her research relies on faked data is considerable and may far outweigh her costs of spot-checking. She may, nonetheless, abstain from such an announcement, because the research assistant is likely to interpret such checks as a signal of distrust regarding his scientific dedication. The feeling that the scientist distrusts him destroys the assistant's motivation in other parts of the relationship. For instance, the assistant may expect a lower success from some potential, mutually beneficial, joint research projects if the scientist doubts his scientific dedication. He would therefore invest less effort in searching for such joint projects - to the disadvantage of the scientist.

More generally, consider a principal ("she") who is interested in the success of a project that she can only realize with the help of an agent ("he"). There are two types of agents. The *trustworthy* type has an intrinsic interest in the success of the project and is therefore willing to exert effort even without any contractual incentives.[2] The *untrustworthy* type is not intrinsically interested in the success of the project. Only explicit contractual incentives can motivate him to exert effort.

The principal may hold different beliefs about the type of the agent depending on some private signals.[3] We call the belief of the principal that the agent is trustworthy the "*trust*" of the principal in the agent. More trust means that the principal considers it more likely that the agent exerts effort even in absence of explicit incentives. The more a principal trusts the agent the lower are her expected costs (and her expected marginal costs) from contractual incompleteness. A principal can therefore try to separate herself from less trusting types by using contractual incompleteness as a signaling device.

Why should the principal have an interest in signaling trust? In our model trust is relevant in some parts of the relationship which are non-contractible by assumption. The more the principal trusts the agent, the more she is willing to rely on the agent.

[2] Equivalently, the agent could be motivated by some fairness motives.

[3] Such a signal could be private information about the agent, could stand for past experiences of the principal with other agents, or could symbolize a more or less optimistic nature of the principal.

She may, for instance, follow his advice more often or give the agent more discretion in his decisions. This, in turn, increases the productivity of the agent's effort in the joint project. Thus, if a trustworthy agent believes to be distrusted by the principal, he invests less effort and the project is less successful.

Our model focuses on the point that the concern to avoid a signal of distrust may spread contractual incompleteness from non-contractible parts to the contractible parts of a relationship.[4] Even in a perfectly contractible world, however, the concern to signal distrust causes contractual incompleteness whenever the belief to be distrusted leads to a negative reaction of the agent. We discuss other, more psychological, reasons for such reactions after presentation of our main model.

The literature on the foundations of incomplete contracts is extensive. A recent survey of this literature is Tirole [91]. Spier [87] points out most explicitly that signaling can cause contractual incompleteness.[5] In her model a risk-averse principal hires a risk-neutral agent. The principal has private information on whether the probability that her project results in high profits is high or low. In the refined equilibrium the principal offers an unconditional wage and thereby (inefficiently) forgoes some insurance in order to signal to the agent that the success probability is high. Notice, however, that, in general, asymmetric information at the contracting stage can equally well lead to more instead of less complete contracts.[6] For instance, slightly changing the setting of Spier to an informed risk-averse principal selling a risky asset to a less informed risk-neutral agent (potentially with some transaction cost for writing the contract conditional on outcome) results in a too complete contract: The principal can signal a good risk asset by conditioning the proposed contract on the outcome, although under symmetric information the agent should pay a fixed price for the asset independently of the asset's

[4] Holmström-Milgrom [48] and Bernheim-Whinston [8] give two different arguments, why it can be optimal to leave some verifiable aspects of a relationship unspecified when other aspects cannot be verified. Holmström-Milgrom show in a multi-task setting, that it may be optimal to give no explicit incentives if the agent has some intrinsic motivation, tasks are substitutes and when the unverifiable task is sufficiently important. Bernheim-Whinstons argument is based on the observation, that writing no contract may give both sides more discretion to punish the other side. This can be important in a repeated games framework where harsh out of equilibrium punishments may be necessary to sustain the desired equilibrium.

[5] See also Allen-Gale [3] for similar ideas in the context of financial economics.

[6] See also Tirole [91], p.764 for this point.

outcome. Hence, in general, the effect of signaling concerns on contractual completeness is ambiguous.

In our paper, in contrast, there is a clear prediction that contracts should be less complete when the principal wants to signal trust: The trustworthy type is defined as someone who chooses the desired action even without contractual enforcement. Hence, the more the principal trusts, the lower she estimates the costs of an incomplete contract. More contractual incompleteness therefore signals more trust and the equilibrium contract is distorted towards less completeness whenever the principal wants to signal trust.

Furthermore, our paper adds a new perspective to the literature on the detrimental effects of sanctions and explicit incentives. A number of authors[7] and recent experimental studies[8] suggest that sanctions, control and explicit incentives can crowd out intrinsic motivation and may even be counterproductive. Sanctions seem to have a particularly detrimental effect in cases where they are deliberately designed by one of the involved parties.

Bénabou-Tirole [6] and Fehr-Klein-Schmidt [29] suggest two different channels through which explicit incentives can negatively affect performance. In Bénabou-Tirole [6] a better informed principal wants to give explicit incentives to the agent if a task is unpleasant or the agent has a low ability. The less informed agent understands that explicit incentives are a signal for an unpleasant task or for his low ability. Explicit incentives therefore crowd out intrinsic motivation and are thus less powerful than under symmetric information. The model of Bénabou-Tirole requires that the principal has superior knowledge about the agent's type (or his task). Our model, in contrast, focuses on the better knowledge of the principal about her own beliefs about the agent's type. This seems a natural assumption in almost any setting with asymmetric information.

The paper by Fehr-Klein-Schmidt [29] addresses the question of how fairness concerns affect the choice of contracts. They show experimentally that it may be optimal for a principal to rely on implicit incentives (the promise of a bonus for good per-

[7] See e.g. Etzioni [26], Deci-Ryan [19] and Frey [33].

[8] See Fehr-Rockenbach [30], Gneezy-Rustichini [37],[38], and Fehr-Klein-Schmidt [29] In section 3.3 we briefly discuss the experimental findings of Falk-Kosfeld [27] in the light of our model.

formance) rather than on explicit incentives (a commitment to a limited fine after poor performance). They demonstrate by a calibration that the experimental results are consistent with a heterogeneous population where some players are inequity-averse while others act selfishly. Our argument may complement their explanation: It is natural to define trust in their setting as the belief that the agent is of a fair type. The existence of some fair-types gives implicit incentives their strength. The heterogeneity of preference types and of beliefs about these types can, in addition, make the choice of explicit incentives counterproductive, as they signal distrust.

After presenting and analyzing our model in section 3.2 we discuss our results in section 3.3 before concluding in section 3.4.

3.2 The Model

3.2.1 Setting

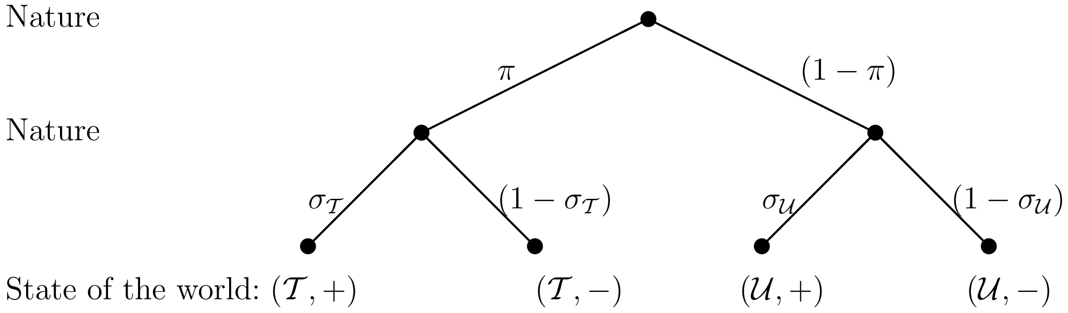
Consider the following principal-agent relationship: A principal needs to hire an agent to realize a project. The agent is one of two types, trustworthy ($\mathcal{T}$) or untrustworthy ($\mathcal{U}$); the proportion of trustworthy types in the economy, π, is strictly between zero and one. The trustworthy type of agent is intrinsically interested in the success of the project,[9] whereas the untrustworthy type does not care about the project per se.

The principal cannot observe the agent's type. She receives, however, a binary, private signal $s \in \{+, -\}$ about the type of the agent. In case the agent is trustworthy, the principal receives a positive signal with the exogenously given probability $\sigma_{\mathcal{T}}$. In case the agent is untrustworthy, the principal receives a positive signal with the exogenously given probability $\sigma_{\mathcal{U}}$, with $\sigma_{\mathcal{U}} < \sigma_{\mathcal{T}}$. These two moves by "nature" are illustrated in figure 3.1

By Bayes' rule a principal with a positive signal believes that she faces the trustworthy type with probability $\pi_+ = \frac{\pi}{\pi + \frac{\sigma_{\mathcal{U}}}{\sigma_{\mathcal{T}}}(1-\pi)} > \pi$ and a principal with a negative signal believes that she interacts with a trustworthy type with probability $\pi_- = \frac{\pi}{\pi + \frac{1-\sigma_{\mathcal{U}}}{1-\sigma_{\mathcal{T}}}(1-\pi)} <$

[9] Instead of being intrinsically interested in the project the agent may also have preferences for fairness and therefore, deliberately, exert high effort.

Figure 3.1: Nature's moves



with $0 < \sigma_{\mathcal{U}} < \sigma_{\mathcal{T}} < 1$.

π. Notice that $\pi_+ > \pi_-$, i.e. a principal with a positive signal has a stronger belief in the agent's trustworthiness. In other words, the principal with a positive signal trusts the agent more strongly than the principal with a negative signal.

The project consists of two parts, the contractible part 1 and the non-contractible part 2. The project's success in the contractible part 1, $B_1 \in \{0, \overline{B}_1\}$ depends only on an unobservable effort $e_1 \in \{0, \overline{e}_1\}$ by the agent. High effort $\overline{e}_1$ benefits the project deterministically by $\overline{B}_1 \equiv B_1(\overline{e}_1) > \overline{e}_1$. Low effort $e_1 = 0$ leads to $B_1 = 0$. A contract can condition on the outcome $B_1(e_1)$ which is realized at the very end of the relationship, i.e. after both players have chosen their actions in the second part of the relationship. Although effort e_1 is not directly observable, it can be inferred from the realized value of B_1.[10] A sufficiently harsh punishment in case of $B_1 = 0$ therefore implements a high effort level $e_1 = \overline{e}_1$ in this contractible part 1.

The success of the project in part 2 depends firstly, on an unobservable effort decision $e_2 \in \mathbb{R}_0^+$ of the agent, secondly, on a decision of the principal whether to *rely* on the agent or whether to *play safe*, and finally, on an unobservable choice by the agent, whether to work *honestly* or *dishonestly.*

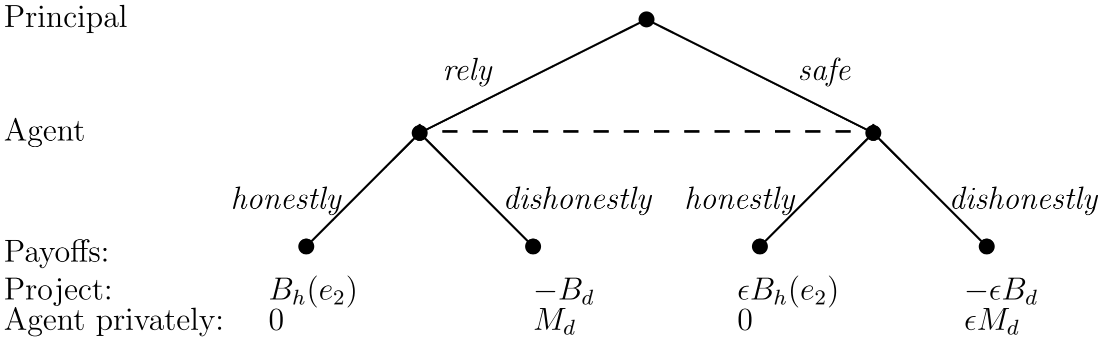
Consider again our introductory example of the scientist and the research assistant. Part 2 of the relationship then corresponds to a jointly beneficial research project.

[10] By assuming that e_1 is only ex post (indirectly) observable we avoid complication of e_1 signalling something about the agent's type.

Firstly, the assistant has to decide how much (unobservable) effort he wants to exert in studying the literature of the research topic. Then, the scientist has to decide whether to share an interesting idea with the assistant and to start a mutually beneficial joint research project. Joint work can be mutually beneficial. Yet, if the the assistant does not work reliably this could ruin the scientist's reputation. If the scientist prefers to play it safe she can instead choose a projet in which the assistant's trustworthiness is of minor importance.

More generally, in case the principal chooses to *rely* on the agent, the project's success is very sensitive to the agent's honesty. If the agent works *honestly* the project benefits by $B_h(e_2)$, where $B_h(\cdot)$ is a twice differentiable function of the agents effort $e_2 \in \mathbb{R}_0^+$, with $B_h' > 0$, $B_h'' < 0$, $B_h(0) \geq \underline{B}_h > 0$, $\lim_{e_2 \to \infty} B_h(e_2) \leq \overline{B}_h \in \mathbb{R}$ and $\lim_{e_2 \to 0} B_h'(e_2) = \infty$, where $\overline{B}_h > \underline{B}_h > 0$ are two, exogenously given boundaries. If the agent works *dishonestly* he gains a private benefit of $M_d > 0$, but causes a damage to the project of $B_d > M_d$. In case the principal chooses to *play safe*, the agent's decision is only of minor importance: all payoffs are multiplied by a small constant ϵ, with $0 < \epsilon \ll 1$. The interaction is illustrated in figure 3.2.

Figure 3.2: Reliance Game



with $B_d > M_d > 0$, $\overline{B}_h > B_h(\cdot) > \underline{B}_h > 0$ and $1 > \epsilon > 0$.

Let $B_2(e_2, \cdot)$ denote the project's success resulting from this interaction in part 2. Due to this second non-contractible part the principal has an interest to signal her trust in the agent.

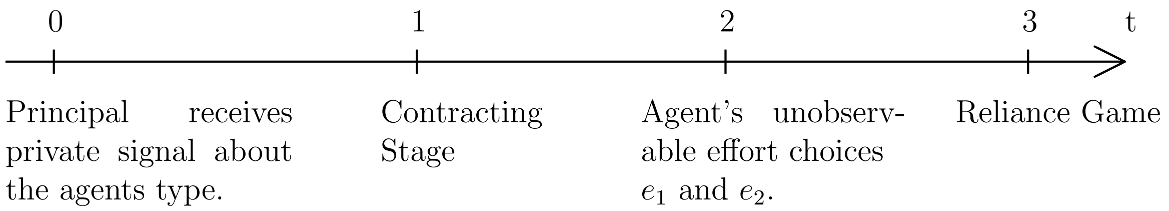
The total success B of the project is the sum of the successes, $B_1(e_1)$ and $B_2(e_2, \cdot)$, in both parts of the relationship, i.e.

$$B(e_1, e_2, \cdot) \equiv B_1(e_1) + B_2(e_2, \cdot). \tag{3.1}$$

Timing of events

The timing of events is illustrated in Figure 3.3.

Figure 3.3: Timeline


0
1
2
3
t
Principal receives private signal about the agents type.
Contracting Stage
Agent's unobservable effort choices e_1 and e_2.
Reliance Game

After nature has randomly chosen the agent's type and the principal's signal, principal and agent sign a contract. Then, the agent chooses his effort levels e_1 and e_2. Both effort choices are unobservable. Finally, principal and agent interact in the reliance game.

Contracting Stage

At the contracting stage, the principal proposes a contract. This contract can, by assumption, only be conditional on $B_1(e_1)$, the project's success in the first part. The most relevant feature of this contract is, whether the contract enforces high effort $\overline{e}_1$ by a sufficiently harsh punishment in case of $B_2(e_2) = 0$. In general, however, the principal can design a sophisticated wage scheme and the agent's decision whether to accept or reject the proposal may potentially reveal information about his type.

Here, we want to demonstrate our main point as concise as possible. For the moment, we thus restrict the set of contractual choices of the principal to a binary choice $\mathcal{C} \in \{contract\ (c), no\ contract\ (n)\}$. In appendix 3.5.2, we demonstrate that our main argument remains valid if we allow for more general contracts and if we take care of the agent's participation constraint.

The *contract* prescribes high effort $e_1 = \overline{e}_1$ in the contractible part 1. A court enforces the contract by the threat of a sufficiently harsh punishment in case of $B_1(e_1) = 0$. In case of writing *no-contract* the principal refrains from such an enforcement.

A possible interpretation for this simple setting is that principal and agent are working together already, and that there exists a binding agreement fixing the wage. In particular, let this existing agreement give the principal the discretion to enforce high effort of the agent in the contractible part 1 of the relationship through a *contract* or to abstain from doing so. Notice that in this simple setting, the agent takes no observable action. After the first exogenous signal s, the principal's belief (i.e. her trust) about the agent's type remains fixed at $\pi_{\pm}$.

Preferences

For simplicity, let the principal and the agent be risk neutral. The *untrustworthy* type of agent maximizes his private, monetary and non-monetary, payoffs M minus his total effort costs $e \equiv e_1 + e_2$. He does not care about the project.

The utility-function of the *untrustworthy*-agent is given by

$$U_{\mathcal{U}}(M, e, B(\cdot)) \quad = \quad M - e. \tag{3.2}$$

The *trustworthy* type of agent is intrinsically interested in the success of the project $B(\cdot)$. We allow for the possibility that the trustworthy agent puts a lower weight, $\kappa \leq 1$, on the project's success than the principal (in monetary units). The utility function of a *trustworthy* agent is therefore

$$U_{\mathcal{T}}(M, e, B(\cdot)) \quad = \quad M - e + \kappa B(\cdot). \tag{3.3}$$

We need, however, that the trustworthy agent sufficiently cares about the project to deliberately exert high effort $\overline{e}_1$ in part 1 and to work honestly in part 2. This is ensured by

Assumption 8

$$1 \geq \kappa \geq \max\left\{\frac{\overline{e}_1}{\overline{B}_1}, \frac{M_d}{B_d}\right\}. \tag{3.4}$$

The *principal's utility* is given by

$$V(W_P, B(\cdot)) = B(\cdot) - W_P, \tag{3.5}$$

where W_P is the wage payed by the principal. Wage payments are relevant only for the extended version in appendix 3.5.2. Here, we normalize $W_P = 0$. Hence, the principal maximizes simply the expected total success of the project $B(\cdot)$.

3.2.2 Analysis of the Principal-Agent Relationship

We analyze this principal-agent relationship backwards.

Reliance Game

Consider the last subgames of the reliance game. An *untrustworthy* agent works *dishonestly* since he does not care about the project and is tempted by the private benefit $M_d > 0$ (or $\epsilon M_d > 0$). A *trustworthy* agent, however, works *honestly*, since $\kappa B_h(e_2) > 0 \geq M_d - \kappa B_d$. A trustworthy agent values the additional success of the project more than his private benefits from working dishonestly.

Should the principal *rely* on the agent? She anticipates that the agent works honestly if and only if he is trustworthy. Depending on her private signal the principal has a belief $\pi_\pm$ that the agent is of a trustworthy type. It is optimal for her to *rely* on the agent if and only if $\pi_\pm B_h(\hat{e}_2) - (1 - \pi_\pm) B_d \geq 0$, or, equivalently, if and only if

$$\pi_\pm \ \geq \ \frac{B_d}{B_h(\hat{e}_2) + B_d}. \tag{3.6}$$

Notice, that in case of

$$\pi_\pm \ > \ \frac{B_d}{\underline{B}_h + B_d} \tag{3.7}$$

the principal plays *rely* independently of her expectations on the effort choice $\hat{e}_2$ of a trustworthy agent. Similarly, if

$$\pi_\pm \quad < \quad \frac{B_d}{\overline{B}_h + B_d} \tag{3.8}$$

the principal plays *safe* independently of her expectations on the effort choice e_2 of a trustworthy agent.

Assumption 9 *We assume:*

$$\pi_+ > \frac{B_d}{\underline{B}_h + B_d} \quad \textit{and} \quad \pi_- < \frac{B_d}{\overline{B}_h + B_d}. \tag{3.9}$$

In other words, a trusting principal with a positive signal *relies* on the agent, whereas a distrusting principal with a negative signal plays *safe*.

The Agent's Effort Choice

An *untrustworthy* agent dislikes effort and does not care about the project. Thus, he always chooses $e_2 = 0$. Furthermore, in the first part of the project he chooses low effort $e_1 = 0$, too, unless high effort is enforced by a *contract*.

A *trustworthy* agent chooses high effort $e_1 = \overline{e}_1$ in the contractible part 1, even without any explicit incentives by a contract. Since $\kappa\overline{B}_1 > \overline{e}_1$, by assumption 8, he cares more about the value added to the project than about his effort costs. In the non-contractible part 2 the trustworthy agent's effort e_2 depends on his belief about whether the principal trusts him. In case he expects the principal to play *safe*, effort e_2 increases the project's success only by $\epsilon B_h(e_2)$; whereas if he expects the principal to *rely* on him, effort e_2 increases the project's success by $B_h(e_2)$.

The agent understands that the principal *relies* on him if and only if she received a positive signal. Ex ante, i.e. before the principal chooses a contract, the trustworthy agent has the rational belief of $\sigma_{\mathcal{T}}$ that the principal received a positive signal and the untrustworthy agent has a belief of $\sigma_{\mathcal{U}}$. The agent may, however, change his belief after observing the principal's contract choice.

Let $\alpha_{\mathcal{T}}$ denote the trustworthy agent's belief that the principal received a positive signal after he observed her contractual choice. In particular, we denote by $\alpha_{\mathcal{T}}^{c}$ the belief of a trustworthy agent if the principal chose to write a *contract* and $\alpha_{\mathcal{T}}^{n}$ the belief of a trustworthy agent if the principal chose to write *no contract.*

Then, a trustworthy agent chooses his effort e_2 in order to maximize

$$\max_{e_2 \in \mathbb{R}_0^+} \left(\alpha_{\mathcal{T}} B_h(e_2) + (1-\alpha_{\mathcal{T}})\, \epsilon B_h(e_2) - e_2\right) \tag{3.10}$$

$$\text{FOC:} \qquad B_h'\left(e_2\right) = \frac{1}{\epsilon + \alpha_{\mathcal{T}}\left(1-\epsilon\right)}. \tag{3.11}$$

$B_h'(\cdot)$ is a strictly decreasing function (since $B_h'' < 0$). Hence, the inverse function $\left(B_h'\right)^{-1}$ exists and is strictly decreasing, too. The optimal effort e_2^* for the trustworthy agent given the belief $\alpha_{\mathcal{T}}$ is therefore

$$e_2^*(\alpha_t) = \left(B_h'\right)^{-1}\left(\frac{1}{\epsilon + \alpha_{\mathcal{T}}\left(1-\epsilon\right)}\right). \tag{3.12}$$

Notice that $e_2^*(\cdot)$ is strictly increasing in $\alpha_{\mathcal{T}}$. Hence, $B_h^*(\alpha_{\mathcal{T}}) \equiv B_h(e_2^*(\alpha_{\mathcal{T}}))$ is strictly increasing in $\alpha_{\mathcal{T}}$, too. The more the trustworthy agent believes to be trusted the more effort e_2^* he exerts. This effort of the agent increases the success of the project and thereby benefits the principal.

Analysis of the Contracting Stage

The principal's decision of whether or not to write a contract is observed by the agent and influences the trustworthy agent's belief $\alpha_{\mathcal{T}}$ of whether the principal trusts him. This belief, in turn, influences the agent's effort decision $e_2^*(\alpha_t)$.

In a perfect Bayesian equilibrium the contractual choice $\mathcal{C}_+ \in \{c, n\}$, of a principal with a positive signal, and $\mathcal{C}_- \in \{c, n\}$, of a principal with a negative signal, must both be optimal given the principal's type (i.e. his belief $\pi_\pm$) and given the trustworthy agent's beliefs $\alpha_{\mathcal{T}}^{c}$ and $\alpha_{\mathcal{T}}^{n}$. On the other hand, the trustworthy agent's beliefs $\alpha_{\mathcal{T}}^{c}$ after observing the choice of a *contract* and $\alpha_{\mathcal{T}}^{n}$ after observing the choice of *no-contract* must be rational given the equilibrium contractual choices $\mathcal{C}_+$ and $\mathcal{C}_-$.

Consider a potential equilibrium candidate which we denote, in a slight abuse of notation[11], by $(\mathcal{C}_+, \mathcal{C}_-, \alpha^c_{\mathcal{T}}, \alpha^n_{\mathcal{T}})$. Let $V^n_\pm$ denote the expected payoff of a $(\pm)$-principal from the choice of writing *no-contract*, and $V^c_\pm$ the expected payoff from the choice of writing a *contract*. Then,

$$\begin{aligned}
V^n_+ &= \pi_+ \left(\overline{B}_1 + B^*_h(\alpha^n_{\mathcal{T}})\right) - (1-\pi_+) B_d && (3.13)\\
V^c_+ &= \overline{B}_1 + \pi_+ B^*_h(\alpha^c_{\mathcal{T}}) - (1-\pi_+) B_d && (3.14)\\
V^n_- &= \pi_- \left(\overline{B}_1 + \epsilon B^*_h(\alpha^n_{\mathcal{T}})\right) - (1-\pi_-) \epsilon B_d && (3.15)\\
V^c_- &= \overline{B}_1 + \pi_- \epsilon B^*_h(\alpha^c_{\mathcal{T}}) - (1-\pi_-) \epsilon B_d. && (3.16)
\end{aligned}$$

The following relations are useful for the further analysis:

$$\begin{aligned}
V^n_+ \gtreqqless V^c_+ &\Leftrightarrow B^*_h(\alpha^n_{\mathcal{T}}) - B^*_h(\alpha^c_{\mathcal{T}}) \gtreqqless \frac{1-\pi_+}{\pi_+}\overline{B}_1 && (3.17)\\
V^n_- \gtreqqless V^c_- &\Leftrightarrow B^*_h(\alpha^n_{\mathcal{T}}) - B^*_h(\alpha^c_{\mathcal{T}}) \gtreqqless \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon} && (3.18)
\end{aligned}$$

The left hand sides of the rearranged inequalities are identical for the trusting and the distrusting principal and depend on the trustworthy agent's beliefs after the contract choice. The right hand side is greater for the trusting principal than for the distrusting one. Thus, the distrusting principal is less willing to chose *no-contract* compared to the trusting principal..

Even a trusting principal only considers to refrain from the option to enforce high effort in part 1 by means of a contract if the expected costs $(1-\pi_+)\overline{B}_1$ are below the maximal possible expected gains from such a signal $\pi_+(B^1_h - B^0_h)$, where we defined

$$B^{\alpha_{\mathcal{T}}}_h \equiv B^*_h(\alpha_{\mathcal{T}}) \qquad \left(\equiv B_h\left(e^*_2(\alpha_{\mathcal{T}})\right)\right), \qquad (3.19)$$

i.e. $B^1_h \equiv B^*_h(\alpha_{\mathcal{T}} = 1)$ and $B^0_h \equiv B^*_h(\alpha_{\mathcal{T}} = 0)$. Otherwise, both types of principal always pool on writing a *contract*. Thus, we concentrate on the interesting case:

[11] The beliefs of the untrustworthy agent are payoff irrelevant. The remaining equilibrium actions are the unique best responses given $(\mathcal{C}_+, \mathcal{C}_-, \alpha^c_{\mathcal{T}}, \alpha^n_{\mathcal{T}})$.

Condition 1

$$\overline{B}_1 < \frac{\pi_+}{1-\pi_+}\left(B_h^1 - B_h^0\right). \tag{3.20}$$

Refinement by the Intuitive Criterion: Signaling games suffer from a multiplicity of perfect Bayesian equilibria that are often sustained by pessimistic out of equilibrium beliefs. We mainly focus on equilibria that are consistent with the intuitive criterion, an equilibrium refinement introduced by Cho and Kreps [17].[12] It constrains the set of out-of-equilibrium-beliefs by the following equilibrium domination argument: Consider an out-of equilibrium action in a perfect Bayesian equilibrium of the game. If one type A can never expect to profit from this deviation (given that all players play best-responses to some out of equilibrium belief) whereas a different type potentially could profit, then any "intuitive" belief should assign probability 0 to type A. The precise definition of the intuitive criterion is in the appendix. We introduce the following

Definition 7 *An* **intuitive equilibrium** *is a perfect Bayesian Equilibrium that is consistent with the intuitive criterion.*

In the appendix we specify and derive all perfect Bayesian equilibria of this game. Here, we concentrate on our main point and summarize the most relevant results in the following propositions.

Proposition 18 *Under assumption 8, 9 and condition 1, there always exists an intuitive equilibrium in which the trusting principal chooses to write no-contract.*

Remark 5 *These intuitive equilibria are*

a) for $\overline{B}_1 \geq \frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^1 - B_h^0\right)$:

the separating equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write a contract.

[12] In our context we need the somewhat more elaborate definition of the intuitive criterion used by Maskin-Tirole [61], because we deal with more stages than the standard two-stage signaling game.

b) for $\frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^{\sigma_\mathcal{T}} - B_h^0\right) < \overline{B}_1 < \frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^1 - B_h^0\right)$:

the hybrid equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write no contract with probability $q = \frac{\sigma_\mathcal{T}}{1-\sigma_\mathcal{T}}\frac{1-\alpha_t^n}{\alpha_t^n}$.

c) for $\overline{B}_1 \leq \frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^{\sigma_\mathcal{T}} - B_h^0\right)$:

the pooling equilibria on writing no contract, in which both types of principal forbear from writing a contract.

To see the intuition for proposition 18 and remark 5, notice that the expected costs of refraining from writing a contract are smaller for the trusting principal, $(1-\pi_+)\overline{B}_1$, than for the distrusting principal, $(1-\pi_-)\overline{B}_1$. In addition, the expected gains from signaling trust are higher for the trusting principal. Firstly, she considers it more likely that she interacts with a trustworthy agent, the type who invests more effort when he believes in being trusted. Secondly, since she plays “rely” in the relaince game, her payoff is more sensitive to the agent’s investment.

Consider a very large $\overline{B}_1$, such that condition 1 does not hold. Then, the expected losses from signaling trust by writing *no contract* are too large compared to the potential gains; the trusting as well as the distrusting principal pool on writing a *contract*.

Decreasing $\overline{B}_1$ until condition 1 just holds, a second equilibrium emerges, the separating equilibrium of remark 5 a. In the range of remark 5 a the trusting principal takes the risk to forgo $\overline{B}_1$ in part 1 of the project to separate herself from the distrusting type. For a distrusting principal writing no contract, to imitate the trusting type, is too expensive.

As we decrease $\overline{B}_1$ further the costs of signaling trust, by forbearing from writing a contract, become smaller. Perfect separation of the trusting and distrusting principal becomes unsustainable when the expected costs of writing no contract for the distrusting type fall below the gains from being mistaken for a trusting type.[13] Then, a fraction of distrusting principals starts imitating the signal. The larger this fraction of distrusting imitators, the lower, in equilibrium, the belief of the rational agent after

[13] Notice that this threshold, $\frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^1 - B_h^0\right)$, is close to zero if ϵ is small.

observing the signal "*no contract*". In the hybrid equilibrium of remark 5 b the fraction of distrusting principals imitating the signal takes exactly the value that makes a distrusting principal indifferent between choosing "*contract*" or "*no contract*".

Decreasing $\overline{B}_1$ further, decreases, in this hybrid equilibrium, the probability that a distrusting principal reveals his type by writing a contract. When this probability reaches zero, we end up in the pooling equilibrium of remark 5 c, sustained by an out of equilibrium belief, that a "*contract*" is a signal of a distrusting type.

In addition to the equilibria in proposition 18 and remark 5 there exist further perfect Bayesian equilibria, in particular the pooling equilibria on writing a *contract.* For an intermediate range of $\overline{B}_1$, however, these additional equilibria do not pass the intuitive criterion and intuitive equilibrium actions are unique.

Proposition 19 *Under assumption 8, 9, condition 1, and in case of*

$$\frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^1 - B_h^{\sigma_\mathcal{T}}\right) < \overline{B}_1 < \frac{\pi_+}{1-\pi_+}\left(B_h^1 - B_h^{\sigma_\mathcal{T}}\right) \tag{3.21}$$

the equilibria of remark 5 are the only intuitive equilibria. In particular, the trusting principal chooses to write no contract in any intuitive equilibrium.

Corollary 10 *Under assumption 8, 9, condition 1, and inequality 3.21 there is a unique outcome in all intuitive equilibria. More precisely,*

a) for $\overline{B}_1 \geq \frac{\pi_-}{1-\pi_-}\epsilon\,(B_h^1 - B_h^0)$**:** *the separating equilibrium, in which the trusting principal chooses to write "*no contract*", is the unique intuitive equilibrium.*

b) for $\frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^{\sigma_\mathcal{T}} - B_h^0\right) < \overline{B}_1 < \frac{\pi_-}{1-\pi_-}\epsilon\,(B_h^1 - B_h^0)$**:** *the hybrid equilibrium in which the trusting principal chooses to write no contract and the distrusting principal chooses to write no contract with probability* $q = \frac{\sigma_\mathcal{T}}{1-\sigma_\mathcal{T}}\frac{1-\alpha_t^n}{\alpha_t^n}$ *is the unique intuitive equilibrium.*

c) for $\overline{B}_1 \leq \frac{\pi_-}{1-\pi_-}\epsilon\left(B_h^{\sigma_\mathcal{T}} - B_h^0\right)$**:** *all intuitive equilibria are pooling equilibria on writing no contract; these equilibria differ only in their out-of-equilibrium beliefs,* $\alpha_\mathcal{T}^c$.

The intuition for proposition 19 follows from two observations. Firstly, under condition 3.21 there exists no hybrid equilibrium in which the trusting principal uses a mixed strategy. Secondly, the pooling equilibrium on writing a *contract* requires that the out of equilibrium belief for writing *no contract*, α_T^n, is smaller than 1. Under condition 3.21, however, the intuitive criterion requires an out of equilibrium belief of 1, as the distrusting principal can never expect to profit, whereas the trusting principal does profit for $\alpha_T^n = 1$.

These propositions confirm the main point of our paper; the principal may choose to leave contracts incomplete to avoid a signal of distrust. Proposition 18 states that, under condition 1, there always exists an equilibrium in which at least the trusting principal abstains from writing a contract, and that these equilibria pass the intuitive criterion. Under condition 3.21 these are the only equilibria that are not excluded by the intuitive criterion - as stated in proposition 19. Then, there is a unique intuitive equilibrium outcome, in which the trusting principal forbears from writing a contact.

.

3.3 Discussion

Our model demonstrates that the fear to signal distrust can endogenously cause contractual incompleteness. The trusting principal, in particular, prefers to to write *no contract*, although under symmetric information she would strictly prefer to write such a contingent contract. She is more afraid of being mistaken for the distrusting principal, than of being exploited by an untrustworthy agent.

In the simple model, with a binary contractual choice, *contract* or *no contract*, such an equilibrium, in which the trusting type refrains from writing a contract, exits only if the exogenously given costs of contractual incompleteness are not to high. The range for uniqueness of the intuitive equilibrium is even smaller.

In appendix 3.5.2 the principal can design more general wage schemes. This complicates the analysis and requires additional assumptions to ensure that the principal can not separate the trustworthy and the untrustworthy agent by a screening contract. In return, however, we get a clear cut prediction about the intuitive equilibrium outcome: In any intuitive equilibrium the distrusting principal chooses the complete contract and the trusting principal chooses the least-cost separating contract. In other words, the trusting principal chooses a degree of contractual incompleteness just sufficient to separate herself from the distrusting type.

Intuitively, the principal can choose any degree of contractual incompleteness by designing the contract properly, i.e. she can choose any loss in case the agent turns out to be untrustworthy.[14] In an intuitive equilibrium, the trusting principal never proposes a contract that is chosen by the mistrusting principal with a strictly positive probability. Marginal and absolute costs of contractual incompleteness are lower for a trusting principal while the gains from signaling trust are larger. In a candidate equilibrium contract chosen by both types of principal the trusting type can separate herself by choosing an additional degree of contractual incompleteness just sufficiently high that the distrusting type strictly prefers the candidate equilibrium contract. By the intuitive criterion the agent should assign a belief of $\alpha_{\mathcal{T}} = 1$ when observing such a deviating contract. Then, however, the trusting principal has a strict preference for this deviation and the candidate equilibrium could not have been an intuitive equilibrium in the first place.

In our model, the value of signaling trust comes from an under-investment of the agent in a non-contractible part of the relationship in case he believes to be distrusted. If we are willing to depart further from standard economic preferences and take a more sociological or psychological standpoint then the negative consequences of signaling distrust become even more relevant. The proposal of a detailed complete contract with sophisticated fines and rewards basically states "I believe you are one of those types who exploits me if he can". Most people would perceive such a statement as an insult and lose any motivation to invest in this relationship. In addition, trust seems to have a strong mutual component. How can I trust you if you do not trust me? In a relation

[14] Losses smaller than $\overline{B}_1 > 0$ can be generated by a lottery over contracts - with some probability complete, with the counter probability incomplete. Losses larger than $\overline{B}_1 > 0$ can be generated by giving the agent no explicit incentives and a commitment by the principal to burn some money in case the agent turns out to be untrustworthy, i.e. in case of $B_1 = 0$.

of similar partners this inference can be rational to some degree and may be amplified by the false consensus effect.[15]

In particular sociologist emphasize that human behavior and well-being is strongly influenced by social status and what other people think about them. Feeling distrusted lowers one's utility directly and a signal of distrust may destroy the potential surplus from a relationship. Only if we feel trusted we attach positive emotions to a relationship, we are willing to invest in it and to forgo some instant profits to maintain the impression of being trustworthy.[16] Luhmann [59] coined the expression of the "self-fulfilling prophecy of distrust". This effect is demonstrated neatly in a recent experimental study by Falk-Kosfeld [27]. An agent can spend any amount from his endowment of 120 token in an investment which benefits only the principal (by the double amount of the investment). Upfront, the principal can choose whether she wants to force the agent to invest at least 10 token, or whether she abstains from any control (then the minimum investment is 0). The large majority of principals chose not to control the agent and, in fact, on average agents invested significantly more when the principal chose to not control. An additional control-treatment where the minimum investment was exogenously given demonstrates that it is not the control per se that leads to lower investments of the agent, but the fact that the principal has deliberately chosen to control. This clearly suggests that a principal choosing to control the agent signals distrust and that this crowds out trustworthiness of the agent.

3.4 Conclusions

This paper demonstrates that the fear to signal distrust can lead to endogenous contractual incompleteness, even when there are no costs of writing a contract and all types of principal would prefer to write a complete contract under symmetric informa-

[15] To see this consider the following stylized model: The world can be either good or bad. In a good world most people are trustworthy while in a bad world most people are untrustworthy. If someone knows only his own preferences he should assign a higher probability to a bad world if he is untrustworthy, himself. Hence, people who distrust others may be more likely to be untrustworthy themselves. See also Englemann [22].

[16] Recent studies suggest (see Sunnafrank-Ramirez [89]) that the first impressions are decisive for the long-term nature of a relationship. Contract proposal are often the starting point in a relation.

tion about their beliefs. Our results are driven, firstly, by asymmetric information on the principal's beliefs about the agent's type and ,secondly, by the importance of trust, and the belief to be trusted, for the relationship.

This fear to signal distrust by proposing to write a contract can cause considerable inefficiencies. Carefully designed policies may therefore become relevant and can help to mitigate these inefficiencies. Many couples are reluctant to sign a prenuptial agreement as they are afraid of signaling distrust to their spouse. A well designed, fair, standard regulation, at least for the case that no contract is written, may therefore be very important. In fact, most states do regulate the consequences in case of a divorce to some degree.

Similarly, parents who funded their children's education should, in return, receive support from their children in case they need help when becoming elderly. Most parents, however, will be very reluctant to insist on any explicit contract with their children - for not showing distrust. Again, some prudential regulation by a government may help to mitigate such problems.

3.5 Appendix

3.5.1 Proofs and all Perfect Bayesian Equilibria

The Perfect Bayesian Equilibria

Separating Equilibrium: Under condition 1, there exists a separating, perfect Bayesian equilibrium if and only if

$$B_h^1 \;\; \leq \;\; B_h^0 + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.22}$$

In this equilibrium the trusting principal with belief π_+ chooses to write *no contract*, (i.e. $\mathcal{C}_+ = n$), and the distrusting principal with belief π_- chooses to write a *contract*, (i.e. $\mathcal{C}_- = c$).

Pooling on writing *No Contract*: There exist perfect Bayesian pooling equilibria on writing *no-contract* if and only if

$$B_h^{\sigma_{\mathcal{T}}} \quad \geq \quad B_h^0 + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.23}$$

On the equilibrium path the belief of the trustworthy agent is $\alpha_{\mathcal{T}}^n = \sigma_{\mathcal{T}}$ and the equilibrium is sustained by an out of equilibrium belief $\alpha_{\mathcal{T}}^c \leq \left(B_h^*\right)^{-1}\left(B_h^*\left(\sigma_{\mathcal{T}}\right) - \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}\right)$.

Pooling on writing a *Contract*: There always exist perfect Bayesian pooling equilibria on writing a *contract.* On the equilibrium path, the belief of the trustworthy agent is $\alpha_{\mathcal{T}}^c = \sigma_{\mathcal{T}}$ and the equilibrium is sustained by an out of equilibrium belief $\alpha_{\mathcal{T}}^n \leq \left(B_h^*\right)^{-1}\left(B_h^*\left(\sigma_{\mathcal{T}}\right) + \frac{1-\pi_+}{\pi_+}\overline{B}_1\right)$.

Hybrid Equilibria with a random contractual choice of the distrusting principal:

Under condition 1, there exists a hybrid perfect Bayesian equilibrium in which the trusting principal chooses to write *no-contract* with certainty and the distrusting principal randomly mixes between both contractual choices if and only if

$$B_h^{\sigma_{\mathcal{T}}} \leq B_h^0 + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon} \leq B_h^1. \tag{3.24}$$

In this equilibrium the trustworthy agent's beliefs are $\alpha_t^c = 0$ and $\alpha_t^n = \left(B_h^*\right)^{-1}\left(\frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon} + B_h^0\right)$. The mixing probability for the distrusting principal to choose *no-contract* is $q = \frac{\sigma_{\mathcal{T}}}{1-\sigma_{\mathcal{T}}}\frac{1-\alpha_t^n}{\alpha_t^n}$.

Hybrid Equilibria with random contract choice of the trusting principal:

Under condition 1 there exists a hybrid perfect Bayesian equilibrium in which the distrusting principal writes a *contract* with certainty and the trusting principal randomly mixes between "*contract*" and "*no contract* if and only if

$$B_h^1 - B_h^{\sigma_{\mathcal{T}}} < \frac{1-\pi_+}{\pi_+}\overline{B}_1. \tag{3.25}$$

In this equilibrium the trustworthy agent's beliefs are $\alpha_t^c = (B_h^*)^{-1}\left(B_h^1 - \frac{1-\pi_+}{\pi_+}\overline{B}_1\right)$ and $\alpha_t^n = 1$. The mixing probability for the trusting principal to choose to write *no-contract* is $q' = 1 - \frac{1-\sigma_{\mathcal{T}}}{\sigma_{\mathcal{T}}}\frac{\alpha_t^c}{1-\alpha_t^c}$.

Refinement by the Intuitive Criterion: Which of the perfect Bayesian equilibria are affected by the refinement of the intuitive criterion? The separating and hybrid equilibria have no out-of-equilibrium belief and therefore pass the intuitive criterion, anyway. Pooling equilibria on writing *no contract* pass the intuitive criterion too, since both types of principal could potentially profit from a deviation to the complete contract.

The only equilibria that potentially fail to pass the intuitive criterion are the pooling equilibria on the complete contract. The intuitive criterion requires the out-of-equilibrium belief after the choice of *no contract* to be $\alpha_t^n = 1$ if for this belief[17] the trusting principal expects to profit from the deviation, whereas the distrusting principal does not. In fact, no pooling equilibrium on the complete contract passes the intuitive criterion if (and only if)[18]

$$\frac{1-\pi_+}{\pi_+}\overline{B}_1 < B_h^1 - B_h^{\sigma_{\mathcal{T}}} < \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.26}$$

In the remainder of this paper we denote a perfect Bayesian equilibrium that passes the intuitive criterion as an **intuitive equilibrium**.

Proofs and Derivation of the Equilibria

Perfect Bayesian Equilibria and Intuitive Equilibria

The following definition is useful in the further analysis:

[17] Clearly, $\alpha_t = 1$ leads to maximal profits for the principal.

[18] In case of $\frac{1-\pi_+}{\pi_+}\frac{\overline{B}_1}{\Delta\delta+\Delta B_h} = 1-\sigma_{\mathcal{T}}$, the intuitive criterion still requires $\alpha_t(cc=0)=1$, but pooling on the complete contract remains incentive compatible.

Definition 8

$$e_2^{\alpha_{\mathcal{T}}} \equiv e_2^*(\alpha_{\mathcal{T}}), \qquad \textit{in particular,} \tag{3.27}$$

$$e_2^0 \equiv e_2^*(\alpha_{\mathcal{T}} = 0) \quad \textit{and} \quad e_2^1 \equiv e_2^*(\alpha_{\mathcal{T}} = 1). \tag{3.28}$$

Separating Equilibrium In the only possible separating equilibrium[19] the trusting principal chooses to write *no contract* and the distrusting principal chooses to write a *contract*. In this perfect-Bayesian-equilibrium the agent has the right beliefs the principal's signal, i.e. he holds the beliefs $\alpha_t^n = 1$ and $\alpha_t^c = 0$.

The trustworthy agent therefore invests e_2^1 if and only if the principal writes no contract.

$\mathcal{C}_+ = n$ and $\mathcal{C}_- = c$ are best responses for both types of principal if and only if

$$(IC_+) \qquad V_+^n \geq V_+^c \tag{3.29}$$

$$(IC_-) \qquad V_-^n \leq V_-^c, \tag{3.30}$$

i.e. by equivalences 3.17 and 3.18

$$B_h^1 - B_h^0 \geq \frac{1-\pi_+}{\pi_+}\overline{B}_1 \tag{3.31}$$

$$B_h^1 - B_h^0 \leq \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.32}$$

Equation 3.31 is already implied by condition 1. Summarizing, under condition 1, there exists a perfectly separating perfect Bayesian equilibrium if and only if condition 3.32 holds.

Pooling Equilibria on "*no contract*" (n): If both types of principal choose in equilibrium to write *no contract* then the agent's belief remains unchanged when he

[19] There can be no separating equilibrium in which the distrusting principal chooses to write *no contract* since incompleteness is costly for her and she prefers to not being separated from the trusting type.

observes this action, i.e. $\alpha^n_{\mathcal{T}} = \sigma_{\mathcal{T}}$ (and $\alpha^n_{\mathcal{U}} = \sigma_{\mathcal{U}}$). Out of equilibrium the trustworthy agent has some belief $\alpha^c_{\mathcal{T}} \in [0,1]$. This forms a perfect-Bayesian-Equilibrium if and only if

$$(IC_+) \quad V^n_+ \quad \geq V^c_+ \tag{3.33}$$

$$(IC_-) \quad V^n_- \quad \geq V^c_-, \tag{3.34}$$

which by equivalence 3.17 and 3.18 corresponds to

$$(IC_+) \qquad B_h^{\sigma_{\mathcal{T}}} - B_h^*(\alpha^c_{\mathcal{T}}) \geq \frac{1-\pi_+}{\pi_+}\overline{B}_1 \tag{3.35}$$

$$(IC_-) \qquad B_h^{\sigma_{\mathcal{T}}} - B_h^*(\alpha^c_{\mathcal{T}}) \geq \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.36}$$

(IC_+) is already implied by (IC_-). Pooling on *no contract* can be sustained as a perfect-Bayesian-Equilibrium by an out-of-equilibrium-belief $\alpha^c_{\mathcal{T}} \in [0,1]$ if and only if the equilibrium can be sustained by $\alpha^c_{\mathcal{T}} = 0$. This is the case if and only if

$$\sigma_{\mathcal{T}} \geq (B_h^*)^{-1}\left(B_h^0 + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}\right). \tag{3.37}$$

Pooling Equilibria on "*contract*" (c): If both types of principal choose in equilibrium to write a *contract* (c) then the belief of the agent remains unchanged when observing this action, i.e. $\alpha^c_{\mathcal{T}} = \sigma_{\mathcal{T}}$ (and $\alpha^c_{\mathcal{T}} = \sigma_{\mathcal{U}}$). Out of equilibrium the agent has some belief $\alpha^n_{\mathcal{T}} \in [0,1]$. These beliefs and contractual choices form a perfect-Bayesian-Equilibrium if and only if

$$(IC_+) \quad V^n_+ \quad \leq V^c_+ \tag{3.38}$$

$$(IC_-) \quad V^n_- \quad \leq V^c_-, \tag{3.39}$$

which by equivalence 3.17 and 3.18 corresponds to

$$(IC_+) \qquad B_h^*\left(\alpha_t^n\right) - B_h^{\sigma_{\mathcal{T}}} \leq \frac{1-\pi_+}{\pi_+}\overline{B}_1 \tag{3.40}$$

$$(IC_-) \qquad B_h^*\left(\alpha_t^n\right) - B_h^{\sigma_{\mathcal{T}}} \leq \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.41}$$

$$(IC_+) \quad \alpha_t^n \quad \leq \left(B_h^*\right)^{-1}\left(B_h^{\sigma_{\mathcal{T}}} + \frac{1-\pi_+}{\pi_+}\overline{B}_1\right) \tag{3.42}$$

$$(IC_-) \quad \alpha_t^n \quad \leq \left(B_h^*\right)^{-1}\left(B_h^{\sigma_{\mathcal{T}}} + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}\right). \tag{3.43}$$

(IC_-) is already implied by (IC_+). Hence, pooling on writing a *contract* can always be sustained as a perfect-Bayesian-Equilibrium, e.g. by the out-of-equilibrium-belief $\alpha_{\mathcal{T}}^n = 0$.

Hybrid-Equilibrium with $\mathcal{C}_+ = n$ and $q \equiv prob(\mathcal{C}_- = n) \in (0,1)$: In this equilibrium the agent knows, when observing a *contract* that the principal is distrusting, i.e. $\alpha_{\mathcal{T}}^c = 0$. In case he observes *no contract* he updates his beliefs by Bayes-rule:

$$\alpha_{\mathcal{T}}^n \equiv prob(+|\mathcal{C} = n) = \frac{prob\left(\mathcal{C} = n|+\right) prob\left(+\right)}{prob(\mathcal{C} = n)} = \frac{\sigma_{\mathcal{T}}}{\sigma_{\mathcal{T}} + \left(1-\sigma_{\mathcal{T}}\right)q}. \tag{3.44}$$

Notice that $\sigma_{\mathcal{T}} < \alpha_{\mathcal{T}}^n < 1$ for every $q \in (0,1)$. Vice versa, for any $\sigma_{\mathcal{T}} < \alpha_{\mathcal{T}}^n < 1$ there exists a $q \in (0,1)$ that leads to this $\alpha_{\mathcal{T}}^n$, namely $q = \frac{\sigma_{\mathcal{T}}}{1-\sigma_{\mathcal{T}}}\frac{1-\alpha_{\mathcal{T}}^n}{\alpha_{\mathcal{T}}^n}$. These beliefs and contractual choices form a perfect Bayesian equilibrium if and only if

$$(IC_+) \quad V_+^n \quad \geq V_+^c \tag{3.45}$$

$$(IC_-) \quad V_-^n \quad = V_-^c, \tag{3.46}$$

which by equivalence 3.17 and 3.18 corresponds to

$$(IC_+) \qquad B_h^*(\alpha_t^n) - B_h^0 \geq \frac{1-\pi_+}{\pi_+}\overline{B}_1 \tag{3.47}$$

$$(IC_-) \qquad B_h^*(\alpha_t^n) - B_h^0 = \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.48}$$

(IC_+) is implied already by (IC_-). Hence, there exists a perfect Bayesian hybrid equilibrium in which the trusting principal chooses "*no contract*" and the distrusting principal mixes between "*contact*" and "*no contract*" if and only if

$$B_h^{\sigma_\mathcal{T}} < B_h^0 + \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon} < B_h^1. \tag{3.49}$$

Hybrid-Equilibrium with $q' \equiv prob(\mathcal{C} = n) \in (0,1)$ **and** $\mathcal{C} = c$**:** In this equilibrium the agent knows, when observing "*no contract*", that the principal is of the trusting type, i.e. $\alpha_\mathcal{T}^n = 1$. In case he observes a *contract* he updates his beliefs by Bayes-rule:

$$\alpha_t^c = \frac{prob\,(\mathcal{C} = c|+)\,prob_\mathcal{T}\,(+)}{prob_\mathcal{T}\,(\mathcal{C} = c)} = \frac{(1-q')\,\sigma_\mathcal{T}}{(1-q')\sigma_\mathcal{T} + (1-\sigma_\mathcal{T})}. \tag{3.50}$$

Notice that $0 < \alpha_\mathcal{T}^c < \sigma_\mathcal{T}$ for every $q' \in (0,1)$. These beliefs and contractual choices form an equilibrium if and only if

$$(IC_+) \qquad V_+^n = V_+^c \tag{3.51}$$

$$(IC_-) \qquad V_-^n \leq V_-^c, \tag{3.52}$$

which by equivalence 3.17 and 3.18 corresponds to

$$(IC_+) \qquad B_h^1 - B_h^*(\alpha_t^c) = \frac{1-\pi_+}{\pi_+}\overline{B}_1 \tag{3.53}$$

$$(IC_-) \qquad B_h^1 - B_h^*(\alpha_t^c) \leq \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}. \tag{3.54}$$

(IC_-) is implied already by (IC_+). Hence, there exists a perfect-Bayesian hybrid equilibrium in which the distrusting principal chooses to write a *contract* and the trusting principal mixes between "*contract*" and "*no contract*" if and only if

$$B_h^0 < B_h^1 - \frac{1-\pi_+}{\pi_+}\overline{B}_1 < B_h^{\sigma_{\mathcal{T}}}. \tag{3.55}$$

The first inequality is guaranteed already by condition 1.

It is straightforward to check that there do not exist any further perfect Bayesian equilibria.

Equilibrium Refinement by the Intuitive Criterion In our context of only two types of principals and more than the standard two stages of a signaling game the intuitive criterion by Cho-Kreps [17] (CK) takes the following form (see also Maskin-Tirole [61]):

Let $T = \{+, -\}$ denote the set of the two types of principals. Let $BR(\mathbf{w}, \alpha_t)$ denote the (unique) equilibrium strategies of the continuation game between the principal and the agent after $\mathbf{w}$ has been offered and has led the trustworthy agent[20] to update his belief to α_t. Consider a candidate perfect Bayesian equilibrium that leads in equilibrium to an expected utility V_i^* for a principal of type i.

We denote an out-of-equilibrium contract proposal $\tilde{\mathbf{w}}$ as **equilibrium dominated for type** $i \in T$, if and only if

$$V_i^* \quad > \quad \max_{\alpha_t \in [0,1]} V_i\left(BR\left(\tilde{\mathbf{w}}, \alpha_t\right)\right). \tag{3.56}$$

Definition 9 *A perfect Bayesian equilibrium passes the* **intuitive criterion** *if and only if the out-of-equilibrium beliefs* $\alpha_t(\tilde{\mathbf{w}})$ *assign zero probability to type* i *(i.e.* $\alpha_t(\tilde{\mathbf{w}}) = 0$ *if* $i = +$ *and* $\alpha_t(\tilde{\mathbf{w}}) = 1$ *if* $i = -$*) whenever* $\tilde{\mathbf{w}}$ *is equilibrium dominated for type* i *and not equilibrium dominated for the other type* j*.*

All separating or hybrid equilibria are not affected by this additional constraint on the out of equilibrium beliefs since "*contract*" and "*no contract*" are both played

[20] Notice that the equilibrium strategies are independent of the belief α_u of the untrustworthy agent.

in equilibrium by some type with a strictly positive probability. Only the pooling equilibria need to be analyzed.

Pooling Equilibria on "*Contract*" (c): In equilibrium the expected payoffs for the trusting principal and distrusting principal are

$$\begin{aligned} V_+^{c,eq} &= \overline{B}_1 + \pi_+ B_h^{\sigma_\mathcal{T}} - (1-\pi_+)\, B_d & (3.57) \\ V_-^{c,eq} &= \overline{B}_1 + \pi_- \epsilon B_h^{\sigma_\mathcal{T}} - (1-\pi_-)\, \epsilon B_d. & (3.58) \end{aligned}$$

In case a principal deviates to writing "*no contract*", the trustworthy agent plays a best response to some belief. For both types of principal the best they can hope for is that this belief is $\alpha_\mathcal{T}^n = 1$ and the trustworthy agent invests $e_2 = e_2^1$. The maximal resulting expected payoffs for the trusting principal and for the distrusting principal are therefore:

$$\begin{aligned} V_+^{n,max} &= \pi_+ \left(\overline{B}_1 + B_h^1\right) - (1-\pi_+)\, B_d & (3.59) \\ V_-^{n,max} &= \pi_- \left(\overline{B}_1 + \epsilon B_h^1\right) - (1-\pi_-)\, \epsilon B_d. & (3.60) \end{aligned}$$

Writing "*no contract*" is equilibrium dominated for the trusting principal if and only if

$$\begin{aligned} & V_+^{c,eq} > V_+^{n,max} & (3.61) \\ \Leftrightarrow \quad & B_h^1 - B_h^{\sigma_\mathcal{T}} < \frac{1-\pi_+}{\pi_+} \overline{B}_1. & (3.62) \end{aligned}$$

For the distrusting principal writing "*no contract*" is equilibrium dominated if and only if

$$\begin{aligned} & V_-^{c,eq} > V_-^{n,max} & (3.63) \\ \Leftrightarrow \quad & B_h^1 - B_h^{\sigma_\mathcal{T}} < \frac{1-\pi_-}{\pi_-} \frac{\overline{B}_1}{\epsilon}. & (3.64) \end{aligned}$$

Hence, for

$$\frac{1-\pi_+}{\pi_+}\overline{B}_1 < B_h^1 - B_h^{\sigma_\mathcal{T}} < \frac{1-\pi_-}{\pi_-}\frac{\overline{B}_1}{\epsilon}, \tag{3.65}$$

the intuitive criterion demands $\alpha_\mathcal{T}^n = 1$ and a pooling-equilibrium on writing a "*contract*" does not pass this refinement due to (IC_+)[21].

The **Pooling Equilibria on "*No Contract*" (n):** all pooling equilibria on "*no contract*" pass the intuitive criterion, since both types could, potentially, profit from a deviation to "*contract*" (e.g. when $\alpha_\mathcal{T}^c = \alpha_\mathcal{T}^n$).

3.5.2 Wage Scheme Contracts

In the second scenario we allow for a more general class of contracts and take the participation constraint of the agent into account. Payments can condition on B_1 and thereby indirectly on the agents effort e_1. We still assume that the contract can not condition on anything in the non-contractible part 2 of the relationship.

We have a problem of two-sided asymmetric information. The focus of this paper is on a signaling story: The principal signals her trust by proposing some contractual incompleteness. In general, however, the principal might also try to screen between both types of agents. Firstly, she could try to propose a contract, that fulfills only the participation constraint of the trustworthy agent. In other words the principal would make the trustworthy agent pay for the right to work for him. In most settings, this is completely implausible and just an artefact of how we modeled the trustworthy agent's preferences. We therefore change our modeling of the trustworthy agent's preferences slightly: Instead of gaining from a successful projects we assume that a trustworthy agent suffers from working for a projet that is less successful than it could be. Then the trustworthy agent still works deliberately, but he is not willing to to pay for his

[21] In case of $\frac{1-\pi_+}{\pi_+}\overline{B}_1 = B_h^1 - B_h^{\sigma_\mathcal{T}}$ equilibrium domination still requires $\alpha_\mathcal{T}^n = 1$, but (IC_+) is nonetheless satisfied.

job.[22] The trustworthy agent's utility is then given by

$$U_{\mathcal{T}}(m, e, B) = m - e + \kappa(B - B_{max}), \tag{3.66}$$

where $B_{max} = \overline{B}_1 + \overline{B}_h$ is the maximal success the project could ever have. Furthermore, we assume that both types of agent have an outside option of 0.

A second way the principal could try to screen between both types of agent is by offering a menu of contracts. In case of perfect screening, the principal would learn the type of the agent by his choice of contract and change her belief (i.e. her trust) correspondingly. Then the agent would understand that the principal knows his type and all asymmetric information would be resolved in equilibrium. Such complications can not arise, however, if being trusted is sufficiently important for the untrustworthy agent (i.e. if M_d is sufficiently large). Then screening is impossible. The untrustworthy type always mimics the behavior of the trustworthy agent at the contracting stage. Under the following assumption the principal can not screen the type of the agent[23]:

Assumption 10 (No Screening Condition (sufficient))

$$M_d \;>\; \frac{\kappa\left(\overline{B}_1 + \overline{B}_h\right)}{\min\{\sigma_{\mathcal{U}}, 1 - \sigma_{\mathcal{U}}\}} \tag{3.67}$$

Lemma 6 *Under assumption 10 the untrustworthy agent chooses the same contract as the trustworthy agent in any perfect Bayesian equilibrium.*

The proof is at the end of the appendix.

The Contract Proposal The contract-proposal is designed by an informed party: The principal has relevant private information when proposing the contract and may therefore signal something about her private information to the agent.

In general, the principal may propose any probability distribution over a set of

[22] The following to arguments lead to the same result: either we can assume that the trustworthy agent has a higher outside option, or that the principal needs to hire always the agent and cannot risk that the untrustworthy agent rejects the offer.

[23] Even if this assumption does not hold, screening may well be too expensive for being optimal.

contracts[24]. Each of these contracts can only condition on the realization of B_1, i.e. a contract specifies a tuple $\mathbf{w} \equiv (\underline{w}_P, \overline{w}_P, \underline{w}_A, \overline{w}_A)$ with $\underline{w}_P \geq \underline{w}_A$ and $\overline{w}_P \geq \overline{w}_A$. $\underline{w}_P$ is the principals payment in case of $B_1 = 0$ and $\overline{w}_P$ her payment in case of $B_1 = \overline{B}_1$. $\underline{w}_A$ is the wage the agent receives in case of $B_1 = 0$ and $\overline{w}_A$ the wage the agent receives if $B_1 = \overline{B}_1$. The requirements $\underline{w}_P \geq \underline{w}_A$ and $\overline{w}_P \geq \overline{w}_A$ capture that the wage of the agent has to be payed by the principal. In case of $w_P > w_A$ the principal commits to"burn some money".

After the principal's contract proposal the agent updates his beliefs about the principal's type. Then, he accepts the contract if and only if his expected utility under the contract equals at least his outside option. Otherwise, the agent receives his outside option of 0. We assume that the principal always wants to hire the agent.

When the agent accepted a court draws a realization from the proposed probability distribution. The court is committed to enforce this realized deterministic contract $(\underline{w}_P, \overline{w}_P, \underline{w}_A, \overline{w}_A)$.[25]

The Reduced Form of the Contract Proposal We simplify the analysis by treating only those contract proposals as different, that lead to different payoffs for at least one of the relevant types.

When the principal offers a contract, she has to care only about the belief and the participation constraint of the trustworthy agent. The untrustworthy agent always accepts the contract, when the trustworthy does and since the untrustworthy agent exerts no effort anyway, his beliefs do not affect the principal. The only way the untrustworthy agent reacts to a contract is that he plays $e_1 = \overline{e}_1$ if and only if $\Delta w_A \equiv \overline{w}_A - \underline{w}_A > \overline{e}_1$.[26] For any given contract the principal can directly incorporate this into

[24] Maskin and Tirole [61] discuss in detail the problem of contract design by an informed principal with common values. They allow the principal to design (almost) any contracting mechanism. In particular the principal can propose a menu of contracts from which she herself chooses one contract after the agent accepted. This gives the signaling game screening properties. However, their analysis is not directly applicable to our setting since here the agents beliefs on the principal's type do matter even after the contract is written. Nonetheless, we conjecture that our results do still hold in this more general setting, since their techniques select also the least-cost separating equilibrium (as we do).

[25] Thus we avoid any potential problems of ex post incentives to renegotiate the contract.

[26] For convenience, we assume the tie breaking rule that the untrustworthy agent plays $e_1 = 0$ if $\Delta w_A = \overline{e}_1$.

her calculation of her expected payoff.

Only the expected payoffs of the trusting principal, the distrusting principal and the trustworthy agent should therefore be relevant for the equilibrium analysis. When the trustworthy agent accepted the contract proposal of the principal, he has a certain belief $\alpha_{\mathcal{T}}$ about the principal's type and we can calculate the expected equilibrium payoffs for each type, given this belief $\alpha_{\mathcal{T}}$. We simplify the analysis by

Definition 10 *Two contract-proposals $\mu(\mathbf{w})$ and $\mu'(\mathbf{w})$ are* **payoff equivalent** *if for every $\alpha_{\mathcal{T}}(\mu) = \alpha_{\mathcal{T}}(\mu')$ both contract proposals lead in equilibrium to the same expected payoffs for the trusting principal, the distrusting principal and the trustworthy agent.*

At the end of this appendix we demonstrate that each equivalence class can be described by a tuple (w, l, l_u) where $w \in \mathbb{R}$, $l \geq 0$ and $l_u \geq \min\{0, \overline{B}_1 - \overline{e}_1 - l\}$. W.l.o.g., consider only contracts in which the trustworthy agent has an incentive to chose[27] $e_1 = \overline{e}_1$: w is the expected wage payment from the principal to the trustworthy agent, l is an unconditional loss of utility for the principal and l_u is a loss in utility for the principal from contractual incompleteness in case the agent turns out to be untrustworthy. The possibility to choose $l > 0$ (i.e. to commit to always burn some money) will play only a minor role for the further analysis and is only included for completeness. The possibility to commit to a loss conditional on meeting an untrustworthy type, however, is crucial. The expected costs of such a commitment are lower for the trusting principal than for the distrusting principal. Choosing a high l_u can therefore serve as a signal of trust.

The expected payoffs from a contract (w, l, l_u) conditional on the belief $\alpha_{\mathcal{T}}(\mathbf{w}) \equiv \alpha_{\mathcal{T}}(w, l, l_u)$ are for the trusting and distrusting principal, respectively :

$$\begin{aligned} V_{+}(w, l, l_u, \alpha_{\mathcal{T}}(\mathbf{w})) &= -(w + l) - (1 - \pi_{+})\, l_u \quad + \pi_{+} B_h^{\alpha_{\mathcal{T}}(\mathbf{w})} \quad + \left[\overline{B}_1 - (1 - \pi_{+})\, B_d\right], \\ V_{-}(w, l, l_u, \alpha_{\mathcal{T}}(\mathbf{w})) &= -(w + l) - (1 - \pi_{-})\, l_u \quad + \pi_{-} \epsilon B_h^{\alpha_{\mathcal{T}}(\mathbf{w})} \quad + \left[\overline{B}_1 - (1 - \pi_{-})\, \epsilon B_d\right]. \end{aligned}$$

The respectively last terms in rectangular brackets are independent of the contract

[27] This no real restriction as we show in appendix 3.5.2 that there is such a contract in each equivalence class.

(and independent of the belief resulting from the contract-proposal). It is convenient to re-normalize the principals utilities to

$$
\begin{aligned}
V_+(w,l,l_u,\alpha_{\mathcal{T}}(\mathbf{w})) &= -(w+l)-(1-\pi_+)\,l_u+\pi_+B_h^{\alpha_{\mathcal{T}}(\mathbf{w})}, \qquad (3.68)\\
V_-(w,l,l_u,\alpha_{\mathcal{T}}(\mathbf{w})) &= -(w+l)-(1-\pi_-)\,l_u+\pi_-\epsilon B_h^{\alpha_{\mathcal{T}}(\mathbf{w})}. \qquad (3.69)
\end{aligned}
$$

The expected utility of the trustworthy agent is

$$
U_{\mathcal{T}}(w,l,l_u,\alpha_{\mathcal{T}}(\mathbf{w})) \;=\; w-\overline{e}_1-e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}\right)\right)+\kappa\left(\left(\alpha_{\mathcal{T}}\left(\mathbf{w}\right)\left(1-\epsilon\right)+\epsilon\right)B_h^{\alpha_{\mathcal{T}}(\mathbf{w})}-\overline{B}_h\right),
$$

where we used that $B_{max}=\overline{B}_1+\overline{B}_h$. Thus, if the trustworthy agent has the belief $\alpha_{\mathcal{T}}\left(\mathbf{w}\right)$ after a contract-proposal $\mathbf{w}$, then he accepts the contract if and only if

$$
w \;\geq\; \overline{e}_1+e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}\right)\right)+\kappa\left(\overline{B}_h-\left(\alpha_{\mathcal{T}}\left(\mathbf{w}\right)\left(1-\epsilon\right)+\epsilon\right)B_h^{\alpha_{\mathcal{T}}(\mathbf{w})}\right). \qquad (3.70)
$$

The right hand side decreases in $\alpha_{\mathcal{T}}$, it is the stronger the belief of the agent that he is trusted, the lower the minimum wage offer that he requires to accept the offer. Thus, a wage $w\geq\overline{e}_1+e_2^0+\kappa\left(\overline{B}_h-\epsilon B_h^0\right)$ assures acceptance of the t-agent for any belief $\alpha_{\mathcal{T}}$.

It is typical for signaling games to suffer from a multiplicity of perfect Bayesian equilibria. We have a (3 dimensional) continuum of contracts, correspondingly many possible out of equilibrium beliefs and, therefore, a large number of equilibria. We derive and specify them in the next section. Given this multiplicity it is remarkable that only one of this equilibria passes the intuitive criterion - the least cost separating equilibrium:

Proposition 20 *The only perfect Bayesian equilibrium passing the intuitive criterion is the least cost separating equilibrium:*

$$
\begin{aligned}
\mathbf{w}^+ &= \left(w^+=\overline{e}_1+e_2^1+\kappa\left(\overline{B}_h-B_h^1\right),l^+=0,l_u^+=\frac{e_2^0-e_2^1+\left(\kappa+\epsilon\pi_-\right)B_h^1-\left(\kappa+\pi_-\right)\epsilon B_h^0}{1-\pi_-}\right)\\
\mathbf{w}^- &= \left(w^-=\overline{e}_1+e_2^0+\kappa\left(\overline{B}_h-\epsilon B_h^0\right),l^-=0,l_u^-=0\right). \qquad (3.71)
\end{aligned}
$$

Before we prove proposition 20, we briefly discuss the result. Intuitively the bite of the intuitive criterion in this more general set of possible contracts comes from the fact that marginally increasing incompleteness of the contract (i.e. the costs l_u of meeting the untrustworthy type) is less costly for the trusting principal. Whenever the agent might mistake her for the distrusting type, she could gradually increase incompleteness until the bad type would not follow her.

l_u was defined as the loss of the principal compared to a complete contract in case the agent turns out to be untrustworthy (for a fixed belief $\alpha_{\mathcal{T}}$). Such a loss exists only when the contract provides insufficient incentives for the untrustworthy type to exert high effort e_1. An $l_u > 0$ means therefore that the contract is incomplete (at least with some strictly positive probability).

In the unique intuitive equilibrium the trusting principal chooses to propose an incomplete contract to signal her trust in the agent's trustworthiness.

Derivation of the perfect Bayesian equilibria in pure strategies and proof of proposition 20 in scenario 2:

Separating equilibria Let $\mathbf{w}^+ \equiv (w^+, l^+, l_u^+)$ denote the contract proposal of the (+)-principal and $\mathbf{w}^- \equiv (w^-, l^-, l_u^-)$ the contract proposed by the (-)-principal. Then in any perfect Bayesian separating equilibrium: $\alpha_{\mathcal{T}}(\mathbf{w}^+) = 1$ and $\alpha_{\mathcal{T}}(\mathbf{w}^-) = 0$.
The proposal of $\mathbf{w}^-$ can only be optimal for the (-)-principal in such a separating equilibrium if

$$\mathbf{w}^- = (w = \overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right), l = 0, l_u = 0). \tag{3.72}$$

There exist some out of equilibrium beliefs that sustain $\mathbf{w}^+$ and $\mathbf{w}^-$ as a separating equilibrium if and only if the beliefs $\alpha_{\mathcal{T}}(\mathbf{w}) \equiv 0 \ \forall_{\mathbf{w} \neq \mathbf{w}^+}$ sustain the equilibrium. Notice that with these out of equilibrium beliefs $\mathbf{w}^-$ is more attractive for both types of principal than any out of equilibrium contract proposal.
In the separating equilibrium the t-agent accepts the proposal $\mathbf{w}^+$ if and only if

$$w^+ \geq \overline{e}_1 + e_2^1 + \kappa\left(\overline{B}_h - B_h^1\right). \tag{3.73}$$

Furthermore, it must be optimal for each type of principal to choose the corresponding proposal, i.e.

$$(IC_+) \qquad \tilde{V}_+(w^+, l^+, l_u^+, \alpha_{\mathcal{T}}(\mathbf{w}^+)) \geq \tilde{V}_+(w^-, l^-, l_u^-, \alpha_{\mathcal{T}}(\mathbf{w}^-)) \tag{3.74}$$

$$(IC_-) \qquad \tilde{V}_-(w^+, l^+, l_u^+, \alpha_{\mathcal{T}}(\mathbf{w}^+)) \leq \tilde{V}_-(w^-, l^-, l_u^-, \alpha_{\mathcal{T}}(\mathbf{w}^-)). \tag{3.75}$$

Rearranging leads to

$$(IC_+) \qquad \left(w^+ + l^+\right) + (1 - \pi_+)\, l_u^+ \leq \overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right) + \pi_+\left(B_h^1 - B_h^0\right) \tag{3.76}$$

$$(IC_-) \qquad \left(w^+ + l^+\right) + (1 - \pi_-)\, l_u^+ \geq \overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right) + \epsilon\pi_-\left(B_h^1 - B_h^0\right). \tag{3.77}$$

Hence, $\mathbf{w}^+$ and $\mathbf{w}^-$ can be sustained as an perfect Bayesian separating equilibrium if and only if $\mathbf{w}^- = \left(w = \overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right), l = 0, l_u = 0\right)$ and $\mathbf{w}^+$ fulfills conditions 3.73, 3.76 and 3.77.

Separating equilibria passing the intuitive criterion The intuitive criterion selects the least cost separating equilibrium with $w^+ = \overline{e}_1 + e_2^1 + \kappa\left(\overline{B}_h - B_h^1\right)$, $l^+ = 0$ and l_u^+ having the value just sufficiently high to fulfill condition 3.77, i.e.

$$l_u^+ = \frac{e_2^0 - e_2^1 + (\kappa + \epsilon\pi_-)\, B_h^1 - (\kappa + \pi_-)\, \epsilon B_h^0}{1 - \pi_-} \tag{3.78}$$

First we proof by contradiction that all perfect Bayesian separating equilibria with $w^+ + l^+ > \overline{e}_1 + \overline{e}_2$ do not pass the intuitive criterion. The intuition is that it is simply cheaper for the trusting principal to separate via l_u^+ than by w^+ or l^+ as the latter two losses have for both types of principal the same expected value, whereas in case of l_u^+ the trusting principal expects a lowers loss than the distrusting one. More formally, suppose (w^+, l^+, l_u^+) fulfills conditions 3.76 and 3.77 and $\xi \equiv w^+ + l^+ - \overline{e}_1 - \overline{e}_2 > 0$. Then the contract $\hat{\mathbf{w}}^+ \equiv (\hat{w}^+ = \overline{e}_1 + \overline{e}_2, \hat{l}^+ = 0, \hat{l}_u^+ = l_u^+ + \frac{\xi}{1 - \frac{\pi_+ + \pi_-}{2}})$ must, by the intuitive criterion lead to the belief $\alpha_{\mathcal{T}}(\hat{\mathbf{w}}^+) = 1$ and thus $V_+(\hat{\mathbf{w}}^+) > V_+(\mathbf{w}^+)$ which contradicts optimality of $\mathbf{w}^+$.

Hence, we must have $w^+ = \overline{e}_1 + e_2^1 + \kappa\left(\overline{B}_h - B_h^1\right)$, $l^+ = 0$, and

$$\frac{e_2^0 - e_2^1 + (\kappa + \pi_+) B_h^1 - (\kappa\epsilon + \pi_+) B_h^0}{1 - \pi_+} \overset{(IC_+)}{\geq} l_u^+ \overset{(IC_-)}{\geq} \frac{e_2^0 - e_2^1 + (\kappa + \epsilon\pi_-) B_h^1 - (\kappa + \pi_-)\epsilon B_h^0}{1 - \pi_-}.$$

Under the intuitive criterion a contract with $l_u^+ > \frac{e_2^0-e_2^1+(\kappa+\epsilon\pi_-)B_h^1-(\kappa+\pi_-)\epsilon B_h^0}{1-\pi_-}$ cannot be an equilibrium, as for sufficiently small $\epsilon > 0$ the contract with $\hat{l}_u^+ \equiv l_u^+ - \epsilon > \frac{e_2^0-e_2^1+(\kappa+\epsilon\pi_-)B_h^1-(\kappa+\pi_-)\epsilon B_h^0}{1-\pi_-}$ has to lead to $\alpha_{\mathcal{T}}(\hat{l}_u^+) = 1$ and leads therefore to a higher payoff to the (+)-principal.

Hence $l_u^+ = \frac{e_2^0-e_2^1+(\kappa+\epsilon\pi_-)B_h^1-(\kappa+\pi_-)\epsilon B_h^0}{1-\pi_-}$.

Lemma 7 *The only separating equilibrium passing the intuitive criterion is the least cost separating equilibrium:*

$$\begin{aligned}\mathbf{w}^+ &= \left(w^+ = \overline{e}_1 + e_2^1 + \kappa\left(\overline{B}_h - B_h^1\right), l^+ = 0, l_u^+ = \frac{e_2^0 - e_2^1 + (\kappa + \epsilon\pi_-) B_h^1 - (\kappa + \pi_-)\epsilon B_h^0}{1 - \pi_-}\right) \\ \mathbf{w}^- &= \left(w^- = \overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right), l^- = 0, l_u^- = 0\right). \end{aligned} \tag{3.79}$$

Pooling Equilibria Here, a contract $\mathbf{w} = (w, l, l_u)$ can be sustained by some beliefs as a pooling equilibrium if and only if it can be sustained by the out of equilibrium beliefs $\alpha_{\mathcal{T}}(\mathbf{w}') = 0\ \forall_{\mathbf{w}' \neq \mathbf{w}}$. In case all out of equilibrium beliefs are 0 the most attractive alternative to the pooling contract $\mathbf{w}$ is the contract $\left(w' = \left(\overline{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right)\right), l' = 0, l_u' = 0\right)$. In equilibrium the belief of the trustworthy agent is $\alpha_{\mathcal{T}}(\mathbf{w}) = \sigma_{\mathcal{T}}$. The equilibrium contract is therefore accepted by the t-agent if and only if

$$(PC_{\mathcal{T}}) \qquad w \geq \overline{e}_1 + e_2^*(\sigma_{\mathcal{T}}) + \kappa\left(\overline{B}_h - (\sigma_{\mathcal{T}}(1-\epsilon) + \epsilon) B_h^{\sigma_{\mathcal{T}}}\right). \tag{3.80}$$

The incentives to deviate from the equilibrium contract are higher for the distrusting principal as the trusting principal expects lower costs from an $l_u > 0$ and values more when the agent beliefs to be trusted. The (-)-principal has no incentive to deviate from the equilibrium contract if and only if

$$(IC_-) \qquad V_-\left(\mathbf{w}, \sigma_{\mathcal{T}}\right) \geq V_-\left(w' = \left(\bar{e}_1 + e_2^0 + \kappa\left(\overline{B}_h - \epsilon B_h^0\right)\right), l' = 0, l'_u = 0, \alpha_{\mathcal{T}} = 0\right).$$

Hence a contract $\mathbf{w} = (w, l, l_u)$ is sustainable as a perfect Bayesian pooling equilibrium if and only if

$$(PC_{\mathcal{T}}) \qquad w \geq \bar{e}_1 + e_2^*(\sigma_{\mathcal{T}}) + \kappa\left(\overline{B}_h - \left(\sigma_{\mathcal{T}}\left(1-\epsilon\right) + \epsilon\right) B_h^{\sigma_{\mathcal{T}}}\right), \tag{3.81}$$

$$(IC_-) \qquad (w + l) + (1 - \pi_-) l_u \leq \bar{e}_1 + e_2^0 + \pi_- \epsilon\left(B_h^{\sigma_{\mathcal{T}}} - B_h^0\right) + \kappa\left(\overline{B}_h - \epsilon B_h^0\right) \tag{3.82}$$

In such a pooling equilibrium the payoffs for the (+) and (-) principal are respectively

$$\tilde{V}_+^{pool}(w, l, l_u) = -(w + l) - (1 - \pi_+) l_u + \pi_+ B_h^{\sigma_{\mathcal{T}}} \tag{3.83}$$

$$\tilde{V}_-^{pool}(w, l, l_u) = -(w + l) - (1 - \pi_-) l_u + \pi_- \epsilon B_h^{\sigma_{\mathcal{T}}}. \tag{3.84}$$

None of the pooling equilibria passes the intuitive criterion In these pooling equilibria the intuitive criterion demands that an out of equilibrium belief $\alpha_{\mathcal{T}}(w', l', l'_u) = 1$ if the (-)-principal does for any belief strictly worse under the alternative contract (w', l', l'_u) compared to her equilibrium payoff and if the (+)-principal may profit for some beliefs. In other words $\alpha_{\mathcal{T}}(w', l', l'_u) = 1$ if

$$(IC_-) \qquad \tilde{V}_-^{pool}(w, l, l_u) > \tilde{V}_-(w', l', l'_u, \alpha_{\mathcal{T}} = 1)$$

$$(IC_+) \qquad \tilde{V}_+^{pool}(w, l, l_u) \leq \tilde{V}_+(w', l', l'_u, \alpha_{\mathcal{T}} = 1),$$

or equivalently

$$(IC_-) \qquad (w' + l') - (w + l) + (1 - \pi_-)(l'_u - l_u) > \pi_- \epsilon\left(B_h^1 - B_h^{\sigma_{\mathcal{T}}}\right) \tag{3.85}$$

$$(IC_+) \qquad (w' + l') - (w + l) + (1 - \pi_+)(l'_u - l_u) \leq \pi_+\left(B_h^1 - B_h^{\sigma_{\mathcal{T}}}\right). \tag{3.86}$$

Consider e.g. the contract $(w' \equiv w, l' \equiv l, l'_u \equiv l_u + \frac{\pi_+}{1-\pi_+}\epsilon\left(B^1_h - B^{\sigma_\mathcal{T}}_h\right)$. The intuitive criterion demands $\alpha_\mathcal{T}(w', l', l'_u) = 1$ and then the trusting principal does strictly better when proposing contract $\mathbf{w}'$ then in the pooling equilibrium. Hence, none of the pooling equilibria survives the intuitive criterion.

No Hybrid equilibrium passes the intuitive criterion The argument is similar to the pooling-equilibria. In a hybrid equilibrium there must exist at least one contract (w, l, l_u) that is chosen by both types of principal with a strictly positive probability, and hence $0 < \alpha_\mathcal{T}(w, l, l_u) < 1$. Consider e.g. the contract $(w' \equiv w, l' \equiv l, l'_u \equiv l_u + \frac{\pi_+}{1-\pi_+}\epsilon\left(B^1_h - B^{\alpha_\mathcal{T}(w,l,l_u)}_h\right)$. The intuitive criterion demands $\alpha_\mathcal{T}(w', l', l'_u) = 1$ and then the trusting principal does strictly better when proposing contract (w', l', l'_u) then in the equilibrium contract (w, l, l_u). Hence, none of the hybrid equilibria passes the intuitive criterion.

Proof of Lemma 6

Let $\mathbf{w} \equiv (\underline{w}, \Delta w)$ denote a contract that pays the agent a wage of $\underline{w}$ in case of $B_1 = 0$ and a wage of $\overline{w} = \underline{w} + \Delta w$ in case of $B_1 = \overline{B}_1$. We want to show that under the No-Screening Condition 10 the principal can not even partially separate the trustworthy from the untrustworthy agent by a menu of contracts $\{(\underline{w}_\mathcal{U}, \Delta w_\mathcal{U}), (\underline{w}_\mathcal{T}, \Delta w_\mathcal{T})\}$ (from which the agent can choose his preferred contract). Let $\pi_\pm(\mathbf{w_i})$ denote the belief of the $(\pm)$-principal that the agent is trustworthy when the agent chose the contract $\mathbf{w_i}$ from the menu. Furthermore, let $\alpha_{\mathcal{T}/\mathcal{U}}(\mathbf{w_i})$ denote the belief of the $(\mathcal{T}/\mathcal{U})$-agent, that the principal trusts him (in the trust-game at the last stage) when he has chosen contract $\mathbf{w_i}$. The expected utility of the agent from choosing contract $\mathbf{w_i}$ are in dependence of his type

$$\begin{aligned} U_\mathcal{U}(\mathbf{w_i}) &= \underline{w}_i + \max\{0, \Delta w_i - \overline{e}_1\} + \left(\alpha_\mathcal{U}(\mathbf{w_i})(1-\epsilon) + \epsilon\right) M_d \\ U_\mathcal{T}(\mathbf{w_i}) &= \underline{w}_i + \max\{0, \Delta w_i - \overline{e}_1 + \kappa \overline{B}_1\} - e^*_2\left(\alpha_\mathcal{T}(\mathbf{w_i})\right) \\ &\quad + \kappa B^{\alpha_\mathcal{T}(\mathbf{w_i})}_h \left(\alpha_\mathcal{T}(\mathbf{w_i})(1-\epsilon) + \epsilon\right) - \kappa B_{max}. \end{aligned}$$

Necessary conditions for the menu $\{(\underline{w}_{\mathcal{U}}, \Delta w_{\mathcal{U}}), (\underline{w}_{\mathcal{T}}, \Delta w_{\mathcal{T}})\}$ to screen (or partially screen) between both types of agents are

$$(IC_{\mathcal{U}}) \qquad U_{\mathcal{U}}(\mathbf{w}_{\mathcal{U}}) \geq U_{\mathcal{U}}(\mathbf{w}_{\mathcal{T}}) \tag{3.87}$$

$$(IC_{\mathcal{T}}) \qquad U_{\mathcal{T}}(\mathbf{w}_{\mathcal{U}}) \leq U_{\mathcal{T}}(\mathbf{w}_{\mathcal{T}}). \tag{3.88}$$

Equivalently

$$\begin{aligned} &\max\{0, \Delta w_{\mathcal{T}} - \bar{e}_1\} - \max\{0, \Delta w_{\mathcal{U}} - \bar{e}_1\} + (1-\epsilon)\left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right)\right) M_d \overset{(IC_{\mathcal{U}})}{\leq} \underline{w}_{\mathcal{U}} - \underline{w}_{\mathcal{T}} \\ \overset{(IC_{\mathcal{T}})}{\leq} &\max\{0, \Delta w_{\mathcal{T}} - \bar{e}_1 + \kappa \overline{B}_1\} - \max\{0, \Delta w_{\mathcal{U}} - \bar{e}_1 + \kappa \overline{B}_1\} - \left(e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)\right) - e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)\right)\right) \\ &+ \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{T}})} - \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{U}})} \end{aligned}$$

A necessary condition for the existence of an $\underline{w}_{\mathcal{T}}$ and $\underline{w}_{\mathcal{U}}$ for which $(IC_{\mathcal{U}})$ and $(IC_{\mathcal{T}})$ both hold is that

$$\begin{aligned} \left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right)\right) M_d \leq\ & \max\{0, \Delta w_{\mathcal{T}} - \bar{e}_1 + \kappa \overline{B}_1\} - \max\{0, \Delta w_{\mathcal{T}} - \bar{e}_1\} \\ & + \max\{0, \Delta w_{\mathcal{U}} - \bar{e}_1\} - \max\{0, \Delta w_{\mathcal{U}} - \bar{e}_1 + \kappa \overline{B}_1\} \\ & - \left(e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)\right) - e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)\right)\right). \\ & + \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{T}})} \\ & - \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{U}})}. \end{aligned}$$

The first line on the right hand side is always greater or equal to $\overline{B}_1$ and the second line is always greater or equal to 0. A necessary condition for that $(IC_{\mathcal{U}})$ and $(IC_{\mathcal{T}})$ can both hold is therefore

$$\begin{aligned} \left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_u\left(\mathbf{w}_{\mathcal{U}}\right)\right) M_d \leq\ & \kappa \overline{B}_1 + \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{T}})} \\ & - \left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)(1-\epsilon) + \epsilon\right) \kappa B_h^{\alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{U}})} \\ & - \left(e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{T}}\right)\right) - e_2^*\left(\alpha_{\mathcal{T}}\left(\mathbf{w}_{\mathcal{U}}\right)\right)\right) \\ \leq\ & \kappa \overline{B}_1 + \kappa \overline{B}_h, \end{aligned} \tag{3.89}$$

where we used $0 \leq \alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{U}}) \leq \alpha_{\mathcal{T}}(\mathbf{w}_{\mathcal{T}}) \leq 1$ and $B_h^1 < \overline{B}_h$ for the last inequality. If the trustworthy agent would be fully separated from the untrustworthy, then Bayesian updating requires: $\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) = 1$ and $\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right) = 0$ and therefore $\left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right)\right) = 1$.

In the hybrid case in which the trustworthy agent chooses contract $\mathbf{w}_{\mathcal{T}}$ for sure and the untrustworthy agent is indifferent between both contracts and mixes with some probability we have $\alpha_{\mathcal{U}}(\mathbf{w}_{\mathcal{U}}) = 0$ and $\alpha_{\mathcal{U}}(\mathbf{w}_{\mathcal{T}}) \geq \sigma_{\mathcal{U}}$, hence $\left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathbf{u}}\right)\right) \geq \sigma_{\mathcal{U}}$.

In the hybrid case in which the untrustworthy agent chooses contract $\mathbf{w}_{\mathcal{U}}$ for sure and the trustworthy agent is indifferent between both contracts and mixes with some probability we have $\alpha_{\mathcal{U}}(\mathbf{w}_{\mathcal{U}}) \leq \sigma_{\mathcal{U}}$ and $\alpha_{\mathcal{U}}(\mathbf{w}_{\mathcal{T}}) = 1$, hence $\left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right)\right) \geq 1 - \sigma_{\mathcal{U}}$.

In any case holds $\left(\alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{T}}\right) - \alpha_{\mathcal{U}}\left(\mathbf{w}_{\mathcal{U}}\right)\right) \geq \max\{\sigma_{\mathcal{U}}, (1 - \sigma_{\mathcal{U}})\}$. Together with condition 3.89 we have shown that screening is impossible if

$$M_d > \frac{\kappa\left(\overline{B}_1 + \overline{B}_h\right)}{\min\{\sigma_{\mathcal{U}}, (1 - \sigma_{\mathcal{U}})\}}. \tag{3.90}$$

Mapping on the Reduced Form Contract Proposal

After writing down the expected payoffs for the different contracts for the (+)-principal, the (-)-principal, and the trustworthy agent we show in **step 1** that for any contract-proposal $\mu((\underline{w}_P, \overline{w}_P, \underline{w}_A, \overline{w}_A))$ there exist a payoff equivalent reduced-form contract (w, l, l_u). In **step 2** we show that for every reduced-form contract (w, l, l_u), there is at least one payoff equivalent contract-proposal $\mu((\underline{w}_P, \overline{w}_P, \underline{w}_A, \overline{w}_A))$ that gives at least the trustworthy agent the incentive to choose $e_1 = \overline{e}_1$. In **step 3** we show that (for a given $\alpha_{\mathcal{T}}$) the expected payoffs under two reduced-form contracts (w, l, l_u) and (w', l', l_u') are equal for the trusting principal, the distrusting principal, and the trustworthy agent only if $(w, l, l_u) = (w', l', l_u')$.

Within this subsection we only want to establish payoff equivalences for given beliefs $\alpha_{\mathcal{T}}$. Then neither the second investment e_2 nor the behavior in the trust game are influenced by the contract. For this section we can therefore simplify the analysis by re-normalizing each players utility by substracting the expected payoff resulting from investment e_2 and the trust-game. Then the expected (re-normalized) payoffs of an original-form deterministic contract $(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A)$ (with $\Delta w_A \equiv \overline{w}_A - \underline{w}_A$):

1. If $\Delta w_A > \overline{e}_1$ (complete contract: both type of agents: $e_1 = \overline{e}_1$):

$$\hat{V}_+(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = -\overline{w}_P + \overline{B}_1 \tag{3.91}$$

$$\hat{V}_-(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = -\overline{w}_P + \overline{B}_1 \tag{3.92}$$

$$\hat{U}_\mathcal{T}(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = \overline{w}_A - \overline{e}_1 \tag{3.93}$$

2. If $\overline{e}_1 \geq \Delta w_A > \overline{e}_1 - \overline{B}_1$ (incomplete contract: t-agent: $e_1 = \overline{e}_1$, u-agent: $e_1 = 0$):

$$\hat{V}_+(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = \pi_+(-\overline{w}_p + \overline{B}_1) - (1 - \pi_+)\underline{w}_P \tag{3.94}$$

$$\hat{V}_-(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = \pi_-(-\overline{w}_p + \overline{B}_1) - (1 - \pi_-)\underline{w}_P \tag{3.95}$$

$$\hat{U}_\mathcal{T}(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = \overline{w}_A - \overline{e}_1 \tag{3.96}$$

3. If $\Delta w_A < \overline{e}_1 - \overline{B}_1$ (both type of agents choose $e_1 = 0$):

$$\hat{V}_+(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = -\underline{w}_p \tag{3.97}$$

$$\hat{V}_-(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = -\underline{w}_p \tag{3.98}$$

$$\hat{U}_\mathcal{T}(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A) = \underline{w}_A - \overline{B}_1. \tag{3.99}$$

Notice any contract $(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A)$ of the last category $(\Delta w_A < \overline{e}_1 - \overline{B}_1)$ has always a corresponding payoff-equivalent contract $(\underline{w}'_p, \overline{w}'_p, \underline{w}'_A, \overline{w}'_A)$ in the first category $(\Delta w_A > \overline{e}_1)$, e.g. $\overline{w}'_P \equiv \underline{w}_p + \overline{B}_1$, $\overline{w}'_A \equiv \underline{w}_A + \overline{e}_1 - \overline{B}_1$ and $\underline{w}'_P \equiv \underline{w}'_A < \overline{w}'_A - \overline{e}_1$. Hence, we can restrict the analysis to contracts of categories 1 and 2. By "complete contract" we denote a contract of category 1 and by "incomplete contract" we denote a contract of category 2.

Now consider the general case of a probability distribution μ over a set of contracts of category 1 or 2. Let μ^c denote the total mass of complete contracts and therefore $1 - \mu^c$ the total mass of incomplete contracts. Let $\overline{w}_A$ denote the expected value of the $\overline{w}_A$ over all contracts. Furthermore, let $\underline{w}_P^i$ and $\overline{w}_P^i$ denote the expected values of $\underline{w}_P$ and $\overline{w}_P$ conditional on having an incomplete contract. Correspondingly, let $\overline{w}_P^c$ denote the expected values of $\overline{w}_P$ conditional on having a complete contract. Then the expected payoffs of a random contract-proposal $\mu(\underline{w}_p, \overline{w}_p, \underline{w}_A, \overline{w}_A)$ are

$$\begin{aligned}
\hat{V}_+(\mu) &= \mu^c(-\overline{w}_P^c + \overline{B}_1) + (1-\mu^c)\left(\pi_+\left(-\overline{w}_P^i + \overline{B}_1\right) - (1-\pi_+)\,\underline{w}_P^i\right) && (3.100)\\
\hat{V}_-(\mu) &= \mu^c(-\overline{w}_P^c + \overline{B}_1) + (1-\mu^c)\left(\pi_-\left(-\overline{w}_P^i + \overline{B}_1\right) - (1-\pi_-)\,\underline{w}_P^i\right) && (3.101)\\
\hat{U}_{\mathcal{T}}(\mu) &= \overline{w}_A - \overline{e}_1, && (3.102)
\end{aligned}$$

or equivalently

$$\begin{aligned}
\hat{V}_+(\mu) &= -\left(\mu^c\overline{w}_P^c + (1-\mu^c)\,\overline{w}_P^i\right) - (1-\pi_+)\,(1-\mu^c)\left(\overline{B}_1 - \Delta w_P^i\right) + \overline{B}_1 && (3.103)\\
\hat{V}_-(\mu) &= -\left(\mu^c\overline{w}_P^c + (1-\mu^c)\,\overline{w}_P^i\right) - (1-\pi_-)\,(1-\mu^c)\left(\overline{B}_1 - \Delta w_P^i\right) + \overline{B}_1 && (3.104)\\
\hat{U}_{\mathcal{T}}(\mu) &= \overline{w}_A - \overline{e}_1. && (3.105)
\end{aligned}$$

The expected (re-normalized) payoffs from a reduced-form contract (w, l, l_u) are for the (+)-principal, (-)-principal and the trustworthy agent are

$$\begin{aligned}
\hat{V}_+(w,l,l_u) &= -(w+l) - (1-\pi_+)\,l_u + \overline{B}_1, && (3.106)\\
\hat{V}_-(w,l,l_u) &= -(w+l) - (1-\pi_-)\,l_u + \overline{B}_1. && (3.107)\\
\hat{U}_{\mathcal{T}}(w,l,l_u) &= w - \overline{e}_1. && (3.108)
\end{aligned}$$

Step 1: For a given original-form contract proposal μ we choose $w \equiv \overline{w}_A$, $l \equiv \left(\mu^c\overline{w}_P^c + (1-\mu^c)\,\overline{w}_P^i\right) - \overline{w}_A$, and $l_u \equiv (1-\mu^c)(\overline{B}_1 - \Delta w_P^i)$. Then $l \geq 0$ and $l_u \geq 0$ and the payoffs are equal to the original-form contract for both types of principal and the trustworthy agent.

Step 2: For a given reduced-form contract (w, l, l_u) distinguish two cases

Case 1: If $l_u \geq \overline{B}_1 - \overline{e}_1$

we can choose the deterministic incomplete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$,

$\underline{w}_P \equiv w + l_u - \overline{B}_1$ and $\underline{w}_A \equiv w - \overline{e}_1 - \epsilon$ with $\epsilon \in]0, \overline{B}_1 - \overline{e}_1[$.

Case 2: If $l_u < \overline{B}_1 - \overline{e}_1$

we can choose a stochastic contract proposal mixing between two contracts: With probability μ^c the complete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$, $\underline{w}_P \equiv \underline{w}_A \equiv w - \overline{e}_1 - \epsilon$ with $0 < \epsilon < \overline{B}_1 - \overline{e}_1$ is drawn.

With the counter probability $(1 - \mu^c) \equiv \frac{l_u}{(-1)\overline{B}_1 + l + \overline{e}_1}$ the incomplete contract $\overline{w}_A \equiv w$, $\overline{w}_P \equiv w + l$, $\underline{w}_P \equiv w + \overline{B}_1 - \overline{e}_1$, and $\underline{w}_A \equiv w - \overline{e}_1 + \epsilon$ with $0 < \epsilon < \overline{B}_1$.

Step 3: Consider two reduced-form contracts (w, l, l_u) and w', l', l'_u with equal expected payoffs for (+)-principal, (-)-principal and (t)-agent. Then $w = w'$ due to equation 3.108 and $l' + (1 - \pi_+)(l'_u - l_u) = l = l' + (1 - P - -)(l'_u - l_u)$, due to equations 3.106 and 3.107. Since $\pi_+ \neq \pi_-$ this equations can only hold if $l'_u = l_u$ and $l' = l$, q.e.d.

Bibliography

[1] Aghion, Philippe, Alberto Alesina, and Francesco Trebbi (2004): Endogenous Political Institutions, *Quarterly Journal of Economics* 119(2), 565-611

[2] Aghion, Philippe, and Patrick Bolton (2003): Incomplete Social Contracts, *Journal of the European Economic Association* 1(1), 38-67

[3] Allen, F. and D. Gale 1994 " Financial Innovation and Risk Sharing" Cambridge: MIT Press

[4] Andreoni, J., Harbaugh, W. and Vesterlund, L. (2003),"The Carrot or the Stick: Rewards, Punishments, and Cooperation" *American Economic Review* Vol. 93, 3, 893-902

[5] Baron, David, and John Ferejohn (1989): Bargaining in Legislatures, *American Political Science Review* 83(4), 1181-1206

[6] Bénabou and Tirole 2003 "Intrinsic vs. Extrinsic Motivation", *Review of Economic Studies*, Vol 70(3), 489-521

[7] Benaim, M. and Weibull, J.W. (2003),"Deterministic Approximation of Stochastic Evolution in Games" *Econometrica* Vol. 71, 873-903

[8] Bernheim, B. Douglas and Whinston, Michael D. 1998 "Incomplete Contracts and Strategic Ambiguity" *American Economic Review*, Vol. 88, No. 4, 902-932

[9] Bergstrom, Theodore C. (2001), "The Algebra of Assortative Encounters and the Evolution of Cooperation", to appear in: *International Game Theory Review*

[10] Bergstrom, Theodore C. (2002), “Evolution of Social Behavior: Individual and Group Selection”, *Journal of Economic Perspectives* Vol. 16, 2, pp. 67-88

[11] Bester, Helmut and Güth, Werner (1998),“Is altruism evolutionary stable?”, *Journal of Economic Behavior and Organization* 34, 193-209

[12] Bolton, G. and Ockenfels, A. (2000), “ERC: A theory of equity, reciprocity and competition”, *American Economic Review*, 90, 166-193

[13] Bowles, Samuel and Gintis, Herbert (2004),“The evolution of strong reciprocity: cooperation in heterogeneous populations”, *Theoretical Population Biology*, 65, 17-28

[14] Boylan, R. (1992),“Laws of large numbers for dynamical systems with randomly matched individuals”, *Journal of Economic Theory* 57, 473-504

[15] Brennan, Geoffrey, and Alan Hamlin (1994): A Revisionist View on the Separation of Powers, *Journal of Theoretical Politics* 6(3), 345-368

[16] Buchanan, James (1991): Constitutional Economics, Oxford: Basil Blackwell

[17] Cho and Kreps 1987 “Signaling Games and Stable Equilibria”, *Quaterly Journal of Economics*, Vol. CII, Iss. 2

[18] Cooper, Ben and Wallace, Chris (2001), “Group Selection and the Evolution of Altruism”, *Oxford Discussion Paper Series*, Nr. 67

[19] Deci, E. and R. Ryan 1985 “Intrinsic Motivation and Self-Determination in Human Behavior”, New York: Plenum Press

[20] Dekel, Eddie, Jeffrey C. Ely and Okan Yilankaya, (2004), “Evolution of Preferences”, mimeo

[21] Ely, Jeffrey C. and Okan Yilankaya, (2001) “Nash Equilibrium and the Evolution of Preferences”, *Journal of Economic Theory* 97, 255-272

[22] Engelmann, D. 2000 “Trust and Trustworthiness - Theoretical Explanations and Experimental Evidence” Disseration, Shaker Verlag

[23] Epstein, David, and Sharyn O’Halloran (1996): Divided Government and the Design of Administrative Procedure: A Formal Model and Empirical test, *Journal of Politics* 58, 373-397

[24] Erlenmeier, Ulrich, and Hans Gersbach (2001): Flexible Majority Rules, *CESifo Working Paper* No. 464

[25] Eshel, Ilan, Larry Samuelson and Avner Shaked (1998) “Altruists, Egoists, and Hooligans in a Local Interaction Model”,*American Economic Review*, 88,1, 157-179

[26] Etzioni,A. 1971 “Modern Organizations”, Englewood Cliffs, N.J.: Prentice-Hall

[27] Falk and Kosfeld 2004“Distrust - The Hidden Cost of Control”, *IZA Discussion Paper* No. 1203

[28] Fehr, Ernst and Gächter, Simon (2000),“Fairness and retaliation: The economics of reciprocity”, *Journal of Economic Perspectives*, 14, 159-181

[29] Fehr, E., A. Klein and K.M. Schmidt 2004 “Contracts, Fairness and Incentives”, *Munich Economics Discussion Paper* 2004-07

[30] Fehr, E. and B. Rockenbach 2003 “Detrimental effects of sanctions on human altruism” *Nature*, Vol. 422, 13 March 2003

[31] Fehr, Ernst and Schmidt, Klaus (2000),“Theories of Fairness and Reciprocity - Evidence and Economic Applications”, (paper prepared for the invited session of the 8th World Congress of the Econometric Society)

[32] Frank, R.H. (1987),“If homo economicus could choose his own utility function, would he want one with a conscience?”, *American Economic Review* 77, 593-604

[33] Frey, B. 1997 “Not Just For the Money. An Economic Theory of Personal Motivation”, Edward Elgar, Cheltenham, 1997)

[34] Friedman, D. and Singh, N. (1999), "On the viability of vengance", UC Santa Cruz, Mimeo

[35] Gersbach, Hans (2002): Democratic Mechanisms: Double Majority Rules and Flexible Agenda Costs, *CESifo Working Paper*, No. 749

[36] Gintis, Herbert (2000), "Strong Reciprocity and Human Sociality", *Journal of Theoretical Biology* 206, 169 - 179

[37] Gneezy, U. and A. Rustichini 2000a "A Fine is a Price" *Journal of Legal Studies*, 29(1), Part 1, 1-17

[38] Gneezy, U. and Rustichini, A. 2000b "Pay Enough or Don't Pay at All", *Quaterly Journal of Economics*, 115(3), 791-810

[39] Grofman, Bernard, and Donald Wittman (eds.)(1989): The Federalist Papers and the New Institutionalism, New York: Agathon Press

[40] Güth, Werner (1995), "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives", *International Journal of Game Theory* 24, 323-44

[41] Güth, Werner and Yaari (1992), Menahem, "An evolutionary approach to explain reciprocal behavior in a simple strategic game," in: U. Witt(Editor), *Explaining Process and Change: Approaches in Evolutionary Economics*, Ann Arbor: The University of Michigan Press, 23-34

[42] Guttman, Joel M. (2003), "Repeated Interaction and the Evolution of Preferences for Reciprocity", *The Economic Journal*, 113, 631-656

[43] Hamilton, Alexander, James Madison, and John Jay [1788](1961): The Federalist Articles, New York: New American Library

[44] Hart, Sergiu (2000), "Evolutionary dynamics and backward induction", *Games and Economic Behavior* 41, 227-264

[45] Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath (2001), “In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies”, *American Economic Review*

[46] Höffler, Felix (1999),“Some play fair, some don’t. Reciprocal fairness in a stylized principal-agent problem”, *Journal of Economic Behavior & Organization*, Vol.38, 113-131

[47] Holmström, B. 1982 “Moral Hazard in Teams”, *Bell Journal of Economics*, Vol. 13, 324-340

[48] Holmstöm, B. and P.R. Milgrom 1992 “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design”, *Journal of Law, Economics and Organization*, 7, Special Issue, pp. 24-52

[49] Huck, S. and Oechssler, J. (1999),“The Indirect Evolutionary Approach to Explaining Fair Allocations”, *Games and Economic Behavior* 28, 13-24

[50] M. Kandori, G. Mailath and R.Rob (1993) “Learning, Mutation, and Long-Run Equilibria in Games.”, *Econometrica*, 61:29-56

[51] Kuzmics, C. (2003), “Stochastic Evolutionary Stability in Generic Extensive Form Games of Perfect Information”,*Games and Economic Behavior*, forthcoming

[52] Kuzmics, C. (2003),“Individual and Group Selection in Symmetric 2-Player Games”, Mimeo

[53] Laffont, Jean-Jacques (2000): Incentives and Political Economy, Oxford: Oxford University Press

[54] Laffont, Jean-Jacques (1995): Industrial Policy and Politics, *International Journal of Industry Organization* 12, 1-27

[55] Laffont, J. and J. Tirole 1993 “A Theory of Incentives in Procurement and Regulation”, Cambridge MA: The MIT Press

[56] Levine, D.K., (1998),"Modeling altruism and spitefulness in experiments", *Review of Economic Dynamics*, 1, 593-622

[57] Lijphart, Arendt (1992)(ed.): Parliamentary versus Presidential Government, Oxford: Oxford University Press

[58] Lijphart, Arendt (1994): Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies 1945-1990, Oxford: Oxford University Press

[59] Luhmann, N. (1968) "Vertrauen", 4th Edition 2000, Lucius & Lucius, Stuttgart

[60] Mailath, G. , S. Morris and A. Postlewaite 2001 "Laws and Authority", Mimeo

[61] Makin, E. and J. Tirole 1992 "The Principal-Agent Relationship with an Informed Principal, II: Common values", *Econometrica*, 60, 1-42

[62] Mas-Colell, Whinston and Green 1995 "Microeconomic Theory" Oxford University Press

[63] Maynard Smith, J. (1964), "Group Selection and Kin Selection", *Nature*, March 14, 201, 1145-147

[64] Messner, Matthias, and Matthias Polborn (2004): Voting on Majority Rules, *Review of Economic Studies* 71(1), 115-132

[65] Montesquieu, Charles de Secondat [1748](1991): The Spirit of the Laws, Littleton, Colo: Rothman

[66] Mueller, Dennis (1996): Constitutional Democracy, Oxford: Oxford University Press

[67] Nöldeke, G. and Samuelson, L. (1993), "An evolutionary analysis of backward and forward induction", *Games and Economic Behavior*, 5, 425-454

[68] Ok, E.A. and Vega-Redondo,F. (2001),"On the evolution of individualistic preferences: an incomplete information scenario", *Journal od Economic Theory*, 97, 231-254

[69] Olson, William J., and Alan Woll (1999): Executive Orders and National emergencies: How Presidents Have Come to "Run the Country by Usurping Legislative Power", *Policy Analysis* No. 358, October 28

[70] Persson, Torsten, Gérard Roland and Guido Tabellini (1997) 'Separation of Powers and Political Accountability', *Quarterly Journal of Economics*, 1163-1202

[71] Persson, Torsten, and Guido Tabellini (2000): Political Economics: Explaining Economic Policy, Cambridge, Mass.: MIT Press

[72] Persson, Torsten, and Guido Tabellini (2003): The Economic Effects of Constitutions, Cambridge, Mass.: MIT Press

[73] Price, G.R. (1970),"Selection and Covariance", *Nature* 277, 520-521

[74] Rawls, John (1971): A Theory of Justice, Cambridge, Mass.: Belknap Press of Harvard University Press

[75] Robson, Arthur J. (1990),"Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake", *Journal of theoretical Biology*, 144, 379-396

[76] Robson, Arthur J. and Fernando Vega-Redondo (1996), "Efficient Equilibrium Selection in Evolutionary Games with Random Matching", *Journal of Economic Theory*, 70, 65-92

[77] Romer, Thomas, and Howard Rosenthal (1979): Bureaucrats versus Voters: On the Political Economy of Resource Allocation by Direct Democracy, *Quarterly Journal of Economics* 93(4), 563-587

[78] Romer, Thomas, and Howard Rosenthal (1983): A Constitution for Solving Asymmetric Externality Games, *American Journal of Political Science* 27, 1-26

[79] Samuelson,L.(2001),"Analogies, Adaptation, and Anomalies", *Journal of Economic Theory*, 97, 320-366

[80] Schramm, Peter W., and Wilson, Bradford P. (1994)(eds.): Separation of Powers and Good Government, Boston: Rowman and Littlefield

[81] Schultz, Peter L. (1994): Congress and the Separation of Powers Today: Practice in Search of a Theory, in: Wilson, Bradford P., and Schramm, Peter W. (eds.): Separation of Powers and Good Government, Boston: Rowman and Littlefield, 185-200

[82] Sethi, R. (1996), "Evolutionary stability and social norms", *Journal of Economic Behavior and Organization*, 29, 113-140

[83] Sethi, Rajiv and Somananthan, E. (1996), "The Evolution of Social Norms in Common Property Resource Use", *American Economic Review* Vol. 86, 4, 766-788

[84] Sethi, Rajiv and Somananthan, E. (2001), "Preference Evolution and Reciprocity", *Journal of Economic Theory*, Vol. 97, 273-297

[85] Sethi, Rajiv and Somananthan, E. (2003), "Understanding Reciprocity", *Journal of Economic Behavior and Organization*, 50, 1-27

[86] Sober, E. and Wilson, D.S. (1998),"UNTO OTHERS - The Evolution and Psychology of Unselfish Behavior", Cambridge(M.A.): Harvard University Press

[87] Spier, Kathryn E. 1992. "Incomplete contracts and signalling" *RAND Journal of Economics*, Vol. 23, No. 3, 432-443

[88] Spence, A.M. 1974 "Market Signalling" Harvard University Press

[89] Sunnafrank, M. and A. Ramirez, Jr. 2004 "At the first sight: Persistent relational effects of get acquainted conversations", *Journal of Social and Personal Relationship*, Vol.21(3), 361-79

[90] Tabellini, Guido (2002): Principles of Policymaking in the European Union: an Economic Perspective, *CESifo Economics Studies*, Vol. 49, 1/2003, 75-102

[91] Tirole, Jean 1999 "Incomplete Contracts: Where Do We Stand?", *Econometrica*, Vol. 67, No. 4 741-781

[92] Voigt, Stefan (1997): Positive Constitutional Economics: A Survey, Public Choice 90, 11-53

[93] Voigt, Stefan (2003)(ed.): Constitutional Political Economy, Northampton: Edward Elgar

[94] Volden, Craig (2002): A Formal Model of the Politics of Delegation in a Separation of Powers System, *American Journal of Political Science* 46(1), 111-133

[95] Weibull, Jörgen W. (1995) “Evolutionary Game Theory”, MIT Press

Curriculum Vitae

13. Juli 1973	geboren in Münster, Westfalen
Juni 1992	Abitur am Kardinal-von-Galen-Gymnasium in Münster
Juli 1992 - Sep. 1993	Zivildienst im Kinderkrankenhaus (Lachnerklinik) in München
Nov. 1993 - April 1995	Vordiplom Physik, LMU-München
Sep. 1995 - Sep. 1996	Erasmus Stipendium, Universidad de Sevilla, Spanien
Mai 1995 - Dez. 1998	Hauptstudium Physik, LMU-München
Jan. 1999 - Feb. 2000	Diplomarbeit in Statistischer Physik an der LMU-München
seit Mai 2000	Doktorandenstudium in Volkswirtschaftlehre an der LMU-München
seit Sep. 2000	Wissenschaftlicher Mitarbeiter, Seminar für Wirtschaftstheorie, LMU-München
Oct. 2001 - Juli 2002	Marie Curie Fellow, Faculty of Economics and Politics, University of Cambridge, GB
9. Februar 2005	Promotion zum Doctor oeconomiae publicae