Kosfeld, Michael

Working Paper

The Role of Leaders in Inducing and Maintaining Cooperation: The CC Strategy

IZA Discussion Papers, No. 12540

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kosfeld, Michael (2019) : The Role of Leaders in Inducing and Maintaining Cooperation: The CC Strategy, IZA Discussion Papers, No. 12540, Institute of Labor Economics (IZA), Bonn

This Version is available at:
http://hdl.handle.net/10419/207366

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12540

The Role of Leaders in Inducing and Maintaining Cooperation: The CC Strategy

Michael Kosfeld

AUGUST 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12540

The Role of Leaders in Inducing and Maintaining Cooperation: The CC Strategy

Michael Kosfeld
Goethe University Frankfurt and IZA

AUGUST 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Role of Leaders in Inducing and Maintaining Cooperation: The CC Strategy*

I discuss recent findings from behavioral economic experiments in the lab and in the field on the role of leaders in human cooperation. Three implications for leadership are derived, which are summarized under the notion *CC strategy*. Firstly, leaders need to trust to not demotivate the motivated. Secondly, leaders need to punish to motivate the non-motivated. Finally, leaders shall (and can) attract motivated types. The discussion is embedded in a more general attempt to promote and stimulate interdisciplinary exchange of both methods and ideas in leadership research.

JEL Classification: C90, D90, M5

Keywords: leadership, cooperation, experiments

Corresponding author:
Michael Kosfeld
Chair of Organization and Management
Goethe University Frankfurt
Grüneburgplatz 1
60323 Frankfurt/Main
Germany

E-mail: kosfeld@econ.uni-frankfurt.de

* Based on the lecture given at the "Economics & Leadership" conference in Groningen, June 7-9, 2017; forthcoming in *The Leadership Quarterly*. I am grateful to three anonymous referees as well as the editors for helpful comments and suggestions.

Introduction

Leaders play an important role in society and organizations. Much of our knowledge of how leaders, as well as different forms of *leadership*, affect outcomes and decision making of individuals, groups, and society at large comes from important research in psychology and management, but there exists both classic and relevant new work on leadership in economics as well. A recent series of papers aims at reducing the knowledge gap between these disciplines by highlighting the common objective (why and how does leadership matter?) and comparing methodological advantages and disadvantages associated with various methods used in the different disciplines (e.g., lab and field experiments, surveys, statistical methods, quantitative and qualitative theory), also investigating potential complementarities. See, e.g., Antonakis et al. (2010), Bolton et al. (2013a), Hermalin (2013), and Zehnder et al. (2017).

This article aims at contributing to this important interdisciplinary exchange by analyzing and discussing the role of leaders in human cooperation. Based to a large part on my own work in this area, I develop what I call the *CC strategy*, which summarizes important evidence behavioral economists have accumulated in recent years, highlighting particular dimensions of how leaders can successfully induce and maintain cooperation in groups and organizations.

The basic organizational set-up I have in mind for analyzing the "problem of cooperation" is the following (see details in the next section). There exist gains from cooperation but collective action faces a dilemma in which self-interested individuals can benefit from free riding on the cooperative behavior of others. Contractual solutions – either implicit via relational contracts or explicit via legally binding agreements – are assumed to be limited or impossible. Thus, in terms of classic leadership instruments

neither transactional instruments nor relational incentives are available. There are many examples for such a set-up. E.g., teamwork in firms and organizations, where individual input is hard to measure or identify, project groups working together for a limited time as for example in open source programming, or common property management.

The main argument behind the *CC strategy* is then as follows. Empirical evidence, in particular from behavioral economic experiments, unambiguously shows that individual motivations for cooperation with others are heterogeneous. While some of us follow their own self-interest, which often goes against mutual cooperation due to free rider incentives, others are willing to cooperate voluntarily, and thus behave altruistically, even if this is individually costly. This behavioral heterogeneity, which is not primarily driven by differences in situational circumstances but as argued in detail below, indeed by differences in individual tastes, or – economically speaking – "revealed preferences", has important implications for leadership. Firstly, leaders want to make sure that if they interact with cooperatively motivated individuals they do not destroy this motivation, i.e., they *do not demotivate those who are motivated*. This requires trust, in particular trust in the other party's willingness to cooperate voluntarily. Secondly, since both type of motivations, self-interest and cooperative motivations, typically co-exist in groups, leaders need to sanction or punish non-cooperation to ensure that the cooperation of those who are willing to cooperate voluntarily is sustained. The reason is that most individuals who are willing to cooperate voluntarily cooperate only if others cooperate as well. Thus, leaders need to *motivate the non-motivated*. Finally, leaders have a strong interest in attracting voluntary cooperators. However, since true motivations are hard to identify – every applicant says he is a team player – the question is how leaders can achieve this. While there exists less empirical work on this issue so far (at least in the economics literature),

there do exist new theoretical results that highlight important mechanisms thereby suggesting particular leadership strategies that seem likely to be successful. Thus, leaders do have the possibility to *attract cooperative individuals*.

By focusing largely on evidence from behavioral economics in this article, I obviously take an "economic perspective" on the role of leadership. But see also the closely related work in psychology on leadership in social dilemma situations (e.g., Van Vugt and De Cremer, 1999; De Cremer, 2002; De Cremer and Van Knippenberg, 2002). The reason for doing so is not that I find other perspectives or evidence from other disciplines less convincing, but mostly that I am less well aware of it. Further, I believe that this economic perspective is actually not very different from what other social science researchers think about leadership. True, most economic models of leadership are rather abstract and often very simplistic (e.g., a leader simply being a party who decides first). However, my main impression is that this is more a matter of taste and differences in scientific approach than an expression of fundamental differences in what leadership ultimately *is*. For example, economists typically love to reduce the complexity of human decision-making as much as possible, whereas psychologists seem often much happier in keeping a substantial degree of this complexity as a valuable ingredient in their analysis. In fact, most economists would probably agree with Yukl's (2013) characterization of leadership as "a process whereby intentional influence is exerted over other people to guide, structure, and facilitate activities and relationships in a group or organization" (p. 18). Notable differences emerge once this characterization is implemented and operationalized in a concrete scientific model.

Finally, my personal view on fruitful interdisciplinary exchange (in particular in the social sciences) is that our main goal should *not* be that the different disciplines shall

all converge towards each other, trying to become "one large scientific discipline", but instead that we do well to keep a sufficient distance that enables everybody to see better the key elements and defining characteristics of one's own discipline as well as of the other disciplines together with their pros and cons. As George Bernard Shaw writes: "Do not let us fall into the common mistake of expecting to become one flesh and one spirit. Every star has its own orbit; and between it and its nearest neighbor there is not only a powerful attraction but an infinite distance. When the attraction becomes stronger than the distance the two do not embrace: they crash together in ruin."[1] Of course, this shall not mean that we should stop talking and listening to (as well as writing and reading) each other. On the contrary, I believe that only a continuous and deep interdisciplinary exchange is likely to move us forward and help each discipline make significant progress in its field.[2] This is where this article (hopefully) can contribute.

The remainder of the article is organized as follows. In the next section I define what I mean and understand by the problem of cooperation and I discuss the available evidence on the co-existence of heterogeneous motivations for cooperation based on behavioral experiments in the lab and in the field. I then explore three implications for leadership and develop what I call the *CC strategy*. Finally, I conclude discussing some open issues and the more general question what both behavioral economists and organizational psychologists may potentially draw from this research.

---

[1] Although Shaw in his play *The Apple Cart* speaks about the relationship between women and men, I believe that the quote nicely captures also (my view on) the relationship between different scientific disciplines.

[2] Fortunately, there exist a number of places where this interdisciplinary exchange takes place: For example, the *CLBO* in Frankfurt (www.clbo-frankfurt.de) and the center *In the Lead* in Groningen (www.rug.nl/inthelead/).

The problem – and the solution

CC – Cooperation is Conditional!

Game theory provides a powerful toolbox to study the problem of cooperation. The classic workhorse model invented by Al Tucker in 1950 to illustrate the social undesirability of Nash equilibrium (Kuhn et al., 1996) is the so-called Prisoners' Dilemma. Its payoff matrix is illustrated in Table 1.

	C	D
C	2,2	0,3
D	3,0	1,1

Table 1: Prisoners' Dilemma

In this game, there are two players who can either cooperate (C) or defect (D). In case both players cooperate, everybody earns a payoff of 2 (say, e.g., euros). If both players defect, everybody earns 1. If only one player cooperates and the other defects, the defecting player earns 3 and the cooperating player earns 0. As is easily seen, strategy D dominates strategy C because it gives a strictly higher payoff to an individual player independent of what the other player does. Therefore, in the unique Nash equilibrium of this game both players choose D. However, the resulting outcome (1,1) is inefficient as each player could earn twice as much – (2,2) – if everybody chose C. Thus, the Nash equilibrium is socially undesirable; individual payoff maximization does not lead to a social optimum. Put differently, the social dilemma of cooperation is that no player has an individual incentive to cooperate although joint cooperation maximizes social welfare. Every player is individually better off by choosing D even if the other cooperates, rather than choosing C as well.

An important assumption maintained in the above analysis is that every player tries to maximize his individual payoff and that no side contracts can be written, i.e., players are unable to sign binding agreements in which they commit themselves to mutual cooperation. Once such agreements are possible, be it via implicit contracting sustained by repeated interaction or via explicit pre-play negotiations that result in a legally binding treaty, the scope for cooperation of course increases. Without such possibility, however, no player can trust the other *assuming that everybody maximizes his own individual payoff*. Whether this assumption is correct or not, however, is an empirical question. In recent years, a large number of studies have analyzed to what degree human behavior is in line with this assumption and to what degree other concerns in particular with regard to the payoff and welfare of others are taken into account (see, e.g., Fehr and Schmidt, 2006; Dhamit, 2016; or Chapter 4 in Kagel and Roth, 2016 for excellent overviews).

In the case of the Prisoners' Dilemma and other social dilemma games with a similar incentive structure (e.g., public goods games and common pool resource games), the main empirical results are as follows. To illustrate, suppose the game in Table 1 is played sequentially (like in Miettinen et al., 2018). One player decides first whether to cooperate or defect and then the other player can make his decision contingent on the first player's choice. This modification leads to a new game form that is illustrated in Figure 1.

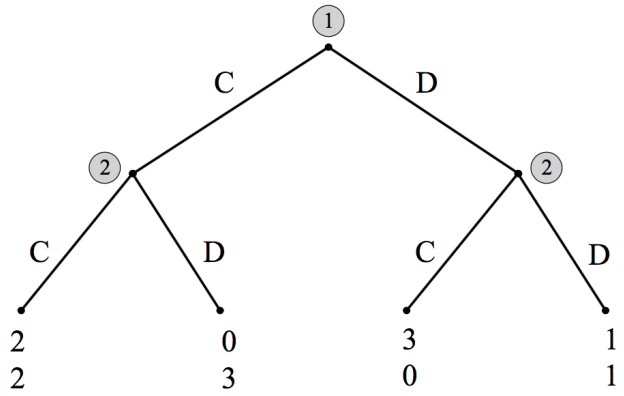

Figure 1: Prisoners' Dilemma played sequentially

Note that the incentive structure for the second player in the sequential game form is the same as before. Under the assumption of individual payoff maximization, player 2 chooses D independent of what player 1 does.[3] However, this is not what is found when the game is played in laboratory experiments. For example, in Miettinen et al. (2018) only 47 percent of the subjects in role of player 2 choose D independent of the first player's choice.[4] 38 percent choose D if player 1 chooses D but choose C if player chooses C. 9 percent always choose C, and 6 percent always choose the opposite of what player 1 does. See Figure 2.

Thus, while about half of the subjects behave in line with individual payoff maximization by always choosing D, thereby revealing a so-called *free-rider* preference, an almost equally large fraction of subjects reveal a preference for so-called *conditional*

[3] The situation for player 1 is trickier, because his optimal behavior depends on his belief about player 2's reaction. If player 1 believes that player 2 maximizes his individual payoff (i.e., always defects), choosing D is the optimal choice.

[4] In Miettinen et al. (2018), payoffs are given by 10 if both players defect, 30 if both players cooperate, and 50 and 5 if one player defects and the other player cooperates, respectively.

cooperation: they cooperate voluntarily but only if the other player cooperates as well. The precise shares of these two behavioral types may vary across experiments (Fischbacher et al., 2001; Kurban & Houser, 2005; Kocher et al., 2008; Herrmann & Thöni, 2009; Rustagi et al. 2010, Gächter et al., 2012), but a robust result from all of these studies is that free riders and conditional cooperators together represent the two main behavioral types: Both types are present and each with a considerable share in the overall population. The other two, empirically less relevant, types can be classified as *altruists* (always choose C) and *mismatcher*, or *contrarian,* (choose the opposite of what the other player chooses).

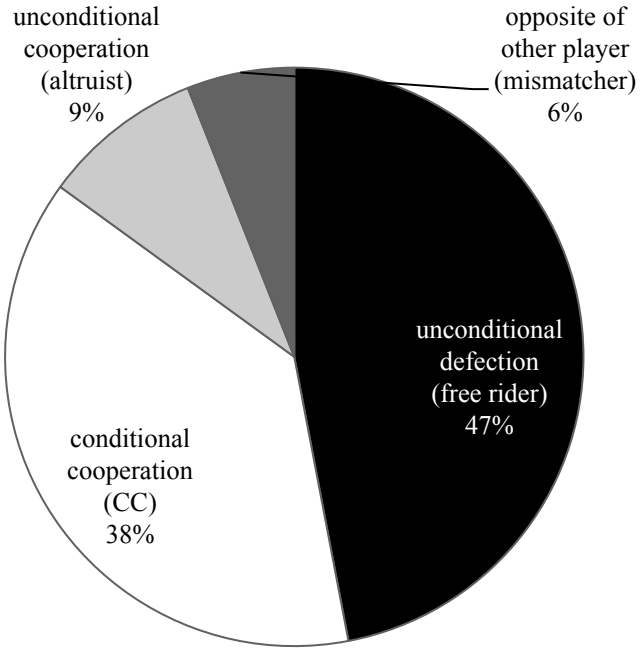

Figure 2: Distribution of revealed preferences in the Prisoners' Dilemma

(data from Miettinen et al., 2018)

What does this imply? On the one hand, the data show that some individuals do *behave exactly as "homo economicus" suggests*, i.e., they follow material incentives and choose to defect or free ride if they can. On the other hand, this is *not true for everybody*. There exists an important group of individuals who are willing to cooperate even if they have no material incentive to do so. Interestingly, van der Weele et al. (2014) show that in contrast to simple dictator-game giving (Dana et al, 2007) this form of *reciprocal motivation,* i.e., a motivation that is conditional on the behavior of others, is more robust to manipulations that give participants in an experiment the option to behave egoistically without allowing the other party to learn about their behavior (so-called "moral wiggle room"). Unfortunately, most economic research so far, including research on leadership, has focused more or less exclusively on the first type – the free rider (henceforth *FR*). See, e.g., the large contract theory literature in economics (Laffont and Martimort, 2002; Bolton and Dewatripont, 2005, Gibbons and Roberts, 2013) that is based on the important premise that basically *all* economic agents follow material incentives and minimize costs of effort. While this has produced relevant insights (Bolton et al., 2013a; Zehnder et al., 2017), the existence of the second type – the conditional cooperator (*CC* in the following) – as well as the interaction between the two types leads to a number of new implications for leadership. As I will argue in detail below, the existence of the *CC* type offers a powerful solution to the cooperation problem, yet only if leaders are prepared to take it into account. In what way this can be done is what I call the *CC strategy*. But before we get there, let us take a quick tour to Ethiopia.

Conditional cooperation (CC) in the field

Laboratory data is valuable because the lab offers a degree of control that is often unreachable in the field. Field data is valuable because it is typically much richer in context. A fruitful yet still relatively novel research strategy is to combine lab and field data in a way that allows researchers to take the best from both worlds: control *and* context richness. This research strategy may also open new possibilities to address the important issue of context in leadership research (cf. Liden and Antonakis, 2009; Dinh et al., 2014). In Rustagi et al. (2010) we asked ourselves whether the relevance of conditional cooperation, hitherto documented only in the lab, could also be identified in the field. As I will come back to the field set-up when discussing the *CC strategy*, let me provide some information about what we did and found in this study. For details, see Rustagi et al. (2010).

In 2000, the Ethiopian government together with the German *Gesellschaft für Internationale Zusammenarbeit (GIZ)* implemented a so-called participatory forest management program to fight deforestation in the Bale Mountains in Ethiopia. As a result of the program, around 50 forest user groups were formed by 2005 each receiving the exclusive right to govern a well-defined area of local forest as a common property. It was decided that every five years the number of "potential crop trees" (PCT) per hectare – an established measure of forest quality – would be assessed to provide an indicator for the performance of each group. The first round of data collected in 2005 reveal a high variation in forest management outcomes with outcomes ranging from a minimum of 13 to a maximum of 162 PCT per hectare (the average being 67). Since commons management is one of the classic examples of an important social dilemma, if not *the* example (cf. the "tragedy of the commons", Hardin, 1968), we investigated to what degree

variation in the presence of conditional cooperators can explain differences in forest management outcomes. To measure the share of *CC* and *FR* types in each group, we conducted a lab experiment in the spirit of the sequential game form described in Figure 1 with more than 600 members from all forest user groups.

In the experiment, two members from the same group play a one-shot linear public goods game, where each player can contribute between 0 and 6 Ethiopian Birr to a public good. Denoting by C_i the contribution of player i (i = 1, 2) to the public good, the payoff (in Birr) of player i is defined as

$$\Pi_i = 6 - C_i + 0.75\ (C_1 + C_2). \qquad (1)$$

Because 0.75 < 1, the marginal individual return of contributing to the public good is negative. Hence, the dominant strategy for each player is not to contribute. However, since 2*0.75 > 1, the marginal social return of contributing to the public good is positive. Both players together are better off if a player contributes. Suppose, for example, that no player contributes, then everybody earns 6 Birr. However, if both players contribute their full endowment, everybody earns 9 Birr.

In our experiment, the public goods game was played anonymously, so players knew that the other player was from their group but they did not know the identity of the other player. Importantly, each player made two decisions: a conditional decision, where a player's own contribution is made contingent on the contribution of the other player (just like the second player in the sequential game form in Figure 1), and an unconditional decision (similar to the first player in Figure 1). After every player has made his choice, for one of the two players the unconditional decision is taken and for the other player the conditional decision is taken (evaluated at the particular unconditional choice of the first player) and payoffs are calculated accordingly. For whom of the two players the

unconditional decision and for whom the conditional decision is taken is randomly determined.

This design allows us to identify each group member's revealed preference. In particular, the conditional decision tells us whether a player reveals himself as a free rider type *FR*, who contributes zero independent of the contribution of the other player, or as a conditional cooperator *CC*, whose contribution correlates positively with the other player's contribution, i.e., who contributes more to the public good the more the other player contributes. Overall, our results show that about 45 percent of the overall population of forest users are of the *CC* type while about 11 percent are of the *FR* type. Importantly, these shares vary significantly *across* groups, hence the question is: Do groups with more *CC* types achieve better forest management outcomes in terms of PCT? The answer is, yes! Ceteris paribus, a 10 percent increase in the share of *CC* types is associated with a significant increase in forest management outcomes by about five PCT per hectare on average. Similarly, a 10 percent increase in the share of *FR* types comes with a significant decrease of about seven PCT per hectare on average.

These results from the forest management context document two important findings: first, free rider and conditional cooperator types (identified by a controlled lab experiment) constitute, once again, a large share of the overall population; second, their individual shares in a group matter for group cooperation outcomes (identified in a natural field context). In the following, I will discuss the implication of these findings for leadership.

Implications for leadership: *The CC strategy*

In sum, what the empirical evidence on cooperation shows is that groups consist of both *motivated* and *non-motivated* types. Motivated types cooperate voluntarily, even if cooperation is individually costly; however, they cooperate only if others cooperate as well. Cooperation is Conditional – *CC*! Non-motivated types free ride, despite this imposing a negative externality on others. On the one hand, this message may sound hardly revolutionary, as probably everybody has made some personal experience with any or both of these types in one form or another. On the other hand, as mentioned already above it is a mere fact that most economic research on leadership so far has focused almost entirely on the second of these types: the non-motivated, or *FR* type. An exception is Rotemberg and Saloner (1993), who analyze the role of empathy (yet, on the part of the leader) assuming that the leader cares about the welfare of the follower. Besides this, most leadership studies in economics indeed take for granted that decision-making (on the part of both follower and leader) is predominantly guided by material incentives (Zehnder et al., 2017). It therefore presents an open question whether and if so how, the consideration of the motivated, or *CC* type matters for leadership research and applications.

Why should it not? Or figuratively speaking, why isn't the motivation of the *CC* type not just like some extra quantity of water in an otherwise empty glass? Why should leadership need to take it into account? If anything, since the glass is not entirely full, i.e., it is a rare case that all agents are fully motivated and motivations are perfectly aligned, instruments that increase and coordinate motivations are probably needed and this is exactly what classic economic research as well as, for example, transactional leadership theory has focused on (e.g., incentive pay, promotion plans, etc.). The answer is that the cooperative motivation of the *CC* type (the extra quantity of water in the otherwise empty

glass) might be there but as is argued below, the motivation can be very volatile. Leaders simply cannot take it for granted and classic incentives may even backfire, the very reason being that *CC* motivation is *conditional on the behavior of others*.

In what follows, I will lay out three implications for leadership that are direct consequences of the existence of the *CC* type, and that I therefore call the *CC strategy*:

1. Trust – to not demotivate the motivated
2. Punish – to motivate the non-motivated
3. Attract *CC*s – if you can

Trust – to not demotivate the motivated

Trust is essential for successful leadership. It is the first part of the *CC strategy*. The main reason why trust is key is the conditionality of cooperative behavior of the *CC* type. Cooperative agents cooperate only if others cooperate as well. If others (including leaders) defect, joint defection is the outcome, and this even if everybody was actually cooperatively motivated. In consequence, trust is necessary to sustain cooperation, because it upholds the (equilibrium) belief that others cooperate, too.

To illustrate, suppose a leader interacts with a follower in the sequential Prisoners' Dilemma in Figure 1. The leader moves first, the follower second. Suppose further that the follower is a conditional cooperator, i.e., a *CC* type. Only if the leader trusts in this game, i.e., cooperates himself because he beliefs that the follower will cooperate, cooperation can be sustained. Player 1 chooses C and player 2 follows. However, if the leader distrusts and decides to choose D, because he (wrongly) beliefs that the follower was a free rider, the follower responds by defecting as well: DD is the outcome. The situation reveals a very important general element of leadership: *signaling*. By acting in a particular manner,

leaders signal their belief about how they consider the situation to look like, in this case how they expect player 2 to behave (Sliwka 2007). They thus either implicitly or explicitly communicate a message that influences followers' behavior, with significant, sometimes surprising, implications on organizational outcomes. One such implication is that a leader's prior belief about the behavior of others may turn into a so-called *self-fulfilling prophecy*: independent of whether the leader's prior belief is actually correct, due to follower's responses the belief turns out to be correct ex post (even if it is false ex ante).

Several economic papers have studied this question in the above described, so-called *leading-by-example* framework. The classic paper is Hermalin (1998). Bolton et al. (2013b) consider the effect of leader signaling in situations that involve follower coordination. For experimental work see, e.g., Clark and Sefton (2001), Güth et al. (2007), Potters et al. (2007), Gächter et al. (2012), Drouvelis and Nosenzo (2013), and Gächter and Renner (2014). These papers document two important things: Firstly, sequential decision-making in the form of one *leader* deciding first and one or more *follower(s)* deciding subsequently helps, i.e., cooperation rates are typically higher compared to when decisions are made simultaneously. Secondly, leader behavior matters. Leaders, who do not cooperate themselves, see less cooperation from followers than leaders, who cooperate and thereby send a signal that cooperation is expected.

The question how leader behavior more generally signals beliefs and how these beliefs in turn influence follower reactions in the context of trust has been analyzed by Falk and Kosfeld (2006) in a novel principal-agent game. The principal-agent framework is, perhaps, the classic workhorse model in economics focusing on the important role of incentives, a key instrument also of transactional leadership (Gibbons and Roberts, 2013; Zehnder et al., 2017). The main elements of this framework are that a so-called *principal*

can decide about a so-called *agent's* incentives to provide effort, where effort is beneficial to the principal but costly to the agent. Depending on various conditions of the situation (e.g., with regard to the information both parties have at the time of contracting, the length of the relationship, etc.), the central question is, how incentives can and should be set such that economic welfare is maximized. In the game we implemented in Falk and Kosfeld (2006), a principal interacts with an agent in a one-shot encounter, where the agent can decide how much of his private resources to invest in a project that is costly to the agent but beneficial to the principal. More precisely, the agent has an endowment of 120 experimental units from which he can invest x into a project that gives $2x$ to the principal and costs the agent x. The principal's endowment is zero. As in other economic experiments, experimental units are exchanged into money at the end of the experiment. The two parties' payoff functions are thus

$$\Pi_a = 120 - x \text{ for the agent,} \qquad (2)$$

$$\Pi_p = 2x \text{ for the principal.} \qquad (3)$$

Before the agent decides about x, the principal can determine the agent's choice set. In particular, he can decide whether to impose a binding minimum investment $\underline{x} > 0$ the agent has to comply with (the level of which is exogenously given by the experimenter). In this case, the agent's choice set is equal to $[\underline{x}, 120]$, i.e., the agent can invest more but he cannot invest less. Alternatively, the principal can leave the agent's choice set unaffected, in which case the agent can choose any value x between 0 and 120.

Note that this simple game captures in a nutshell various key elements of typical principal-agent relationships: The efficient outcome (the outcome that maximizes the sum of all parties' payoffs) requires behavior on the part of the agent that is in the interest of the principal but not in the interest of the agent (here: positive investment x). This conflict

of interests typically leads to inefficient results (here: low x). The principal therefore wants to make use of available instruments to align the agent's incentive in order to reach a better outcome (here: a minimum level $\underline{x}$). Often such instruments still leave a lot of freedom to the agent, so that the outcome the principal can ensure is typically only second best. For example, the principal can fix working hours from 9 to 5 and monitor the agent's actual working time via an employee card to identify misbehavior. Still, the agent can decide how much effort to invest when being present at work.

Why did we think this is an interesting game that tells us something relevant about leadership? The reason is the following: Under the assumption that the agent maximizes his individual payoff, his optimal choice is $x = 0$. In this case, imposing a minimum $\underline{x}$ is the best option for the principal ensuring him a payoff of $2\underline{x}$. This is the equilibrium outcome predicted by classic economic reasoning based on the *non-motivated* agent type. Now, suppose the agent did not maximize his individual payoff but was instead *motivated* to invest some positive $x_m > \underline{x}$ voluntarily. For example, the agent could be fair-minded considering the unequal starting position in the game, in which the player in the role of the agent has all and the player in the role of the principal has nothing.[5] Alternatively and closer to firm contexts, the agent might care intrinsically about the project and invest effort even in the absence of material incentives. What should such an agent infer if the principal imposes a minimum $\underline{x}$? Clearly, as the implementation of $\underline{x}$ makes sense only if the principal does not believe the agent to choose $\underline{x}$ or more, the agent is likely to conclude exactly this: the principal does not trust the agent but expects him to be non-motivated. Yet, if the agent cares about his (self- and/or social-) image of being a motivated type, he may be unwilling to act in the interest of such a principal. In consequence, he may save

[5] In the game described, this would be a choice of $x = 40$ resulting in a payoff of 80 for both parties.

his effort costs and chooses $\underline{x}$ instead of x_m. Thus, distrust by the principal triggers non-trustworthiness by the agent. Trust may therefore be the better choice (Ellingsen and Johannesson, 2008).

The results in Falk and Kosfeld (2006) support this hypothesis. For example, in one of our main treatments, in which $\underline{x}$ is exogenously set equal to 10, the data show that 68 percent of the agents are of the motivated type, i.e., they voluntarily invest more than 10 even though this is materially costly (see Figure 3). About a third of them (24 percent) invest even the payoff-equalizing amount $x = 40$. However, a substantial fraction of these motivated types invest more than 10 only if the principal does not impose the minimum of 10, in other words, *only if the principal trusts them to be trustworthy*. If instead the principal distrusts and imposes the minimum of 10, many of these motivated agents choose $x = 10$. At the same time, 32 percent of the agents reveal to be non-motivated, i.e., they choose less than 10 if they can. If the principal imposes the minimum on them, they also choose 10, because they have to. Thus, principals in the experiment face, once again, a heterogeneous environment of both motivated and non-motivated types. While incentives in form of a minimum requirement induce non-motivated agents to invest more (just as classic economic analysis predicts), they backfire with regard to the motivated type.

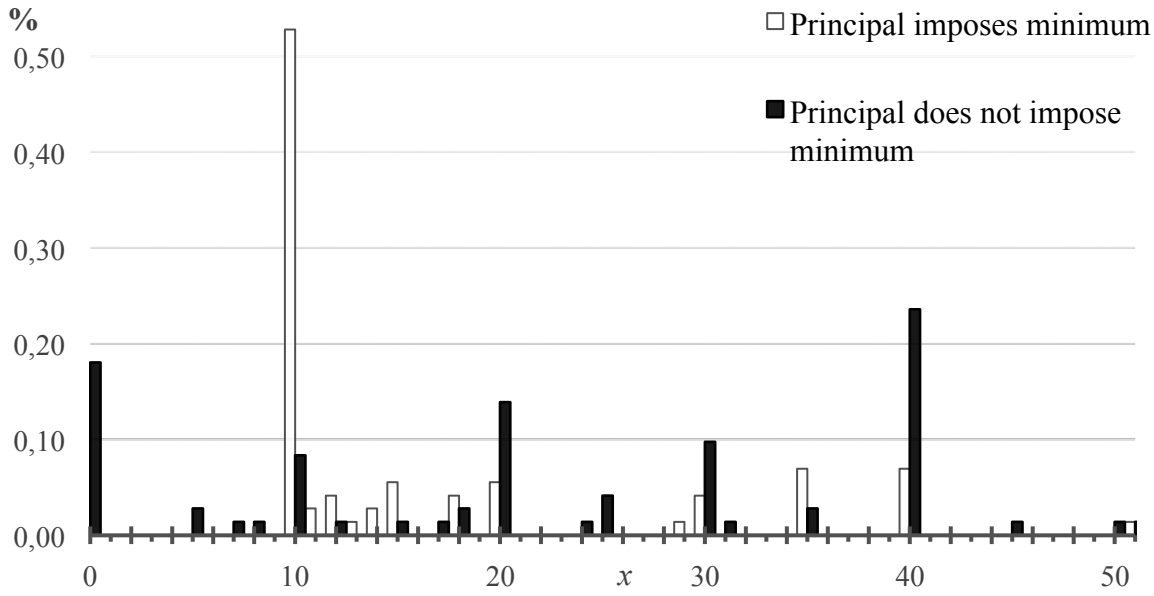

Figure 3: The hidden costs of distrust

(data from the C10 treatment in Falk and Kosfeld, 2006)

In Falk and Kosfeld (2006) the data show that, on average, the principal is better off if he trusts than if he imposes the minimum. In the latter case, he earns 35 experimental units on average; if he trusts, he earns 46 units on average, i.e., 30 percent more. Do participants in the role of the principal anticipate this? 71 percent do, the others don't.

Why do almost a third of the principals decide not to trust but impose the minimum thereby earning a significantly lower expected payoff? Intuition suggests that these principals may have pessimistic beliefs, i.e., they do not expect the agent to be motivated. Our results show that this indeed the case. Asked about how much they expect the agent to invest conditional on their own (i.e., the principal's) choice, basically every participant in the role of the principal reports subjectively rational beliefs: Participants, who decide to impose the minimum, expect on average that the agent invests a bit more

than 10 if they impose the minimum and that he invests less than 10 if they do not impose the minimum. Principals, who decide to trust, hold the same belief in case they impose the minimum, but expect the opposite in case of trust. In this case, they expect the agent to invest significantly more than 10. They thus believe the agent to be motivated.

Let us stop for a second and re-state this. Although all participants in the role of the principal face *exactly the same situation* in the experiment, we observe a remarkable heterogeneity with respect to the belief about the other party's motivation: principals, who believe the agent to be non-motivated, impose the minimum; principals, who believe the agent to be motivated, trust. It is as if the two groups of principals see the world through different glasses, a trusting and a distrusting one. Unfortunately, in the context of our study we can only speculate where these differences in prior beliefs come from (e.g., personal experience or a training in neoclassical economics), but the point I want to make is that not only principals' prior beliefs differ but – more importantly – they trigger different behavior on the part of the agent, each justifying the principal's belief ex post, even if the belief is ex ante false. Beliefs thereby become a self-fulfilling prophecy (Luhmann, 1968).

A number of follow-up studies have analyzed the "hidden costs of control" documented in Falk and Kosfeld (2006) in various scenarios. See, e.g., Gerlach (2008), Schnedler and Vadovic (2011), Ziegelmeyer et al. (2012), Burdin et al. (2015), Riener and Wiederhold (2016), and Kessler and Leider (2016). One important finding from these studies is that trust does not always pay financially, i.e., the principal is not always better off on average waiving available instruments of control. This finding should not come as a surprise as trust can, of course, only pay if the expected motivation on the agent's side is high enough. If motivation is low, blind trust is naïve, because it will be exploited too

often. This is also seen in Falk and Kosfeld (2006), where in another treatment the principal can compensate the agents' investment costs with a fixed wage (and in addition can decide to impose a minimum or not). If the chosen fixed waged is low, the agent's average motivation is low, as well. In this case, trust (i.e., not imposing a minimum) does not pay off. If the fixed wage is higher, however, the agent's motivation is also higher. At some point, the negative effect of not trusting on motivated agents dominates the positive effect on the non-motivated. Then trust starts to pay.

Punish – to motivate the non-motivated

In the previous section I argue that trust is an essential element of leadership, because it communicates beliefs that sustain cooperation. In this section I argue, perhaps at first glance counterintuitively, that punishment – in a sense a contrary of trust – is important, because it similarly upholds the belief of cooperation. The rationale is indeed the same. Whereas trust can sustain the belief of cooperation in vertical relationships (between a follower/agent and a leader/principal), punishment of non-cooperation can uphold beliefs in horizontal relationships (between followers/agents), that is, in teams, groups, and communities.

Recall the interaction between *CC* and *FR* types in cooperation problems such as team work, commons management or, more abstractly, any Prisoners' Dilemma-like situation. To fix ideas, suppose that we can describe the situation as a simultaneous two-player cooperation problem, in which one of the two players is a *CC* type and the other is a *FR* type. What outcome shall we expect behavior in this game to settle on? The answer is clear. Since the *CC* type cooperates only if the other player cooperates and the free rider defects unconditionally, mutual defection is the unique Nash equilibrium in this situation.

Thus, although one of the two players is cooperatively motivated, cooperation is no equilibrium outcome. In fact, mutual defection also arises as an outcome even if *both* players are of a *CC* type but everybody has pessimistic beliefs, i.e., each player expects the other to not cooperate. In this case, the situation has multiple Nash equilibria, and while prospects of cooperation are certainly better compared to the first case in the sense that mutual cooperation is now an equilibrium outcome, the players' problem is still to coordinate on this equilibrium. Depending on the payoffs in the game, this may or may not be an easy task (I will come back to this below). The analysis shows that sometimes trust, i.e., cooperative beliefs, is also needed in horizontal relationships. However, this is true only if both players are of the *CC* type.

What these arguments show is that even if there are *CC* types in a group, this does not guarantee mutual cooperation. (A variation of this proposition is that even if there is no cooperation, this does not mean that there are no *CC* types.) What is needed in heterogeneous groups is the possibility to enforce cooperation by the punishment of non-cooperation. Only then *FR* types will be motivated to cooperate and consequently *CC* types will cooperate (voluntarily) as well. In the Ethiopian forest management set-up discussed above this is achieved by voluntary forest patrols group members organize to monitor free rider behavior. Our data show that *CC* types participate significantly more often in these activities than *FR* types (Rustagi et al. 2010). Thus, the positive effect of *CC* types on forest management outcomes in this field case is based not only on *CC* types being willing to cooperate voluntarily, but also revealing a higher motivation to monitor free riding and thus contribute to the successful enforcement of cooperation in this set-up.

A number of experimental studies have shown that second-order punishment, i.e., the possibility for group members to punish each other, can indeed be a real "game

changer" in the sense that it transforms environments in which zero or little cooperation is the norm without punishment opportunity, to highly cooperative environments when punishment is possible (e.g., Fehr and Gächter, 2000; Gürerk et al., 2006, 2014). An important question in these studies is, why do people actually punish. If punishment is individually costly, the mere opportunity to punish others should not imply that it is actually used, nor that it is used effectively or efficiently. Several studies have therefore also pointed to the limits of second-party punishment (Nikiforakis, 2008; Nikiforakis et al., 2012; Herrmann et al, 2008). One possible way out is to centralize punishment by putting it in the hands of a leader. The question is then whether groups are willing to transfer authority voluntarily, and whether leaders are able and motivated to use punishment effectively. Both issues are obviously linked to each other. The first question has been analyzed by a number of recent experimental papers documenting that institutional change in the form of an endogenous implementation of centralized punishment institutions is well possible and welfare improving (e.g., Tyran and Feld, 2006; Kosfeld et al., 2009; Markussen et al. 2014). The second question with regard to a leader's motivation to punish has been less studied so far. O'Gorman et al. (2009) show in the lab that leader punishment can in principle be an effective instrument to maintain group cooperation, but they do not study variation in the motivation to punish. Kosfeld and Rustagi (2015) analyze this question in the context of the Ethiopian forest management case study. The context provides us with a unique possibility to investigate the role of leader punishment, and in particular a leader's motivation for punishment, on group cooperation outcomes. While group members are responsible for the monitoring (via forest patrols), it is the leader of a group who decides about the punishment of free riding or, more generally, norm violations.

But, how can we measure the motivation of a leader to punish norm-violating behavior? One approach could be to look at *actual* punishment data in the field, by collecting information about cases where group members violated some local norm and analyzing if and how the leader punished group members in these cases. The problem with such an approach is that actual punishment and the incidence of norm violations are statistically jointly determined. In groups, where the leader is known, for example, to be a tough punisher violations are less likely to occur, and hence there will also be little punishment, than in groups where the leader is known to punish only little or not at all. The problem would be similar to estimating the effect of police on crime rates, knowing that cities with a high crime rate typically (need to) invest more in policing. A different approach is to elicit a leader's motivation by his revealed preference for punishment in an experimental game. This is the way we took in this paper (Kosfeld and Rustagi, 2015).

We invited leaders of all forest user groups to participate in a third-party punishment game that consists of two stages. In the first stage of the game, two members of the leader's group participate in a linear public goods game of the same type as introduced above: each group member can contribute (this time simultaneously) up to six Ethiopian Birr to a public good with each member's payoff being defined by equation (1). In the second stage, the leader can now punish each individual member depending on members' contributions to the public good. Precisely, he can allocate so-called "deduction points" to each of the two members. Each deduction point costs the leader 1 Birr and reduces a member's payoff by 3 Birr. To finance his decision, the leader receives an endowment of 10 Birr.

Importantly, we did not simply "play the game" sequentially but we asked the leader to make his punishment decision for *all possible outcomes* of the first stage without

knowing what contributions the two group members actually choose. To keep such a decision manageable, we restricted each group members' choice set in the first stage to the set {0, 2, 4, 6}. Thus, for each pair of contributions from this set, we asked the leader to decide how much he wanted to punish each of the two members. Then, after group members and the leader had made their decisions, payoffs were realized based on these decisions. In this way the punishment decision of the leader becomes payoff relevant and is not purely hypothetical.

Note that with this elicitation method we obtain a valid measure of a leader's revealed preference for punishment. This would not have been the case if we had simply played the game sequentially, i.e., group members had decided first and then the leader had decided after having observed the particular group members' decisions. Firstly, the leader would have reacted only to one outcome of the first stage. Secondly, and more importantly, this outcome would have been affected by group members' anticipation of the leader's punishment. Thus, punishment and the outcome to which punishment is observed would, again, have been jointly determined. A proper comparison of punishment patterns across leaders would not be possible.

What type of revealed preferences for punishment can we expect in this third-party punishment game? Since punishment is costly to the leader (each deduction point allocated to a group member costs the leader 1 Birr), material motives can be ruled out. Leaders, who want to maximize their monetary earnings, don't punish. Further, we can rule out reputational motives, as we made sure in the experiment that group members did not learn the actual punishment decisions of their leader.[6] This also protected a leader

[6] We achieved this by paying out all the money a participant earned in the experiments at the end of the study so that nobody could deduce individual decisions in any of the experiments.

from possible reactions from group members after the experiment. Based on the experimental literature, we hypothesized that there exist three possible punishment motives:

- Efficiency motive: A leader punishes contributions that are inefficient, i.e., that do not maximize the total payoff of the two group members. This requires punishment of all contributions less than six Birr.
- Equality motive: A leader punishes contributions that generate payoff inequality. In this case, the leader punishes the group member, who contributes less than the other member.
- Antisocial motive: A leader punishes even if neither of these two norms is violated, i.e., group members are punished even if they contribute the maximum amount of six Birr.

Our results in Kosfeld and Rustagi (2015) show that the majority of leaders (29 out of 51) do not punish at all. These leaders thus reveal a money-maximizing preference, which was also emphasized by leaders when we asked them about their reasoning in making decisions (e.g., one leader said: "I prefer to have money in my pocket."). 14 leaders (27.5 percent) reveal an equality motive when punishing group members, i.e., these leaders punish members who contribute less than the other but do not punish when both members contribute equally (a statement was, e.g.: "Make payoffs nearly equal."). Four leaders (7.8 percent) punish in case of inequality and in addition also punish if group members contribute equally but less than six Birr. They thus reveal an additional motive for efficiency. Finally, four leaders (7.8 percent) punish antisocially: they punish players even if they contribute six Birr to the public good. When asked about their reasoning, these leaders stated, e.g., that, "it is so much fun to reduce income".

The results document, once again, an important heterogeneity in participants' behavior in the experiment, this time involving "natural" group leaders. Despite facing the exact same experimental situation, leaders behave very differently in the third-party punishment game, revealing both a classic money-maximizing motive (similar to a *non-motivated* type) and intrinsic motives – prosocial (equality and efficiency driven) as well as antisocial. Intriguingly, this heterogeneity observed in the game is correlated with ratings group members gave us on their leader in an independent household survey. Here, antisocial leaders are significantly more likely to be rated as a "bad leader".

The key question is whether leader types, in terms of revealed punishment motives in the behavioral game, make any difference for group cooperation outcomes in the field, in terms of average PCT per hectare. Our results show that this is indeed the case. Table 2 summarizes the results of linear regressions with average PCT per hectare in a group as the dependent variable on group leaders' types (weighted by the level of punishment), first without any controls (column 1), then with group level controls (column 2), village fixed effects (column 3), and finally with additional leader controls (column 4). The benchmark leader type in all regressions is the money-maximizing leader who does not punish in the game.

As Table 2 shows, leaders who reveal an antisocial motive are associated with a significantly worse group performance. The effect size is 20 PCT per hectare on average. Leaders who punish inequality and inefficiency have groups with significantly higher PCT per hectare, with an effect size of 29 PCT per hectare on average. Both effects are large given that average performance of all groups is 67 PCT per hectare. Interestingly, no significant association can be found for leaders who reveal an equality motive alone. One potential explanation is that punishment of inequality does not push groups who

coordinate on low-cooperation outcomes towards higher cooperation levels and higher efficiency. If everybody cooperates on the same low level, there is no inequality. Hence, leaders who do not punish in this case will not exert any influence on groups in terms of aiming to reach higher cooperation and efficiency levels. In a sense, the situation resembles that of a coordination game. See Weber et al. (2001), Brandts and Cooper (2007), and Brandts et al. (2015) on the role of leadership in such settings. Another explanation is that groups with equality-motivated leaders may need more time to reach higher cooperation outcomes. Our results in Kosfeld and Rustagi (2015) based on second-round forest assessments support this view, as groups with equality-motivated leaders are shown to eventually see higher group cooperation outcomes compared to non-punishing leaders.

Dependent variable: Average PCT per hectare

	(1) No controls	(2) Group level controls	(3) Village fixed effects	(4) Leader controls
Equality motive	-1.186	0.097	-0.638	-0.484
	(1.896)	(1.460)	(1.303)	(1.259)
Equality & efficiency	3.200*	2.349**	2.494***	2.494***
motive	(1.595)	(0.898)	(0.875)	(0.827)
Antisocial motive	-9.795***	-6.834***	-7.404***	-8.355***
	(3.315)	(1.809)	(2.396)	(2.329)
N	51	51	51	51
Adj. R^2	0.11	0.74	0.77	0.78

Table 2: Leader types and group cooperation outcomes (Kosfeld and Rustagi 2015)

Let me conclude this section with a few remarks. Recall that group members in the first stage of the third-party punishment game know that their leader has the possibility to punish them in the second stage. Using our data on group members' types (*CC* and *FR*), we find that *CC* types contribute significantly less to the public good, if their leader is of an antisocial type. This corroborates the negative association observed in the field data in Table 2 by documenting a similarly negative effect on cooperation outcomes in the

behavioral game. Next, the incidence of antisocial punishment we find in the Ethiopian context is actually not very different from the incidence of antisocial punishment in different western locations, where similar experiments have been conducted, but actually much lower than in other locations in the world (see Herrmann et al., 2008). While the ultimate determinants of antisocial punishment, or antisocial motives more generally, are still far from understood, the available evidence suggests that these motives are clearly present and there exists a large heterogeneity across locations and contexts.

Finally, an alternative instrument to the punishment of non-cooperation may be the reward of cooperation. Interestingly, empirical studies show that, at least in the context of cooperation, the effectiveness of rewards is not the same (e.g., Gürerk et al., 2014; Homonoff, 2018). While the overall discussion seems still unsettled, one important difference between punishments and rewards is that effective punishment does not need to be executed in equilibrium, as it successfully deters free riding, while rewards have to be paid. Thus, rewards may be more costly in equilibrium. Further, punishment seems indeed more frequently be used in the context of norm enforcement (like, e.g., cooperation; see Balliet et al., 2011) while rewards are particularly important, for example, to motivate innovative and explorative behavior (Manso, 2011).

Attract CCs – if you can

The first two leadership dimensions of the *CC* strategy rest on the co-existence of cooperative and non-cooperative types observed in many organizational set-ups: leaders need to trust followers in order to not demotivate those who are motivated; but leaders also need to punish non-cooperation in order to motivate those who are not motivated, and thereby sustain cooperation by the motivated, as well. More generally, successful

leadership relies on *motivating the non-motived without demotivating the motivated*. This can become a quite complex task.

Wouldn't it thus be great, if groups consisted only of one type? At best, of course, of the motivated type! But even if everybody were a free rider, the complexity of leadership would be much reduced as, in principle, classic economic instruments could be applied. The third dimension of the *CC strategy* therefore considers the question whether it is possible that motivated and non-motivated agents separate – via self-selection – in different groups and organizations and what leaders can do, if anything, to promote and sustain such separation. Leaders clearly have an interest in attracting *CC*s and avoiding *FR*s (cf. Gächter and Thöni, 2005; Page et al., 2005; Cinyabuguma et al., 2005), so it would be good to understand if they can!

A priori, it seems unclear whether the sorting of *motivated* and *non-motivated* types via self-selection is possible and, more importantly, whether it is also sustainable given that the allocation of individuals, i.e., types, across organizations and firms in the modern world is the result of market interactions with free individual decisions. Some papers have argued that sorting is impossible, as labor markets will force firms that benefit from the presence of motivated types to pay higher wages, which attract non-motivated types (Lazear, 1989; Kandel and Lazear, 1992). As long as firms cannot identify types directly (e.g., by personality tests) firms will therefore be unable to benefit from a workforce of motivated types alone, at least in equilibrium. This suggests that there is little leaders can do to attract the motivated.

However, in Kosfeld and von Siemens (2009, 2011) and von Siemens and Kosfeld (2014) we show that the above argument is not entirely correct and that separation via self-selection is very well possible. While pooling of motivated and non-motivated types

cannot always be ruled out, results show that there always exists a separating equilibrium in which types self-select into different organizations that differ from each other both in terms of incentives and in terms of effort level and cooperation. The main mechanism behind this separation result is very intuitive: if motivated types care about being together with other motivated types (because they are able to achieve personal and organizational goals better), they can be attracted by organizations that are unattractive for non-motivated types. One possibility to achieve this is to pay (slightly) lower wages. As firms benefit from such a strategy as well, they will be willing to do so, also in competitive markets. Thus separation can be sustained.

On the one hand, this separation of motivated and non-motivated types offers a new explanation for the often surprising heterogeneity we see between firms with respect to, for example, the provision of incentives, the level of team work or, more generally, the organizational culture, even between firms that operate in the same industry (see, e.g., Gittell, 2000; Gittell et al., 2004; Ichniowski et al, 1997). On the other hand, it also opens possibilities for leadership to play an important role in this respect as well. Two elements are needed for leaders to be able to become *points of attraction* for motivated types: Firstly, motivated types need to have an interest in interacting with other motivated types. This can come from general complementarities between workers' effort and input in an organization's production function, or from explicit teamwork and worker cooperation, more specifically. Secondly, leaders need to provide incentives, or more generally shape the organizational environment and culture such that non-motivated types are unwilling to self-select into the organization. One possibility to achieve this is to impose constraints on paying out high wages or to emphasize other non-material dimensions of the work

environment (see below). Alternatively, leaders may also build up a reputation that non-cooperation will be punished and not be tolerated.

In Bauer et al. (2017), we test the underlying mechanisms behind these ideas in a lab experiment. The experiment runs over several rounds. In each round, participants receive a private resource of 10 experimental units, from which they can make an investment to generate a monetary donation to the *Deutsche Krebshilfe*, a charity that funds cancer research in Germany. Before participants decide about their investment, every participant has to choose between two teams, team A and team B. Participant i's effective donation d_i in a given round is then generated by multiplying i's investment x_i with the average investment of all other participants who are in the same team, i.e.,

$$d_i = x_i \text{ average } x_j, \tag{4}$$

where x_j is the investment of any participant j who has chosen the same team as participant i. Resources that are not invested by a participant have a marginal value of 5 (but see below). Thus, participant i's monetary payoff in any round is given by

$$5(10 - x_i). \tag{5}$$

Payoff function (5) implies that participants, who do not care about the *Deutsche Krebshilfe* (in other words, who are *non-motivated*), will maximize their payoff by investing $x_i = 0$. Participants play in total 20 rounds in the experiment with feedback in each round about the number of participants as well as the average investment in both teams in the previous round. At the end of the experiment, one round is randomly drawn and participants are paid and donations made according to the decisions in this round.

We consider three different treatments. In the first treatment T1, team A and team B are identical. In particular, each unit that is not invested to generate a donation is worth 5 to any participant, independent of whether he is team A or team B. In the second

treatment T2, we make team B materially more attractive by increasing the marginal value of each unit that it is not invested to 7. Everything else is kept the same, i.e., donations are again determined by multiplying individual investments with the average investment of other participants who are in the same team. In the third treatment T3, marginal values differ as in treatment T2 but we no longer allow participants to self-select into teams. Instead, participants are randomly assigned to teams in each round in this treatment.

What shall we expect in this experiment? Note that while the generation of donations may appear artificial, it captures an important element highlighted above: a strong complementarity between individual investments. If others in my team invest a lot, the donation I generate with any investment is higher compared to if others invest only little. Ceteris paribus, the more the others invest the higher is my donation. For example, if I invest 5 units and all others in my team invest 5 units as well, my donation is equal to 25. If instead others' average investment in the team equals 1, my donation is only equal to 5. And if average investment is zero, my donation is zero as well. Thus, motivated participants who care about generating donations to the *Deutsche Krebshilfe* have an interest in being in a team together with other motivated participants, and they want to avoid participants who are non-motivated. The question is whether they can achieve this.

Without going into theoretical details it should have become clear by now that treatment T2 is the one in which we may expect separation to be observed. The intuition is that only here there exists a team (team A) that is relatively unattractive for non-motivated agents and *therefore* potentially attractive for motivated agents. In treatment T1, marginal values in both teams are identical. In treatment T3, marginal values are different (and hence also opportunity costs which may have an effect on investments as well) but self-selection is ruled out due to random assignment of participants into teams.

Our results in the experiment confirm the above reasoning. In treatment T1 participants invest, on average, a bit more than two units in both teams. As teams are identical, behavior is indeed indistinguishable. Furthermore, over all rounds participants distribute roughly 50:50 across the two teams. In contrast, in treatment T2 average investments increase to about five units in team A, while they stay at the level of two units in team B. On average, about 20 percent of the participants choose team A and 80 percent choose team B. The positive effect on investments is due to self-selection, because in treatment T3 where selection is ruled out by design (but the difference in marginal values between team A and team B is kept constant), no such effect is observed. Instead, here average investments are again at a level of a bit more than two units in both teams.

These results suggest that the sorting of *motivated* and *non-motivated* types via self-selection into different organizations (here, teams) is possible. Thus, there is scope for leadership to play an important role here, as well. What precise instruments will prove best is something I expect future research to show. However, NGOs and non-profit organizations already provide a useful example. Because what these organizations have in common, besides being characterized by a particular "mission" (e.g., to fight cancer or poverty) that also attracts a particularly motivated workforce, is that these organizations often face, or implement, explicit constraints on re-distributing surplus within the organization. In line with their mission these organizations are credibly committed to spend a significant part of their surplus on a particular non-profit goal or some public good or service. They thus have less leeway to pay their workers high wages compared to a "normal" profit-maximizing firm. This commitment creates an important advantage in attracting motivated workers, not because the latter care particularly about the non-profit goal per se (they may well do, and probably the more the better) but because non-

motivated workers are kept away and *therefore* motivated have an incentive to come (and stay).

Discussion

This article has two main goals: Firstly, to show that there exists an important heterogeneity in individual motives to cooperate in social dilemma situations. While some (the *CC*) are willing to cooperate voluntarily conditional on the cooperation of others, others (the *FR*) are self-interested and free ride if they can. Secondly, to argue that this heterogeneity, i.e., the co-existence of these different types, has important implications for leadership that, to the best of my knowledge, have not been addressed in the literature so far. I call these implications the *CC strategy*. The first implication is that leaders need to trust in order to not demotivate the motivated. Since the voluntary cooperation of motivated types is conditional on the cooperation of others (including leaders), distrust can lead to a self-fulfilling prophecy in which beliefs are confirmed ex post (i.e., no cooperation is the outcome) although they are false ex ante (i.e., agents are cooperatively motivated). But leaders also need to punish, and this is the second implication of the heterogeneity of types. Because only if the non-motivated are motivated to cooperate, due to the punishment of non-cooperation, the motivated will cooperate, as well. Otherwise, no cooperation is, again, the outcome. Finally, leaders have an interest in attracting motivated types. Whether they can, depends on the degree to which motivated types care about being together with others who are also motivated and whether leaders manage to shape an organization's environment such that the non-motivated are indeed kept away. The available evidence suggests that this is possible, though, perhaps, not always straightforward.

What can organizational psychologists and economists learn from this research (and from another)?

The research described in this article can clearly be characterized as a "typical economists' approach": Abstract models based on tools from mathematical game theory are used to analyze the organizational set-up, and empirical methods – here, a combination of lab experiments and field data – are employed to compare the theoretical predictions to the observations made in the data. This raises the question, what organizational psychologists, who often have a different methodological perspective on leadership research, can learn from this work. At the same time, what is it that economists may learn from leadership research in psychology?

Firstly, I hope of course that the main message and arguments of the *CC strategy* itself are both of interest and convincing to organizational psychologists. Where questions or doubts remain I am happy to learn, hoping that future research will be able to provide answers.

Secondly, on a more general level I think a useful lesson for non-economists pursuing leadership research is that the results and studies in this article clearly show that economists are actually not too far away from other social scientists in the sense that leadership is considered to play an important role in shaping (and understanding) human behavior in firms, groups, and organizations. This itself may have a positive effect on the willingness and motivation of researchers of both sides to engage in fruitful interdisciplinary exchange on this topic. One important element of the research described here, for example, is a rather "follower-centric" perspective on leadership focusing on the implications of different follower types (motivated vs. non-motivated, CC vs. FR) on effective leadership. In contrast, most leadership theories in psychology appear to take a

more “leader-centric” approach, although there exist recent exceptions (e.g., Kohles et al., 2012; Tee et al., 2013). Classic situational leadership theory (SLT) (Hersey and Blanchard, 1972) seems particularly related to my arguments, as the key idea of SLT is also that different follower types need to be treated differently. While the empirical support for SLT seems to be relatively scarce (cf. Thompson and Vecchio, 2009), its continued popularity (in particular among business practitioners) suggests that a deeper investigation of its theoretical and empirical foundation – potentially linking also to the arguments made here – seems worthwhile.

Thirdly, I hope the studies in this article illustrate that behavioral economists have developed quite a powerful toolbox of both analytical and empirical methods that in combination offer great new opportunities to take a fresh look at important leadership questions. For example, the role of trust is a topic that has obviously been studied in lots of research in management and psychology in the past. One hypothesis that is sometimes put forward is that trust breeds trustworthiness, i.e., cooperation in a leader-follower framework is high, because trust *increases* the motivation of followers to behave cooperatively. Our results in Falk and Kosfeld (2006) show that this assumption is unwarranted. Without going into details, the experimental data together with a rigorous game-theoretic analysis of players’ incentives and equilibrium behavior clearly show that an underlying prior motivation of agents to cooperate voluntarily is a necessary condition to reach cooperation as an equilibrium outcome. Therefore, trust does not breed trustworthiness but rather, distrust in the form of non-cooperation by a leader *destroys* trustworthiness and thus brings about non-cooperation on the part of followers. This is, I believe, an important message that in my view is difficult to obtain without the use of game theoretic analysis in combination with clean experimental data.

In fact, one major reason why I think that behavioral economics has become so successful within the economic literature over the recent years is that it is exactly the close link between game theoretic analysis and empirical, in particular experimental, methods that has allowed researchers to test existing theories by clever experimental designs (both in the lab and in the field), which then provide the basis for better theories that again are tested by new data, and so forth.

From an outside perspective this suggests that also non-economic researchers, including organizational psychologists, may want to use these methods and concepts, contrast them with their own research and results, and see what new lessons can potentially be drawn. Here, I would find the heterogeneity of leader motives that are uncovered by behavioral experiments, as for example in Kosfeld and Rustagi (2015), a promising starting point. Obviously, there exists a multitude of interesting field set-ups as well as administrative data and surveys from relevant organizations and natural leadership contexts. It would be great to see more behavioral experiments implemented along these lines in leadership research in the future.

Another exciting avenue I think is to compare more systematically different established leadership concepts (e.g., transformational, relation-oriented, or task-oriented leadership) with the behavior of natural leaders in various experimental games. This may be achieved by combing established scales from organizational psychology with experimental games leaders play in the same study. Here, I would think there is also a lot to learn for economists. For example, economists typically interpret participants' behavior in a particularly designed experiment as a "revealed preference" (see the discussion in this article). Yet, we actually know only very little about the true stability of such "preferences" both across contexts and across time. It is thus unclear, whether behavior in

these games can indeed be interpreted as a stable personality characteristic, i.e., as an individual *trait*, or better as a *state*. More research is needed to answer this question and it seems obvious that economists can benefit a lot from the work of psychologists in this area. Another direction is to explore further the possibilities leaders have to motivate followers by acting in a particular way, e.g., via leading-by-example (Hermalin, 1998) or as transformational or charismatic leaders (Bass, 1985; Antonakis et al., 2016; Zehnder et al., 2017).

In sum, what the analysis and discussion in this article shows is that there is scope (or at least hope) for fruitful and exciting interdisciplinary exchange in future leadership research, in which economists and psychologists not only acknowledge and take into account the results and evidence from the other discipline, but where both disciplines, each with its own methodological idiosyncrasies (and with the sufficient distance between each other to judge and recognize the other discipline's pros and cons), also contribute methodologically to our *joint* understanding of how humans, including leaders, interact and make decisions.

References

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3,* 293–319.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*, 1086–1120.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137,* 594–615.

Bass, B. M. (1985). *Leadership and performance beyond expectations*. New York: The Free Press.

Bauer, K., Kosfeld, M., & von Siemens, F. A. (2017) Self-selection in the lab. Working Paper Goethe University Frankfurt.

Bolton, P., Brunnermeier, M. K., & Veldkamp, L. (2013a). Economists' perspectives on leadership. *Handbook of leadership theory and practice: An HBS centennial colloquium on advancing leadership*. Harvard Business Press.

Bolton, P., Brunnermeier, M. K., & Veldkamp, L. (2013b). Leadership, coordination, and corporate culture. *Review of Economic Studies, 80,* 512–537.

Bolton, P. & Dewatripont, M. (2005). *Contract theory*. Cambridge, MA: MIT Press.

Brandts, J. and Cooper, D. (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association, 5,* 1223–1268.

Brandts, J., Cooper, D., & Weber, R. A. (2015). Legitimacy, communication, and leadership in the turnaround game. *Management Science, 61,* 2627–2645.

Burdin, G., Halliday, S., & Landini, F. (2015). Third-party vs. second-party control: Disentangling the role of autonomy and reciprocity. *IZA Discussion Paper* No. 9251.

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics, 89,* 1421–1435.

Clark, K. & Sefton, M. (2001). The sequential prisoner's dilemma: Evidence on reciprocation. *Economic Journal, 111,* 51– 68.

Dana, J., Weber, R. A., & Xi Kuang, J. (2007). Exploiting moral wiggle room: experiments demonstrating an illusionary preference for fairness. *Economic Theory, 33*, 67–80.

De Cremer, D. (2002). Charismatic leadership and cooperation in social dilemmas: A matter of transforming motives? *Journal of Applied Social Psychology, 32,* 997–1016.

De Cremer, D. & Van Knippenberg, D. (2002). How do leaders promote cooperation? The effects of charisma and procedural fairness. *Journal of Applied Psychology, 87,* 858– 866.

van der Weele, J., Kulisa, J., Kosfeld, M., & Friebel, G. (2014). Resisting moral wiggle room: How robust is reciprocal behavior? *American Economic Journal: Microeconomics, 6*, 256–264.

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

Dinh, J. E., Lord, R. G., Gardner, W. L., Meuser, J. D., Liden, R. C., & Hu, J. (2014). Leadership theory and research in the new millennium: Current theoretical trends and changing perspectives. *Leadership Quarterly, 25,* 36-62.

Drouvelis, M. & Nosenzo, D. (2013). Group identity and leading-by-example. *Journal of Economic Psychology, 39,* 414– 425.

Ellingsen, T. & Johannesson, M. (2008). Price and prejudice: The human side of incentive theory. *American Economic Review, 98*, 990–1008.

Falk, A. & Kosfeld, M. (2006). The hidden costs of control, *American Economic Review, 96*, 1611–1630.

Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*, 980–994.

Fehr, E. & Schmidt, K. (2006). The economics of fairness, reciprocity and altruism – experimental evidence and new theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook on the economics of giving, reciprocity and altruism*, vol. 1, 615–691, Amsterdam: Elsevier.

Fischbacher, F., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from public goods experiments. *Economics Letters, 71*, 397–404.

Gächter, S., Nosenzo, D., Renner, E., & Sefton, M. (2012). Who makes a good leader? Cooperativeness, optimism, and leading-by-example. *Economic Inquiry*, 50, 953-967.

Gächter, S. & Renner, E. (2014). Leaders as role models for the voluntary provision of public goods. *IZA Discussion Paper* No. 8580.

Gächter, S. & Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association, 3,* 303–314.

Gerlach, P. (2008). Experimental studies on incentives, trust, and social preferences in organizations, PhD dissertation, University of Cologne.

Gibbons, R. and Roberts, J. (2013). Economic theories of incentives in organizations. In R. Gibbons, & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton University Press.

Gittell, J. H. (2000). Organizing work to support relational co-ordination. *International Journal of Human Resource Management, 11*, 517–539.

Gittell, J. H., von Nordenflycht, A., & Kochan, T. A. (2004). Mutual gains or zero sum? Labor relations and firm performance in the airline industry. *Industrial and Labor Relations Review, 57*, 163–180.

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

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics, 120*, 220–230.

Güth, W., Levati, M. V., Sutter, M., & van der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics, 91,* 1023–1042.

Hardin G. (1968). The tragedy of the commons. *Science, 162*, 1243.

Hersey, P. & Blanchard, K. (1972). *Management of organizational behavior*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Hermalin, B. E. (1998). Toward an economic theory of leadership: Leading by example. *American Economic Review, 88,* 1188–1206.

Hermalin, B. E. (2013). Leadership and corporate culture. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton University Press.

Herrmann, B., Thöni, C., & Gächter S. (2008). Antisocial punishment across societies. *Science, 319*, 1362–1367.

Herrmann, B. & Thöni, C. (2009). Measuring conditional cooperation: A replication study in Russia. *Experimental Economics, 12*, 87–92.

Homonoff, T. A. (2018). Can small incentives have large effects? The impact of taxes versus bonuses on disposable bag use. *American Economic Journal: Economic Policy, 10,* 177–210.

Ichniowski, C., Shaw, K., & Prennushi, G. (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *American Economic Review, 87*, 291–313.

Kagel, J. H. & Roth, A. E. (2016). *The handbook of experimental economics, vol. 2*. Princeton University Press.

Kandel, E. & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy, 100*, 801–817.

Kessler, J. B. & Leider, S. (2016). Procedural fairness and the cost of control. *Journal of Law, Economics, and Organization, 32*, 685–718.

Kocher, M., Cherry, T., Kroll, S., Netzer, R., & Sutter, M. (2008) Conditional cooperation on three continents. *Economics Letters, 101*, 175–178.

Kohles, J. C., Bligh, M. C., & Carsten, M. K. (2012). A follower-centric approach to the vision integration process. *Leadership Quarterly, 23,* 476–487.

Kosfeld, M., Okada A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review, 99*, 1335–1355.

Kosfeld, M. & Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia. *American Economic Review, 105*, 747–783.

Kosfeld, M. & von Siemens, F. A. (2009). Worker self-selection and the profits from cooperation. *Journal of the European Economic Association, 7*, 573–582.

Kosfeld, M. & von Siemens, F. A. (2011). Competition, cooperation, and corporate culture. *RAND Journal of Economics, 42*, 23–43.

Kuhn, H. W., Harsanyi, J. C., Selten, R., Weibull, J. W., van Damme, E., Nash, J. F., & Hammerstein, P., (1996) The work of John Nash in game theory. Nobel Seminar, December 8, 1994. *Journal of Economic Theory, 69*, 153–185.

Kurzban, R. & Houser, D., (2005) Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 1803–1807.

Laffont, J.-J. & Martimort, D. (2002). *The theory of incentives: The principal-agent model*. Princeton: Princeton University Press.

Lazear, E. P. (1989). Pay equality and industrial politics. *Journal of Political Economy, 97*, 561–580.

Liden, R. C. & Antonakis, J. (2009). Considering context in psychological leadership research. *Human Relations, 62,* 1587–1605.

Luhmann, N. (1968). *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. 4th Edition Stuttgart: Lucius & Lucius, 2000.

Manso, G. (2011). Motivating innovation. *Journal of Finance, 66,* 1823–1869.

Markussen, T., Putterman, L., & Tyran, J.-R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies, 81,* 301–324.

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. W. (2018). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. CESifo Working Paper No. 6358.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*, 91–112.

Nikiforakis, N., Noussair, C., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics, 96*, 797–807.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009) Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 276, 323– 329.

Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. *Economic Journal, 115,* 1032–1053.

Potters, J., Sefton, M., & Vesterlund, L. (2007). Leading-by-example and signaling in voluntary contribution games: An experimental study. *Economic Theory*, 33, 169–182.

Riener, G. & Wiederhold, S. (2016). Team building and hidden costs of control. *Journal of Economic Behavior & Organization, 123*, 1–18.

Rotemberg, J. J. & Saloner, G. (1993). Leadership style and incentives. *Management Science, 39,* 1299–1318.

Rustagi D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management, *Science, 330*, 961–965.

Schnedler, W. & Vadovic, R. (2011). Legitimacy of control. *Journal of Economics & Management Strategy, 20*, 985–1009.

von Siemens, F. A. & Kosfeld, M. (2014). Team production in competitive labor markets with adverse selection. *European Economic Review, 68*, 181–198.

Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review, 97*, 999–1012.

Tee, E. Y. J., Ashkanasy, N. M., Paulsen, N. (2013). The influence of follower mood on leader mood and task performance: An affective, follower-centric perspective of leadership. *Leadership Quarterly, 24,* 496–515.

Thompson, G. & Vecchio, R. P. (2009). Situational leadership theory: A test of three versions. *Leadership Quarterly, 20,* 837–848.

Tyran, J.-R. & Feld, L. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics*, 108, 135–156.

Van Vugt, M. & De Cremer, D. (1999). Leadership in social dilemmas: The effects of group identification on collective actions to provide public goods. *Journal of Personality and Social Psychology, 76,* 587–599.

Weber, R. A., Camerer, C. F., Rottenstreich, Y., & Knez, M. (2001). The illusion of leadership: Misattribution of cause in coordination games, *Organizational Science, 12,* 582–598.

Yukl, G. (2013). *Leadership in organizations*. Essex, UK: Pearson Education.

Zehnder, C., Herz, H., & Bonardi, J. P. (2017). A productive clash of cultures: Injecting economics into leadership research. *The Leadership Quarterly, 28*, 65–85.

Ziegelmeyer, A., Schmelz, K., & Ploner, M. (2012). Hidden costs of control: four repetitions and an extension. *Experimental Economics, 15*, 323–340.