ELSEVIER

European Economic Review 46 (2002) 687–724

EUROPEAN ECONOMIC REVIEW

www.elsevier.com/locate/econbase

Joseph Schumpeter Lecture

# Psychological foundations of incentives☆

Ernst Fehr*, Armin Falk

*Institute for Empirical Economic Research, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland*

**Abstract**

During the last two decades economists have made much progress in understanding incentives, contracts and organizations. Yet, they constrained their attention to a very narrow and empirically questionable view of human motivation. The purpose of this paper is to show that this narrow view of human motivation may severely limit understanding the determinants and effects of incentives. Economists may fail to understand the levels and the changes in behaviour if they neglect motives like the desire to reciprocate or the desire to avoid social disapproval. We show that monetary incentives may backfire and reduce the performance of agents or their compliance with rules. In addition, these motives may generate very powerful incentives themselves. © 2002 Elsevier Science B.V. All rights reserved.

*JEL classification:* J41; C91; D64

*Keywords:* Incentives; Contracts; Reciprocity; Social approval; Social norms; Intrinsic motivation

## 1. Introduction

Economics is based on incentives and it derives its strength from being able to predict how people change their behaviour in response to changes in incentives. Economic theory provides powerful theoretical tools for predicting the effects of changes in incentives – tools that are hardly matched by any other social science. At the same time, however, economists tend to constrain their attention to a very narrow and empirically questionable view of human motivation. Contract theory and principal – agent theory, for example, typically restrict their attention to the motives to achieve income through

☆ Schumpeter Lecture, Annual Conference of the European Economic Association 2001.

* Corresponding author. Tel.: +41-1-257-3709; fax: +41-1-3640366.
*E-mail addresses:* efehr@iew.unizh.ch (E. Fehr), falk@iew.unizh.ch (A. Falk).

0014-2921/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0014-2921(01)00208-2

effort and to avoid risks. *It is the purpose of this paper to show that this narrow view of human motivation may severely limit progress in understanding incentives.*

We will provide evidence suggesting that powerful non-pecuniary motives like the desire to reciprocate or the desire to avoid social disapproval, also shape human behaviour. By neglecting these motives economists may fail to understand the levels and the changes in behaviour. Moreover, we will show that these motives interact in important ways with economic incentives. As a consequence economists may even fail to understand the effect of *economic* incentives on behaviour if they neglect these motives. In particular, we will show that because of the existence of these motives, economic incentives may backfire and reduce the agents' performance or compliance with rules.

In this paper we will discuss the interactions of three important human motives with economic incentives – the motive to reciprocate, the desire for social approval and the desire to work on interesting tasks. The first two motives are social in nature, i.e., by taking them into account one acknowledges human beings as social beings. The third motive is not related to the social nature of man but originates in the nature of certain tasks. There are many tasks providing intrinsic enjoyment for those who perform them and these tasks are therefore undertaken even in the absence of economic incentives. Section 2 provides experimental evidence indicating that reciprocity may severely weaken certain economic incentives while at the same time strengthening other kinds of economic incentives. In addition it is shown that reciprocity by itself constitutes a source of powerful economic incentives. In Section 3 we discuss the complications that arise for incentive provision when social approval is important. The presence of approval motives implies, among other things, that economic incentives may backfire and lead to *permanent* negative effects on rule compliance. Thus, even if the incentive change that caused the negative effect on rule compliance is removed, the extent of rule compliance may have been permanently reduced as a result of the initial change in the incentive. In Section 4 we discuss the psychological literature on the interaction between extrinsic incentives and task-specific intrinsic motivation. We argue that, although the results and the claims of this literature are intriguing and interesting, the *economic* relevance of this literature has yet to be shown. This means that further research will be necessary to remove the prevailing ambiguities regarding the interpretation of results. In addition, it is necessary to test the claims of this literature in economically relevant contexts.

By pointing out the limits of the prevailing economic view of incentives we aim at providing a better psychological foundation of incentives. Thus, despite our criticism our endeavour is constructive rather than destructive. In fact, we share a great admiration for the accomplishments of contract and incentive theory over the past two decades. The theory generated important insights and provides the theoretical tools that are the basis for the rigorous modelling of a larger set of human motives. It is our hope that economists will meet the challenge that is generated by our data. Since there are still important gaps in our empirical and theoretical knowledge much remains to be done.

## 2. Reciprocity and economic incentives

This section discusses the interactions between a particularly important kind of social preference – reciprocity – and economic incentives. During the last 15 years experimental economists have documented the existence of a class of non-pecuniary motives that have been called "social preferences". A person exhibits social preferences if the person does not only care about the material resources allocated to her but also cares about the material resources allocated to relevant reference agents. Depending on the situation, the relevant reference agents may be the colleagues in the firm with whom a person interacts most frequently, or a person's relatives, or a trading partner, or a person's neighbours. In principal – agent situations it is quite likely that the principal constitutes a reference actor for the agent. If there are multiple agents it also seems likely that agents also care about the material resources allocated to the other agents. The experimental evidence indicates that a substantial fraction of the people exhibits social preferences. In this paper we do not attempt to summarize the empirical evidence on social preferences (for surveys see Fehr and Schmidt, 2001; Sobel, 2001). Instead, we single out one kind of social preference that is particularly important for our purposes – the preference for reciprocity. [1]

Reciprocity can be viewed as a contingent social preference because depending on the behaviour of the reference person, e.g., the principal, a reciprocal agent values the principal's material payoff positively or negatively. More specifically, if the agent perceives the actions of the principal as kind, the agent values the principal's payoff positively. If, in contrast, the principal's actions are perceived as hostile, the agent values the principal's payoff negatively. Whether an action is perceived as kind or hostile depends on the consequences and the fairness or unfairness of the intention underlying the action. The fairness of the intention, in turn, is determined by the equitability of the payoff distribution, relative to the set of feasible payoff distributions, caused by the action.

It is important to emphasize that reciprocity is not driven by the expectation of future material benefits. It is, therefore, fundamentally different from "cooperative" or "retaliatory" behaviour in repeated interactions. These behaviours arise because actors expect future material benefits from their actions; in the case of reciprocity, the actor is responding to friendly or hostile actions even if no material gains can be expected. Rabin (1993), Levine (1998), Falk and Fischbacher (1999), Dufwenberg and Kirchsteiger (1999), Segal and Sobel (1999) as well as Charness and Rabin (2000) have developed models of reciprocity. Other authors like, for example, Fehr and Schmidt (1999), have tried to capture important elements of reciprocity in simpler, and hence more tractable, models of inequity aversion.

[1] This does not mean that we believe that other types of social preferences like, e.g., altruism or spitefulness, are unimportant. It reflects, however, our belief that reciprocity is frequently quantitatively more important than other types of social preferences and that it has particularly important consequences in strategic interactions. For more detailed arguments on this see Fehr and Fischbacher (forthcoming).

### 2.1. Reciprocity as a source of voluntary cooperation

In this section we provide evidence indicating that reciprocity induces agents to cooperate voluntarily with the principal if the principal treats them kindly. The evidence is based on a so-called gift exchange experiment conducted by Fehr et al. (1997).[2] In the experiment a subject in the role of an employer (the principal) can make a job offer to the group of subjects in the role of workers (the agents). Each worker can potentially accept the offer. There are more workers than employers to induce competition among the workers. A job offer consists of a *binding* wage offer $w$ and a *non-binding* 'desired effort level' $\hat{e}$. If one of the workers accepts an offer $(w, \hat{e})$ she has to determine the *actual* effort level $e$. In the experiment the choice of an effort level is represented by the choice of a number. The higher the chosen number the higher is the effort and the higher are the monetary effort costs to be borne by the worker. The desired and the actual effort levels have to be in the set $\{e_{\min}, \ldots, e_{\max}\} \equiv \{0.1, 0.2, \ldots, 1\}$ and the wage offer has to be in the set $\{0, 1, \ldots, 100\}$. The higher $e$ the larger is the material payoff for the employer but the higher are also the worker's effort costs $c(e)$. Material payoffs from an exchange are given by $100e - w$ for the employer and $w - c(e)$ for the worker. A party who does not manage to trade earns zero. The effort costs are increasing and convex with $c(e_{\min}) = 0$ and $c(e_{\max}) = 18$.

Note that since $\hat{e}$ is non-binding the worker can choose any $e$ in the set $\{0.1, 0.2, \ldots, 1\}$ (in particular $e < \hat{e}$) without being sanctioned. It is obvious that, since $c(e)$ is strictly increasing in $e$, a selfish worker will always choose $e = e_{\min} = 0.1$. Therefore, a rational and selfish employer, who believes that there are only selfish workers, will never offer a wage above $w = 1$. This is so because the employer knows that the workers will incur no effort costs and, being selfish, will accept a wage offer of $w = 1$. At $w = 1$ the trading worker earns 1 which is more than if the worker does not trade. However, if the employer believes that there are sufficiently many reciprocal workers he has an incentive to offer more generous wages because this induces the reciprocal workers to provide higher effort levels. In addition, the employer may appeal to the workers' reciprocity by being more generous when choosing a higher desired effort level.

Fig. 1 depicts the results of this experiment. The figure shows that higher desired effort levels are indeed associated with more generous offers to the workers. The higher $\hat{e}$ the higher was the rent $w - c(\hat{e})$ offered to the workers. This suggests that employers

[2] In this experiment subjects were not informed about the identity of their trading partner and the parties could not establish repeated interactions. The experimental procedures also ensured that no subject could acquire a reputation for being, for example, cooperative. Trading partners were located in different rooms. These features of the experiment ensured that the exchange really took place between anonymous strangers. In all laboratory experiments discussed in this paper subjects could earn significant amounts of money according to their decisions and the rules of the experiment. Completely anonymous strangers, who never learned the identities of their interaction partners, interacted with each other. The reason for this is not that we believe that anonymous interactions are particularly realistic. Yet, if reciprocity shows up in anonymous interactions it is even more likely to show up in non-anonymous interactions. In addition, non-anonymous interactions are likely to involve a host of confounding factors.

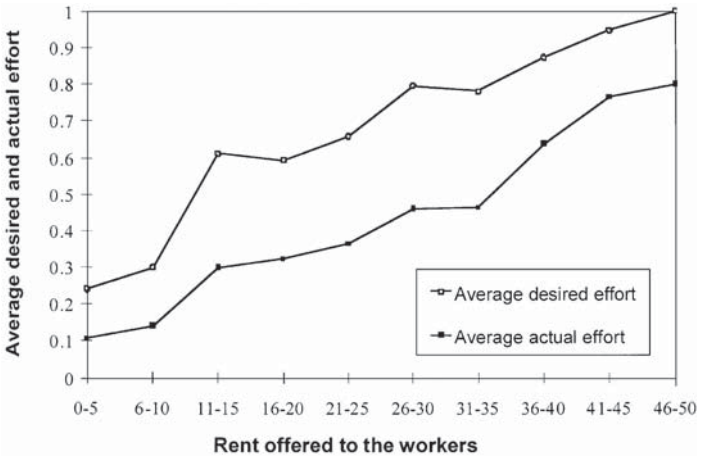


Fig. 1. Relation of desired and actual effort to the rent offered to the workers (*source*: Fehr et al., 1997).

indeed wanted to elicit reciprocal responses from the workers.[3] Moreover, Fig. 1 shows that *on the average* the workers responded reciprocally to the employers' offers. The higher the rent that was offered to the workers the higher was the actual effort level. This means that workers exhibited voluntary cooperation depending on the generosity of the job offer. The existence of reciprocity-based voluntary cooperation should, however, not make us overlook two facts. First, there is still a lot of shirking as indicated by the difference between the desired effort and the actual effort. Second, in addition to the reciprocal workers there is also a substantial fraction of selfish workers who always choose the minimal effort or who rarely respond in a reciprocal manner.[4]

In our view these results are important because voluntary cooperation is relevant in many real world contexts. For example, whenever employees have discretion over the intensity or the type of activity they perform voluntary cooperation is very valuable for the firm. The relevance of voluntary cooperation for the employment relation is neatly

[3] An alternative interpretation is that the experimental employers just wanted to share the surplus that is produced if the worker performs at $\hat{e}$. This interpretation can be ruled out, however, because if effort is fixed exogenously, it turns out that employers pay much less generous wages.

[4] There are also many other studies suggesting the existence of reciprocity-driven voluntary cooperation (see, e.g., Fehr et al., 1993; Berg et al., 1995; Bolle and Kritikos, 1998; Brandts and Charness, 1999; Fehr and Falk, 1999; McCabe et al., 1998, 2000; Charness, 2000; Abbink et al., 2000; Gächter and Falk, 2001). Taken together, the fraction of subjects showing positive reciprocity is rarely below 40 and sometimes even 60 per cent, whereas the fraction of selfish subjects lies also often between 40 and 60 per cent. Moreover, these frequencies of positive reciprocity are observed in such diverse countries as Austria, Germany, Hungary, the Netherlands, Switzerland, Russia and the U.S. It is also worthwhile to stress that *positive reciprocity is not diminished if the monetary stake size is rather high*. In the experiments conducted by Fehr and Tougareva (1996) in Moscow subjects earned on average the monetary income of ten weeks in an experiment that lasted for 2 hours. The monthly median income of subjects was US $17 while in the experiment they earned on average US $45. The impact of reciprocity also does not vanish if the experimental design ensures that the experimenter cannot observe individual decisions but only aggregate decisions (Berg et al., 1995; Abbink et al., 2000).

confirmed by the extensive study of Bewley (1995, 1999). Bewley reports that "managers claim that workers have so many opportunities to take advantage of employers that it is not wise to depend on coercion and financial incentives alone as motivators" (Bewley, 1995, p. 252). In addition, Bewley's results suggest that reciprocity-based voluntary cooperation is the key reason for downward wage rigidity: "In economics, it is normally assumed that people, being self-interested, must be either coerced or bribed into performing tasks. However, the main causes of downward wage rigidity have to do with employers' belief that other motivators are useful as well, which are best thought of as having to do with generosity". Bewley's results nicely confirm the results of the competitive market experiments by Fehr et al. (1993) and Fehr and Falk (1999). These experiments explicitly show that reciprocity-driven voluntary cooperation causes downward wage rigidity because lower wages are associated with lower effort and lower profits.[5] If the experimenter rules out voluntary cooperation by fixing the effort level exogenously, wages converge to the competitive level, while if workers have the opportunity to cooperate voluntarily with their employer, wages remain far above the competitive level.

Reciprocity-driven voluntary cooperation also plays an important role in the context of the provision of public goods. It is shown by Croson (2000), Fischbacher et al. (2001), and Falk and Fischbacher (forthcoming) that many people increase their contribution to a public good if others also increase their contributions, although, in material terms, each individual has a strict incentive to contribute nothing. This kind of *conditional cooperation* thus introduces strategic complementarity into public goods situations. This is important for the management of the employment relation since public goods situations frequently arise within firms. The existence of conditional cooperation renders the management of the workers' beliefs about the other workers' effort important because if a conditional cooperator believes that the others shirk he will also tend to shirk.

One aspect of belief-management is choosing the right members for the organization. A few shirkers in a group of employees may quickly spoil the whole group. Bewley (1999), for example, reports that personnel managers use the possibility of firing workers mainly as a means to remove "bad characters and incompetents" from the group and not as a threat to discipline the workers. The reason is that explicit threats create a hostile atmosphere and may even reduce the workers' general willingness to cooperate with the firm. Managers report that the employees themselves do not want to work together with lazy colleagues because these colleagues do not bear their share of the burden, which is viewed as unfair. Therefore, the firing of lazy workers is mainly used to establish internal equity, and to prevent the unravelling of cooperation. This supports the view that conditional cooperation is important inside firms.

[5] In a recent paper Krueger (2001) provides strong evidence that the quality of Firestone tyres decreased significantly after the management of Firestone announced in January 1994 that it wants to reduce the wages of new hires by 30 per cent. Thus the deterioration of the quality of the tyres occurred although the wage cut was not yet implemented. As a consequence of the low quality of the tyres produced during the industrial conflict between the management and the workers Firestone had to recall 14.4 million tyres. According to the National Highway Traffic and Safety Administration Firestone tyres have been linked to 203 fatalities and more than 900 injuries.

There is a close relation between the notion of reciprocity and the idea that employers often deliberately attempt to change the preferences of their employees in ways that help to achieve the firm's goals. Employers prefer, in particular, loyal employees who take into account the goals of the firm. The very fact that employees have so many opportunities to take advantage of their employer renders loyal workers very valuable for the employer. It is interesting that in their widely known textbook *Economics, Organizations and Management* Milgrom and Roberts (1992) acknowledge this point when they write that "important features of many organizations can best be understood in terms of deliberate attempts to change preferences of individual participants". Yet, despite this their whole book is then based on the assumption that people behave as if they "were entirely motivated by narrow, selfish concerns".[6]

Loyalty means that the workers take into account the interests of their employer, which is just another way of saying that they value the employer's payoff positively. Hence, the notion of loyalty is closely related to the notion of social preferences and, in particular, to the notion of reciprocity because the existence of reciprocal workers means that employers can generate loyalty by being generous to the workers. If one acknowledges that many employees have reciprocal preferences the firms' attempts to change their employees' preferences are thus no longer mysterious. If it is true that some people are more self-interested than others then choosing the "right" people is one way of affecting the preferences of a firm's workforce. For this reason employers have a strong interest in recruiting employees who have favourable preferences and whose preferences can be affected in favourable ways. There is circumstantial evidence for this because the testing and screening of employees is often as much about the employee's willingness to become a loyal firm member as it is about the employee's technical abilities.

### 2.2. Explicit incentives and voluntary cooperation

After we have established the existence of reciprocity-driven voluntary cooperation the next question is how explicit incentives interact with voluntary cooperation. Do explicit incentives leave the willingness to cooperate voluntarily intact, do they increase it or do they decrease it? Moreover, if there are interaction effects, which features of the explicit incentive are driving the interaction? Fehr and Gächter (2000b) studied these questions in the context of the above gift exchange experiment by implementing the following incentive. In addition to $w$ and $\hat{e}$ the experimental employers could also stipulate a fine $f$ that had to be paid by shirking workers in case that shirking could be verified. The fine was constrained by an upper bound $f_{\max}$ and the probability of verifying shirking was equal to $s = 1/3$. Because of the upper bound on the fine the maximal enforceable effort level in the presence of self-interested risk neutral agents was $e = 0.4 > e_{\min} = 0.1$.[7] Thus, in the presence of only self-interested agents the

[6] For a recent attempt to incorporate social preferences in the theory of organisation see Rob and Zemsky (2000).

[7] For this simple incentive the no-shirking condition is given by $sf \geqslant c(\hat{e}) - c(e_{\min})$ where $sf$ is the expected loss from shirking while $c(\hat{e}) - c(e_{\min}) = c(\hat{e})$ is the expected gain from shirking because $c(e_{\min}) = 0$. The maximal enforceable effort can be derived from the equation $sf = c(\hat{e})$.

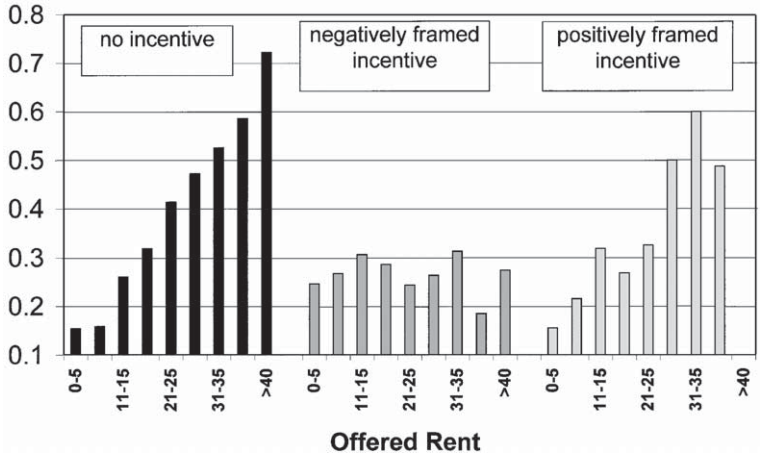


Fig. 2. The impact of explicit incentives on actual average effort (*source*: Fehr and Gächter, 2000b).

employer is always better off by imposing the maximal fine. Moreover, since the surplus is monotonically increasing in the effort level, the surplus is also maximized by imposing the maximal fine.

In the experimental instructions the term "fine" was not used because it was thought that "fine" is a value-laden term. Instead, the fine was described to the subjects as a wage deduction. Since Fehr and Gächter (2000b) were also interested in the impact of the framing of incentives they conducted an additional treatment in which the incentive was described as a bonus payment, i.e., as a wage increase relative to the base wage. In this treatment the employers could stipulate a base wage $w$, a desired effort $\hat{e}$ and a bonus $b$. As in the negatively framed treatment the bonus was constrained by an upper bound equal to $f_{\max}$. The bonus was not paid to a shirking worker in case that shirking could be verified, which happened again with probability $s = 1/3$. Thus, in economic terms the positively framed incentive is exactly identical to a corresponding negatively framed incentive. For example, if in the positive frame $b = f_{\max}$ the expected loss from shirking is $sf_{\max}$, which is exactly identical to the expected loss from shirking in the negative frame in case that $f = f_{\max}$. Thus, from an economic viewpoint, the set of enforceable effort levels does not differ across frames.

Fig. 2 presents the effort results of these experiments. The left graph in Fig. 2 shows the relation between the offered rent and workers' effort levels in the baseline treatment, i.e., when there is no explicit incentive at all. This graph replicates the results displayed in Fig. 1. The graph in the middle indicates how workers' effort levels respond to the offered rent when there is a negatively framed incentive. In 98.5 per cent of all the cases the employers stipulated a fine in this treatment and only in 1.5 per cent of the cases they set $f = 0$. In 69 per cent of the cases the maximal fine was imposed. This graph shows that voluntary cooperation is substantially and significantly weakened by the availability or the actual use of the incentive. The average effort in this treatment is even below $e = 0.4$, the level that can be forced on self-interested agents

by imposing the maximal fine. The reduction in effort is associated with a reduction in the surplus relative to the baseline treatment while – despite the lower surplus – the employers' profits are higher in the treatment with the negatively framed incentive. This is due to the fact that the use of the incentive allowed the employers to substantially change the distribution of the surplus. Instead of relying on costly generosity as an incentive device (i.e., the carrot) employers paid on the average much lower rents and relied on the fine (i.e., the stick) as an incentive device. Overall, the comparison between the left graph and the graph in the middle illustrates the main theme of this paper – that in the presence of non-pecuniary motives there are important and, relative to the predictions of the economic model, unexpected interactions between material incentives and non-pecuniary motives. It is also worth emphasizing that similar results were obtained in the studies of Bohnet et al. (2001), Benz et al. (2001), Evans et al. (2001) and Schulze and Frank (2001).

The notion of reciprocity provides a natural interpretation of the evidence in Fig. 2. Remember that reciprocity means that agents respond in a hostile manner to actions that reveal a hostile intention. In our view the fining of workers may reveal hostile intentions for two reasons. First, the fine per se may be perceived as hostile. Second, threatening to fine a worker is an indication of distrust. To the extent to which trusting actions are perceived as kind and distrusting actions as hostile, a fine will be perceived as a hostile act. Whatever the exact reason for the perception of a hostile intention is, if the workers perceive the fine as a hostile act they are no longer willing to put forward extra effort beyond the level that is dictated by self-interest. In fact, they may even be willing to shirk in response to a hostile contract although the expected cost of shirking exceeds the benefits of shirking. It is interesting that even if the employers pay a rather high rent the workers are no longer willing to provide much extra effort. It seems that the implicit message of a generous contract stipulating a fine is contradictory. Appealing to the workers' generosity and trustworthiness by being generous and, at the same time, expressing distrust by telling them that they will be fined if they do not respond with high effort levels does not seem to go together.

Our interpretation of the evidence in terms of reciprocity raises at least two questions. First, is it possible to affect the perceived kindness or hostility of an incentive by merely changing the framing of the incentive? This question can be answered by the treatment with the positively framed incentive because one might conjecture that the bonus-frame is likely to be perceived as less hostile than the fine-frame. The right graph in Fig. 2 indeed shows that voluntary cooperation is substantially higher when the incentive is framed in terms of a bonus payment. This indicates that the framing of an explicit incentive in terms of extra rewards elicits more effort compared to a frame in terms of punishment. This result suggests that reciprocity motives interact in important ways with cognitive factors. The notion of a kind or a hostile action inevitably depends on a reference point and our evidence suggests that these reference points can be manipulated by the framing of the incentive. In the negative frame the total compensation in case of non-shirking is the natural reference point and the fine focuses attention on the fact that something will be taken away in case of shirking. In the positive frame the base wage is the natural reference point and the bonus focuses attention on the fact that something will be given if the desired effort is provided. It seems that "taking

away something" is perceived as less friendly than "giving something" even if the total compensation is identical. So far there is no model of reciprocity that captures such shifts in the reference point.

Fig. 2 illustrates that positively framed incentives elicit much higher voluntary cooperation than negatively framed ones. However, the figure also indicates that in the absence of any explicit incentive voluntary cooperation is even higher than in the presence of a positively framed incentive. This effect is statistically significant (Fehr and Gächter, 2000b). A similar effect has been observed in a field experiment conducted by Berry and Kanouse (1987). They found that, by first paying physicians a certain sum of money, they could increase the likelihood that the doctors would complete and return a long questionnaire they received in the mail. When they added a check for $20 to the questionnaire 78 per cent of the doctors sent back a completed questionnaire. 95 per cent of those who returned the questionnaire cashed their checks while only 26 per cent of those who did not return the questionnaire did so. When, instead, the receipt of the check was contingent on returning a completed questionnaire only 66 per cent of the doctors returned the questionnaire. The result of this study has also been confirmed by the meta-analysis of Church (1993). Church reports that if the request for the completion and return of a survey is associated with an unconditional advance payment the response rate increases by 19 percentage points relative to surveys without concomitant payment. Moreover, when the payment of money is made contingent upon completion of the survey the response rate does not rise relative to the case where no payment is offered.[8] This suggests that the effects displayed in Fig. 2 also hold in other settings.

The second question that is raised by our interpretation concerns the difference between the availability of a hostile incentive and the actual use of a hostile incentive. If a hostile incentive is available and the employers can deliberately refrain from using this incentive, isn't this a particularly kind action? Again there may be two reasons for this: First, refraining from the explicit threat of punishment may be perceived as kind per se. Second, it also makes trust explicit in a salient way. If our interpretation is correct, then by explicitly *not* using a hostile incentive the employers should be able to elicit even higher effort levels compared to a situation in which no explicit incentive is available. Fehr and Rockenbach (2001) examined this conjecture in the context of a modified trust game (Berg et al., 1995). In this experiment an investor and a responder interact only once and both are endowed with 10 experimental money units (MUs).[9] The investor can send any $x \in \{0, 1, \ldots, 10\}$, to the responder and the experimenter then triples the amount that the responder receives. The responder observes the investor's transfer and can then send back any $y \in \{0, 1, \ldots, 3x\}$. The payoff of the investor is given by $10 - x + y$ and the payoff of the responder is defined as $10 + 3x - y$. In addition to transferring money to the responder the investor also announces a desired

[8] James and Bolstein (1992) report the following extreme case: They found that an unconditional advance payment of $5 elicited a response rate of 52 per cent while the offer to pay $50 contingent upon completion of the survey induced only 23 per cent of the potential respondents to return the survey. When no payment at all was offered the response rate was 21 per cent.

[9] One MU was equal to 0.5 German Marks.

The external validity of experimental results stemming from student populations is sometimes questioned because it could be the case that non-student populations behave in different ways. To address this criticism Fehr and List (2002) have replicated the Fehr–Rockenbach study with chief executive officers from Costa Rica. In addition they conducted a control treatment with students from Costa Rica. The study shows that CEOs are, in general, much more trusting and much more trustworthy than the students because the CEOs transfer more money and, controlling for the transfer $x$, they send back more money.[10] However, the differences across the treatments with and without incentives were qualitatively similar and quantitatively even larger than in the study by Fehr and Rockenbach. Controlling for the transfer levels, the back-transfers are much higher when the incentive is available but not used compared to the baseline treatment. This suggests that the behavioural patterns induced by reciprocal preferences are even stronger among the CEOs compared to student populations.

The same forces that explain the data pattern in Fig. 3 may also explain why so few marriages are accompanied by prenuptial agreements. We believe that prenuptial agreements are likely to introduce distrust into a marriage because they require detailed discussions and specifications of what will happen in case that the relationship will be terminated. As a consequence they may do more harm than good. Since it is impossible to specify all aspects of a marriage in a comprehensive contract, a marriage is always based on implicit agreements and voluntary cooperation. A marriage thus has to be based on mutual trust because otherwise it will not function well. Moreover, it also seems likely that being trusted is in itself valuable for the trustee. Including contingencies about what will happen if one party fails to abide by the contract is likely to be taken as an indication of distrust and perhaps even hostility, which in turn may trigger what the prenuptial agreement attempted to avoid – a lack of mutual trust and cooperation.[11]

### 2.3. *Reciprocity as a source of economic incentives*

In Section 2.1 we mentioned that, although a substantial fraction of experimental subjects exhibits reciprocal behaviour, there is also a large fraction of subjects who behave in a purely selfish manner. The negative side effects of the explicit incentives mentioned above do not apply to selfish subjects because these subjects do not exhibit voluntary cooperation. The interaction between reciprocity and the behaviour of selfish subjects therefore takes a different form. It is based on the economic incentives arising from the existence of reciprocal subjects. To illustrate the creation of economic

[10] Hannan et al. (forthcoming) found that in a gift exchange game MBA-students, who have a regular job, exhibit more trustworthiness compared to students without a regular job. This result and the results of Fehr and List suggest that subjects with more work experience behave in a more trustworthy manner.

[11] Recently, Becker (1998) argued that divorce laws should be replaced by compulsory marriage contracts because the contracts can be tailored to the needs of the marriage partners. However, in our view this would lead to the emergence of a standard marriage contract and discussions about deviating from the standard contract would lead to distrust and lack of cooperation as prenuptial agreements would do today. We owe this idea to David Kreps.

incentives through reciprocating subjects we reconsider the gift exchange experiments conducted by Fehr et al. (1997).

In an extension of the simple experiment discussed in Section 2.1 the authors examined the impact of giving the employers the option of responding reciprocally to the worker's choice of $e$. Each employer was given the opportunity to reward or punish the worker after he observed the actual effort. By spending one MU on reward the employer could *increase* the worker's payoff by 2.5 MUs, and by spending one MU on punishment the employer could decrease the worker's payoff by 2.5 MUs. Employers could spend up to 10 MUs on punishment or on rewarding their worker. The important feature of this design is that if there are only selfish employers they will never reward or punish a worker because both rewarding and punishing is costly for the employer. Therefore, in case that there are only selfish employers there is no reason why the opportunity for rewarding/punishing workers should affect workers' effort choice relative to the situation where no such opportunity exists. However, if a worker expects her employer to be a reciprocator it is likely that she will provide higher effort levels in the presence of a reward/punishment opportunity. This is so because reciprocal employers are likely to reward the provision of $e \geqslant \hat{e}$ and to punish underprovision ($e < \hat{e}$). This is in fact exactly what one observes, on the average. If there is underprovision of effort employers punish in 68 per cent of the cases and the average investment in punishment is 7 MUs. If there is overprovision employers reward in 70 per cent of these cases and the average investment in rewarding is also 7 MUs. If workers exactly meet the desired effort employers still reward in 41 per cent of the cases and the average investment into rewarding is 4.5 MUs.

We also elicited workers' expectations about the reward and punishment choices of their employers. Hence, we are able to check whether workers anticipate employers' reciprocity. It turns out that in case of underprovision workers expect to be punished in 54 per cent of the cases and the expected average investment into punishment is 4 MUs. In case of overprovision they expect to receive a reward in 98 per cent of the cases with an expected average investment of 6.5 MUs. As a result of these expectations workers choose much higher effort levels when employers have a reward/punishment opportunity. The presence of this opportunity decreases shirking from 83 to 26 per cent of the trades, increases exact provision of $\hat{e}$ from 14 to 36 per cent and increases overprovision from 3 to 38 per cent of the trades. The average effort level is increased from $e = 0.37$ to 0.65 so that the gap between desired and actual effort levels almost vanishes. An important consequence of this increase in average effort is that the aggregate monetary payoff increases by 40 per cent – even if one takes the payoff reductions that result from actual punishments into account. Thus, the reward/punishment opportunity considerably increases the total pie that becomes available for the trading parties.

~~We believe that the material incentives that are provided by reciprocal principals help solving one of the key problems in many agency relations, i.e., the problem of the provision of incentives when there are multiple tasks for the agents. In most employment relations the employees typically have to perform several tasks and because of measurement and verifiability problems it is often not possible to target explicit incentives to all tasks. It is well known from practice (Kerr, 1975) and from theory (Holmström and Milgrom, 1991; Baker, 1992) that in this situation explicit performance~~

contrast to this, the mere opportunity of punishing the agent after observing that the agent indeed shirked does not convey such a message. In this case the punishment threat is vague and implicit and nobody is "told" that she is considered as a potential cheater. Moreover, most subjects are likely to consider shirking as unfair if the contract offered the agent a generous share of the surplus. This means that most subjects are likely to consider the punishment of shirking agents, if the contract offer has been fair, as legitimate. The problem, therefore, is how to implement the punishment threat such that sanctioning is considered as legitimate without offending those agents who do not need to be coerced to cooperate.

We believe that reciprocity-based incentives based on the opportunity of punishing the agent ex post exactly achieve this. These incentives discipline the potential shirkers because they know that a certain fraction of the principals is going to punish them in case of shirking without offending those who cooperate voluntarily because there are no explicit threats. For the same reason we believe that the incentives arising from repeated interactions are so effective. The psychological properties of repeated game incentives are quite similar to the properties of reciprocity-based implicit incentives because they are imposed ex post without being explicitly announced ex ante. For example, in the experiments of Brown et al. (2001) the employers could not explicitly threaten to fire shirking workers but in fact they fired them. Our interpretation is that this disciplined the potential shirkers without offending the cooperators.

In our view the powerful effects of implicit incentives in endogenously repeated games also arise from the positive interactions between reciprocity and repeated game incentives. First, there is evidence (Van Dijk et al., forthcoming) that successful cooperation in repeated interactions strengthens the emotional and affective ties between the parties, which is just another way of saying that the parties' willingness to take the other party's interest into account is strengthened. This means that cooperation is self-reinforcing because successful cooperation has the effect that the parties care more for the other's payoff, which, in turn, enhances the willingness to cooperate voluntarily. Second, the presence of reciprocal subjects provides incentives for the selfish subjects to mimic the cooperative behaviour of the reciprocal subjects. This has been shown theoretically (Kreps et al., 1982) and experimentally (Gächter and Falk, 2001). For instance, if it were common knowledge that every actor is selfish, cooperation could not be sustained in the finitely repeated experiments of Brown et al. Yet, in the presence of reciprocal subjects, the selfish subjects can gain a credible reputation for being cooperative by behaving like the reciprocal subjects. In this way they can ensure themselves employment and a higher material payoff.

## 3. Social approval, social norms and economic incentives

Reciprocity is one powerful motive that interacts in important ways with material incentives but there are also other motives for which this is the case. In this section we discuss the interactions between the motive to gain social approval and to avoid social disapproval on the one hand and material incentives on the other hand. Since social (dis)approval is closely related to the enforcement of social norms the interaction

between (dis)approval and incentives is also relevant for the interplay of social norms and incentives.

### 3.1. The relevance of social approval

Circumstantial evidence and introspection suggests that many people like to receive social approval and try to avoid social disapproval. Social approval means that we are the objects of others' admiration while disapproval means that we are the objects of others' disgust and contempt. Approval, therefore, makes us proud and happy while disapproval causes embarrassment and shame and makes us unhappy. These social rewards and punishments are a basic "currency" that induces children and adults alike to perform certain activities and avoid others. What child does not want to receive approval from parents and teachers, what student does not want to be praised for performing well by his professors, and what scientist does not value the approval by her peers. The important role of social approval was already recognized by Smith (1759) in the Theory of Moral Sentiments where he wrote: "We are pleased to think that we have rendered ourselves the natural objects of approbation, ... and we are mortified to reflect that we have justly merited the blame of those we live with". Likewise, Harsanyi (1969) was convinced that social approval is important: "People's behaviour can largely be explained in terms of two dominant interests: economic gain and social acceptance". More recently there is a growing literature, which incorporates concerns for social approval into economic models, or which argues that such steps should be taken (e.g., Akerlof, 1980; Besley and Coate, 1992; Bernheim, 1994; Dufwenberg and Lundholm, 2001; Lindbeck, 1995, 1997; Lindbeck et al., 1997). However, mainstream economics has so far been relatively unmoved by these attempts.

While social approval may be valued positively because it sometimes generates material benefits, we believe that most of us also value social approval positively (and disapproval negatively) for its own sake. There is much circumstantial evidence and questionnaire evidence supporting the view that (dis)approval has behavioural consequences (e.g., Rainwater, 1979; Lindbeck, 1995, 1997). Moffit (1983) provides econometric evidence consistent with this view. In the U.S. as much as 30–60 per cent of the citizens who are eligible for welfare do not apply. The study of Moffit suggests that this is the result of the stigmatization of welfare recipients because living on welfare violates work norms.

Recently, Gächter and Fehr (1999) and Rege and Telle (2001) provided experimental evidence suggesting that social rewards and punishments affect behaviour. Rege and Telle show this in the context of a ten-person public goods experiment in which each contribution to the public good reduces the material payoff of the contributor. Every dollar contributed to the public good increases the material payoff of each of the ten group members by 20 cents, i.e. the contributor loses 80 cents. In the baseline condition of this experiment subjects' contribution to the public good remains anonymous. Neither the experimenter nor the other subjects know a subject's contribution. In the approval-condition both the other subjects and the experimenter can observe each subject's contribution. Note also that in both conditions the experimenters recruited subjects that were strangers to each other. In the baseline condition subjects contributed

34 per cent of their endowment to the public good while in the approval condition the contributions were twice as high. A plausible interpretation of this is that in the approval condition subjects feared the disapproval of the other group members.[14]

This interpretation is supported by the results of Gächter and Fehr (1999) who also found that, given some minimal social contact among strangers, making individual contributions publicly observable raises contributions to the public good substantially. Beyond this Gächter and Fehr explicitly measured the positive and negative emotions that are the basis for social (dis)approval. They show that free riding elicits extremely strong negative emotions among the other group members. Moreover, in the post-experimental group discussions the other group members verbally insulted the free riders.

### 3.2. Social approval and economic incentives

If the desire to gain approval and to avoid disapproval affects people's behaviour it is natural to ask how this desire interacts with economic incentives. We would like to stress that we consider our arguments in this context as quite preliminary and speculative. Apart from a few theoretical and empirical studies little is known in this area. Yet, scientific considerations have to start somewhere and the relevance of the approval motive suggests that this is a potentially fruitful field for further enquiry.

There are cases in which economic rewards and punishments work in the same direction as the approval motive. If an employee publicly receives a bonus for good performance the employee will also often receive the admiration of the colleagues. Likewise, if an employee is denied a bonus for violating legitimate rules at the workplace, and if the colleagues know this, then the monetary sanction will often go together with the colleagues' disapproval. Another example is given by the punishment of free riders in public goods situations. The emotions data in Gächter and Fehr (1999) suggest that free-riding causes a lot of anger among the cooperators and that this anger is anticipated by the potential free-riders. Fehr and Gächter (2000a) and Carpenter (2001) examined the hypothesis that the cooperators' anger will induce them to punish the free riders even if punishment is costly for the cooperators. For this purpose they implemented a public goods experiment with two stages. At stage 1 all group members simultaneously decided how much to contribute to the public good. For every (experimental) dollar invested into the public good each group member earned 40 cents, i.e., the investing member lost 60 cents but the group as a whole benefited from the investment. At stage 2 each group member was informed about the contribution of the others in the group. After this each member could punish the others by assigning points to them. For each point assigned the income of the punished group member was reduced by 10 per cent. Thus, the punishment of free riders constituted a material incentive to

[14] The fact that the experimenter observes the subjects' contributions is not likely to be important. There has been a debate whether observability by the experimenter affects subjects' behaviour in experiments. To our knowledge only Hoffmann et al. (1994) found an effect of experimenter-subject anonymity in dictator games, Bolton et al. (1998) as well as Johannesson and Persson (2000) found none. Bolton and Zwick (1995) found no significant effect of experimenter-subject anonymity in ultimatum games and Laury et al. (1995) found no effect in public goods games, either.

the extent to which it reduced the income of the free riders, and an approval incentive to the extent to which it expressed social disapproval. Fehr and Gächter (2000a) as well as Carpenter (2001) show that this opportunity to punish has a dramatic impact on cooperation. While cooperation unravels to extremely low levels in the absence of a punishment opportunity, almost full cooperation can be established in the presence of a punishment opportunity. The approval dimension of the punishment is supported by the recent study of Masclet et al. (2001). These authors allow the subjects in a public goods experiment to assign "disapproval points" to the other group members after the subjects have been informed about others' contributions. However, the disapproval points have no material consequences – they merely indicate disapproval. It turns out that disapproval alone raises the contributions to the public good relative to the baseline with no punishment opportunities, but the rise is lower compared to a situation where disapproval is associated with a material punishment.

The above examples suggest that economic incentives and approval incentives may reinforce each other. There are, however, reasons to believe that the relation between these two kinds of incentives is not always that straightforward. One complication arises because approval incentives are likely to cause strategic complementarity among the agents' actions, i.e., the strength of approval incentives depends on other people's behaviour. More specifically, the marginal social approval arising from an individual's praise-worthy behaviour is likely to depend positively on the average level of the others' praise-worthy behaviour. This is indicated by the empirical results in Gächter and Fehr (1999). They show that an individual's gain in social approval arising from an increase in the contribution to a public good is the higher the higher the average contribution of the other group members.[15] An important consequence of this is that there may well be many levels of equilibrium contributions (see e.g., Lindbeck et al., 1997; Huck et al., 2001). If, e.g., the average contribution is high each individual faces high approval incentives. Therefore, the individual will also choose a high contribution. Likewise, if average contributions are low, the individual faces low approval incentives and, hence, will choose a low contribution.

Fig. 6 illustrates the case of multiple equilibria. In Fig. 6 we assume for simplicity that individual $i$'s level of compliance with a morally legitimate rule (i.e., the relative frequency of obeying the rule in a given time interval) is higher, the higher the average compliance of the others. If the bold line represents the reaction function of each individual there are three equilibria. There is a stable low-compliance equilibrium (point A), an unstable equilibrium (point B) and a stable high-compliance equilibrium (point C). Fig. 6 also illustrates that small changes in the environment that reduce an individual's compliance level may cause large behavioural effects because the high compliance equilibria may vanish. Suppose, e.g., that initially the high-compliance equilibrium C is played and that an exogenous change then shifts the reaction function of each individual to the dotted line. In this case only the stable low-compliance equilibrium remains so that we can expect a large reduction in the compliance level.

[15] Remember that reciprocity also introduces strategic complementarity among the group members' contributions to a public good (see Section 2.1).

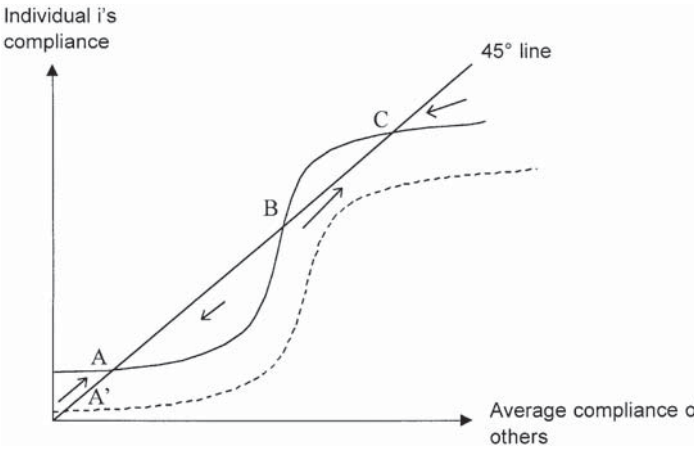


Fig. 6. Multiple equilibria in the presence of approval incentives.

The existence of multiple levels of equilibrium compliance has potentially important consequences. One consequence is that by expressing social values and providing information about compliance with these values the principal may affect the agents' beliefs, which in turn affects the process of equilibrium selection. In this way the law, by expressing certain values, acquires an expressive function (Kahane, 1996; Cooter, 1998; Bohnet and Cooter, 2001). Another interesting question is how the introduction of certain material incentives affects behaviour in the presence of multiple equilibria. A recently published experiment by Gneezy and Rustichini (2000a) suggests that there may be unexpected and intriguing complications. Gneezy and Rustichini studied the parents' response to the introduction of a fixed fine for picking up their children too late from Kindergarten. Parents who have their children in the Kindergarten during the day often are under time pressure and, therefore, they pick up their children too late relative to the established rules. These rules are typically part of the implicitly agreed upon terms of trade between the parents and the Kindergarten. Therefore, if the parents pick up their children too late they violate a legitimate rule. As a consequence, the parents face the disapproval of the principal and of the employees of the Kindergarten, which can be thought of as the non-pecuniary cost for being late.

In the experiment, which lasted for 20 weeks, there were two conditions. In the baseline condition parents just face the disapproval of the employees, i.e., there are no additional costs. In the other condition the experimenters implement a fixed fine after week four for picking up a child too late. The fine is removed after week 16. In weeks 5 and 6 the fine has little impact on the behaviour of the parents although in week 6 there is already a slight *increase* in the number of late comers. Then, from week 7 onwards, there is a steep *increase* in the number of late comers until their number is roughly twice as high as in the baseline condition. Moreover, when the fine is removed at the end of week 16 the number of tardy parents remains roughly twice as high as in the baseline condition.

An important aspect of this experiment concerns the way in which the fine was introduced. After week four parents simply found the following note on the bulletin board of the Kindergarten: "As you all know, the official closing time of the day-care center is 1600 every day. Since some parents have been coming late, we (with the approval of the "Authority for Private Day-Care Centers in Israel") have decided to impose a fine on parents who come late to pick up their children. As of next Sunday a fine of NIS 10 will be charged every time a child is collected after 1610. The fine will be calculated monthly, and it is to be paid together with the regular monthly payment". The parents tended to look at this board every day, since important announcements were posted there. Note that this announcement is quite ambiguous with regard to the moral message that is conveyed. While the term fine indicates that one should not pick up a child too late, the term "official closing time" suggests that in fact it is not so bad. In addition, since the fine is imposed only if somebody is late for more than 10 minutes the implicit message is that being late a little bit is not at all bad. Finally, the sentence that the fine "is to be paid together with the regular monthly payment" suggests to the parents that the fine is nothing else but a price for being late. As a consequence, it seems likely that this way of introducing the fine transformed the act of being late from a rule violation to a market transaction.[16] While in the baseline condition there was no ambiguity about the fact that being late constituted a violation of the rules the imposition of a price conveyed the message that the commodity of "being late" could now be bought. As a consequence, there was no longer a basis for disapproval and parents who were late may no longer have felt bad. Or put differently: Demanding a price for being late decreased the disapproval costs for the parents so that the total costs of being late may have been reduced. Thus, in terms of Fig. 6 the introduction of the fine may be interpreted as a downward shift in individuals' reaction functions which caused the break down of the high-compliance equilibrium C and a gradual shift to the low-compliance equilibrium A'.

The existence of multiple equilibria in situations involving social approval also provides a plausible explanation for the fact that the removal of the fine did not induce the parents to return to pre-fine compliance levels. It is well known from literally hundreds of experiments that behavioural changes to exogenous shifts typically occur gradually. Subjects rarely jump to a new equilibrium but they gradually converge in a piecemeal fashion to a new equilibrium. Thus it seems likely that, after the removal of the fine, the parents were caught in the low-compliance equilibrium A because point A is much closer to point A' than to point C. In fact, if the parents had adaptive expectations this is what one could have expected. Taken together, the stylized facts of Gneezy and Rustichini (2000a) can therefore be neatly explained by the interaction between approval incentives and material incentives.

There is also another experiment by Gneezy and Rustichini (2000b) suggesting that the introduction of explicit monetary incentives may weaken approval incentives. This experiment involves Israeli high school children who are doing volunteer work. Every

[16] This interpretation means that the perception of the fine as a price for being late may depend on the framing of the fine. If the fine is unambiguously associated with the perception that being late constitutes a violation of the rules the fine may have a different effect.

year, on a predetermined day, students go from house to house collecting monetary donations that households make to societies for cancer research, assistance to disabled children, etc. To induce the children to perform these activities they typically receive much social approval from parents, teachers and other people. Note that it is the very fact that they perform these activities voluntarily without monetary compensation that deserves to be approved. Paying the children money for their activity removes, therefore, the basis for social approval. Or put differently: The monetary reward reduces the approval reward. One implication of this argument is that the introduction of a money reward may well reduce the intensity with which the children collect money. This is indeed the finding of Gneezy and Rustichini (2000b). When the children are promised that they can keep 1 per cent of the money collected the amount collected is reduced by 36 per cent and when they are promised that they can keep 10 per cent of the money collected the reduction in the amount collected is still 8 per cent. This is compatible with the view that the introduction of a money reward causes a fixed reduction in the approval reward but that further increases in the monetary incentive have no further detrimental effects on the approval reward.

We believe that the above argument holds for other types of moral behaviour as well. Moral behaviour is often considered to be moral for the very reason that it is undertaken despite pecuniary incentives to the contrary. Paying people for their moral behaviour is, therefore, a contradiction in itself because it means that their behaviour can no longer be considered as moral. For example, if you are paid for your honesty most people will no longer evaluate your honest behaviour as moral behaviour. Since moral behaviour typically is associated with social approval, paying for moral behaviour means that approval incentives will be reduced.

There is one additional complication here. If people know that somebody engages in a moral behaviour *solely* because the person expects to receive social approval they probably will no longer consider the behaviour of the person as moral. We seem to approve of moral behaviour because it is not driven by external incentives. This problem is, however, not as severe as it might seem because the desire for social approval is typically closely connected to the desire to *deserve* social approval. The close link between the desire to receive approval and the desire to deserve approval has already been beautifully described by Smith (1759, p. 166): "Man naturally desires, not only to be loved, but to be lovely; ... He naturally dreads, not only to be hated, but to be hateful; ... He desires not only praise, but praise-worthiness; ... He dreads not only blame, but blame-worthiness". Social approval is therefore closely related to self-approval.[17] An important consequence of this is that moral behaviour is not only exhibited if the actor's behaviour is observed so that the actor can *actually* expect social approval. If actors also want to be worthy of praise they engage in the moral behaviour even when unobserved. Applied to the money collection experiment of Gneezy and Rustichini this means that the introduction of a monetary reward does not only reduce the social approval the children receive, but also the children's self-approval for their activity. The children consider themselves as less praise-worthy when they collect money, which reduces the psychological incentive to perform the activity. Thus, the negative effect of the

[17] Adam Smith basically spelled out elements of a Freudian theory of the superego.

introduction of the money reward may occur irrespective of whether others know that the children are paid. Likewise, if actors not only fear the actual social disapproval but they want to avoid that they are blame-worthy, they tend to avoid violating legitimate rules even in the absence of social disapproval. Applied to the Kindergarten experiment this means that the introduction of the fine not only reduces the disapproval for being late but parents also no longer consider being late as blame-worthy.

### 3.3. The management of social norms

Social (dis)approval is a key element in the enforcement of social norms. Therefore, the interactions between economic incentives and social approval also have implications for the enforcement of social norms. In particular, rewarding people monetarily for obeying social norms may weaken norm enforcement and may, hence, lead to a gradual erosion of norm-guided behaviour. Likewise, giving potential norm violators the opportunity to free themselves from following a social norm by making them pay for the norm violation may backfire for the same reason that making parents pay for being late had a counterproductive effect on parents' behaviour.

This insight has also potentially important implications for the kind of punishment that a society chooses to deter norm violations. From a strictly economic viewpoint it has always been a puzzle why modern societies frequently put norm violators into prison given that imprisonment consumes a lot of resources and deterrence can also be achieved much cheaper by threatening to fine norm violators. However, our considerations suggest that it may be unwise for a society to replace imprisonment by monetary fines to enforce important norms. The reason is that imprisonment and fining may convey very different moral messages. While imprisonment unambiguously conveys the message that the norm violator conducted morally wrongful acts, fining people may transform norm violations into a kind of market transaction.[18] Likewise, giving the convicted norm violators the choice between imprisonment and fining is problematic either because it means that at least those who can afford to pay the fine will prefer the fine while the rest of the people will have to choose imprisonment. This is also likely to be detrimental for most people's willingness to comply voluntarily with the norm because voluntary compliance is conditional on the compliance of other people. Public order and the absence of crime are public goods and we know that people's willingness to contribute to public goods heavily depends on their perceptions of others' contributions (see Section 2.1 and Gächter and Fehr (1999) and Falk and Fischbacher (forthcoming)). Allowing even only a minority of the people to free themselves, although at some cost, from obeying the norm may trigger the unravelling of the social norm. Thus, if a society wants to mobilize the incentives arising from social (dis)approval for the enforcement of norms it should choose forms of punishment that make unambiguously clear that norm violations are morally wrong. This is so because

[18] There are of course also other reasons (e.g., wealth constraints) why imprisonment may be the preferred sanction. See also our discussion on conditional cooperation in Section 2.1: It may not only be wise for organizations to exclude norm violators from interacting with cooperative co-workers but also for the society as a whole to limit the interaction of norm violators and norm followers to a minimum.