Four Essays in Microeconomics: Social Norms and Social Preferences

Vom Fachbereich Wirtschaftswissenschaften
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doctor rerum politicarum (Dr. rer. pol.)
genehmigte

Dissertation

vorgelegt von

M.A. *Simon Niklas Koch*

Tag der mündlichen Prüfung:	23.11.2022
Dekan:	Prof. Dr. Jan Wenzelburger
Vorsitzender:	Prof. Dr. Jan Wenzelburger
Berichterstattende:	1. Prof. Dr. Philipp Weinschenk
	2. Prof. Dr. Daniel Heyen

D 386
2022

Acknowledgments

I am deeply indebted to my supervisor Philipp Weinschenk, for showing me how to do and enjoy economic research and for always offering an open ear and a helping hand. Thank you, Philipp.

I thank my second supervisor Daniel Heyen and the chair of the doctoral committee Jan Wenzelburger for always being supportive in this endeavor.

I thank Valentina Eppich for proofreading the manuscript excellently.

I thank my parents, Elisabeth and Werner Koch, for their generous support in all aspects of life and my sister Anne for teaching me to set one foot in front of the other.

Contents

I Introduction

This thesis contributes to the economic literature on social motivations for behavior. Social motivations and psychological motivations for human behavior had gone missing in economics until they were rediscovered by modern behavioral eoncomics (Kahneman and Tversky, 1979). Social aspects of the human condition as motivating factors for actions had been present in the texts of classical economists like in Adam Smith's Theory of Moral sentiments (cf. Festré, 2010, p. 512.), or the works of Alfred Marshall and John Stuart Mill (cf. Burke and Young, 2011, p. 312.). Cognitive and social motivators complement self-interested motivation as factors that drive human endeavors. With respect to social motivators for human behavior, there are two main lines in modern behavioral economics that incorporate social motivations. On the one hand, social preferences, where the individual is assumed to care to some extend about the payoff of another individual. On the other hand, social norms, which restrict the possible action set of individuals by prescribing and proscribing certain behavior in certain situations.

Social preference analysis developed in economics because empirical results on human behavior could not be reconciled with money-maximizing behavior which neoclassical accounts had predicted (Güth et al., 1982). People seemed to care about the welfare of other players in laboratory experiments. The model of inequity aversion by Fehr and Schmidt (1999) is possibly the most widely known example of models that incorporate social preferences. However, the idea has disseminated into many different subsections of economics. It has gained particular traction with the agency literature. Here, the interplay of social preferences of one or both sides of a contractual arrangement for the actual contract decision is in focus. How do agents of different types, i.e. different social preferences, react to different contractual offers and can the principal devise mechanisms to improve the output or screen for different types of agents are important research questions (Besley and Ghatak, 2005).

Besides social preferences, the second social motivator for human behavior is considered to be social norms. Social norms are a widely used concept in all social sciences. Social norms are seen as social devices that increase conformity in a society. The sociological caricature of homo sociologicus is following every social prescription without reflection, similarly to how homo oeconomicus is a caricature of a human being without social relations only driven by greed (Elster, 1989). Social norms influence human behavior by setting standards for behavior. These social norms are produced in society and permeate society at all levels. From a neoclassical economic perspective, social norms are difficult to understand. Why should homo oeconomicus contribute to the costly maintenance of social norms, when free-riding is always an option (Schurtenberger, 2018)? Divergent economic voices have postulated that social norms are a way for human beings to solve coordination problems and social dilemmas. If norms prescribe to drive on the right hand side of the road it makes sense to follow that rule, since otherwise coordination costs would be high. If norms of cooperation can be sustained in a world of free-riding, beneficial outcomes can be achieved in society (Ullmann-Margalit, [1977] (2015); Sugden, [1986] (2005)).

This thesis contributes to the literature on both categories of social motivators and contains four papers that are presented in chapters II through V. The first three of them, in chapter II, III, and IV, contribute to the agency literature on social preferences. In particular, a simple model of social preferences is proposed that nonetheless delivers interesting new insights with respect to screening, monitoring, and non-monetary incentives. The fourth paper in, chapter V, is a review of theoretical models of social norms in economics. The goal of this paper is to ease access to these economic models for all interested behavioral scientists. The remainder of this introduction will be dedicated to a short synopsis of the four papers that constitute this thesis.

Synopsis:

The paper ”Contract Design with Socially Attentive Preferences” , published almost identically as Koch and Weinschenk (2021) in *Games and Economic Behavior*, is presented in chapter II and introduces the concept of socially attentive preferences, where the agent cares to some extent about the payoffs of the principal and a third party. First, a theoretical contribution of Demougin and Fluet (1998) for selfish agents is generalized by showing that the result stays valid with socially attentive agents. This generalization shows that the optimal contract has a simple binary structure in agency models with limited liability. Monetary incentives are still effective with social preferences, but it might be optimal for the principal to set none. That is because an agent with socially attentive preferences reacts less strongly to monetary incentives than an egoistic agent does. Hence, the principal’s costs of additional incentives to increase the motivation of the agent might be too high to be justified. This constitutes a potential explanation for the puzzle why many employment contracts specify no or only weak monetary incentives. With asymmetric information concerning the agents level of social attentiveness, the principal, surprisingly, does, optimally, not screen for the agent’s type but offers a simple pooling contract. The reason is that with a menu of contracts, in order to guarantee incentive compatibility, the principal would have to reward some types of agents for the unsuccessful outcome and offer them relatively low incentives in case of success. Thus, the incentives to these types are inefficiently weak and the principal can improve by scraping these contracts. This is a new result in the literature on social preferences. The result is empirically relevant, since most firms do not offer menus of contracts but resort to contract posting. This is also a possible explanation why employment contracts display a high degree of uniformity.

The paper ”Contractible and Non-contractible Efforts with Socially Attentive Preferences”, in chapter III, tests the robustness of results in the agency literature by showing whether they stay valid with socially attentive preferences. A principal still implements an inefficiently low effort level with a wealth-constrained agent, but social preferences attenuate the issue. Contractibility of effort allows the principal to implement the efficient effort level in standard agency models. However, this needs not always be the case with socially attentive preferences. Either, for relatively low levels of social attentiveness, because the principal and the agent do not consider the impact of the agent's action on the third party sufficiently, or, for relatively high levels of social attentiveness, because the principal exploits the agent and demands an excessively high effort level. The latter effect can decrease the generated surplus the relationship generates. The generated surplus can be lower in case of contractibility than in case of non-contractibility if the agent is sufficiently socially attentive, again due to the over-implementation of effort by the principal. This provides an efficiency argument for limits on specifications in employment contracts and the principal's monitoring control.

In the paper presented in chapter IV, ”Endogenous Socially Attentive Preferences”, the possibility to increase the level of social attentiveness of the agent and the possibility of a socially attentive principal are introduced. The paper is connected to the literature that focuses on non-monetary forms of motivation for employees. We allow the principal to invest into the socially attentive preferences of an agent. This investment could consist of highlighting the positive impact the agent's work has on society or in devising a mission of the company that aligns with the ideals of the agent. Technically, the costly investment increases the level of social attentiveness of the agent. While it is not always optimal for the principal to incentivize the socially attentive agent to exert higher effort with monetary incentives, it is always optimal to motivate such an agent to exert higher effort by investing into their level of social preferences. However, the principal's investment is, in general, not socially optimal. The reasons for this are, first, that the principal ignores the higher effort costs implied by a higher effort level and, second, the positive effect of a higher effort on the payoff of the third party. This outcome can be improved when allowing for both the principal and the agent to have socially attentive preferences. In this case, the principal will take an investment decision that is closer to the social optimum, because the two effects for the agent and the third party are included into the effort implementation program to the extent that the principal is socially attentive.

Chapter V, ”Economic Modeling of Social Norms”, presents a review of the theoretical economic literature on social norms. In order to arrive at a definition of social norms the points of concordance are condensed and the points of disagreement discussed. Social norms are further distinguished from social preferences and identity economics. The different topics norm analysis has been used for in economics are reviewed. This serves as a step to bring the different behavioral sciences closer together, since it allows for simple access to theoretical economic contributions on the subject. Representative approaches are presented for each norm category of the paper and their merits analyzed. All contributions are presented and contextualized in how they refer to and add to the perspective on social norms in economics. The theoretical focus is softened to some extend to include current issues of empirical economic research with regards to norms, in order to highlight possible future ways of modeling social norms in economics.

II Contract Design with Socially Attentive Preferences

Joint work with Philipp Weinschenk

The standard agency model assumes that the agent does not care how his decisions influence others. This is a strong assumption, which we relax. We find that, although monetary incentives are also effective with socially attentive agents, the principal may optimally set none. This could explain the puzzle why empirically only a fraction of employees experiences monetary incentives. Furthermore, in case the agent's type is private information, the principal optimally offers a single pooling contract, i.e., never screens for different types, no matter how rich the set of possible attentiveness levels is and what shape the underlying distribution function has.

JEL Classification: D82, D91, M52.

Keywords: agency model, socially attentive preferences, incentives.

II.1 Introduction

There is rich evidence that many decision-makers take into account how their actions influence the well-being of others.[1] In standard agency models, it is nonetheless assumed that agents do not care how their decisions influence others. This is a strong assumption, which we relax in this paper.

We study an agency model, where an agent acts on behalf of a principal and the agent's non-contractible effort influences the probability distribution over outcomes. The outcome affects the principal, the agent, and possibly a third party.[2] We augment the agent's preferences by allowing him[3] to be socially attentive, i.e., put weight on others' utilities.[4]

[1] Andreoni and Miller (2002) document that only a quarter of persons are selfish money-maximizers. In Engel's (2011) meta study the share is one third.

[2] Depending on the context, the third party could be interpreted as customers, the ecological environment, or other employees.

[3] We follow the standard convention in the agency literature and talk about a male agent and a female principal.

[4] The literature (see, for example, Andreoni et al., 2007 or DellaVigna et al., 2019) puts forward two reasons for social attentiveness: altruism and warm-glow.

The analysis reveals that monetary incentives are not only effective when the agent is egoistic, but also if he is socially attentive. That is, a contract that specifies a higher remuneration for a successful outcome motivates the agent to exert more effort.[5] Examining the structure of the optimally designed contract, we first show that, in the benchmark of an egoistic agent, the optimal contract includes monetary incentives. This is in contrast to the case with socially attentive preferences: When the agent is sufficiently socially attentive, the principal optimally refuses to provide incentives, even though these are effective. The intuition for this result is that, since a socially attentive agent reacts less strongly to incentives than an egoistic one, the principal's costs of providing incentives may – in comparison to the benefit of incentives, in the form of a more motivated agent – be too high to justify incentives. This is no limit result, i.e., providing no incentives could also be optimal for moderate levels of social attentiveness. Under a regularity condition, the optimal incentives are monotonically decreasing in the agent's social attentiveness. The optimal monetary incentives are hence always lower with socially attentive preferences than with egoistic preferences. In summary, with a socially attentive agent, the principal optimally sets either no incentives or incentives that are rather weak. This result is empirically relevant, since it could explain the puzzle that – contrary to the predictions of the standard theory – many employees experience no financial incentives or rather weak incentives.

We also examine the scenario where not only the agent's effort, but also the agent's level of social attentiveness is private information. That is, there is both moral hazard and adverse selection. We show that screening for the different types of agents is never optimal. The principal thus optimally designs a pooling contract and this holds no matter how rich the set of potential levels of social attentiveness of her agent is and what shape the underlying distribution function has. This result might also be interesting from an empirical perspective, since it provides a theoretical foundation for the uniformity of observed contracts.[6] We further show that, because the average effect of incentives is important for the design of the optimal contract, the principal provides a pooling contract where all types experience either no or relatively weak incentives, in comparison to the benchmark with a surely egoistic agent. We also show that, when higher levels of the agent's social attentiveness are more likely, in the sense of first-order stochastic dominance, the optimal incentives are weaker.

Relation to the literature. The empirical literature documents that many employees experience no financial incentives or rather weak incentives. In the representative US sample of Lemieux et al. (2009), only 37% of workers are in performance-pay jobs and

[5] While monetary incentives are effective in standard agency models, e.g. Hölmström (1979), they could be ineffective in models of crowding out of intrinsic motivations; see, for example, Bénabou and Tirole (2006).

[6] Empirically, most firms do not offer menus of contracts. The majority of employment contracts is determined by wage posting; see, for example, Brenzel, Gartner, and Schnabel (2014).

the median share of performance pay is only about 3.5% of total earnings. Bryson et al. (2012) confirm this finding for the US and document that these payment structures are even less common in Europe. In most EU-15 countries, only 10%-25% of workers are in performance-pay jobs. Gittleman and Pierce (2013) use a different definition of performance-pay jobs and a more recent sample of the same data as Bryson et al. (2012), and show that in the US, only 20% of employees receive a performance-related pay. Moreover, if monetary incentives exist, they are in practice often weaker than predicted by standard theory (Williamson, 1985; Holmström and Milgrom, 1990; Che and Yoo, 2001). By relaxing the assumption that agents do not care how their decisions influence others, our paper generates predictions that are perfectly in line with the empirical findings: with socially attentive preferences, the optimally designed contract provides either no incentives or rather weak incentives.

Our paper also contributes to the theoretical literature that incorporates different forms of social preferences into agency models. Itoh (2004), Englmaier and Wambach (2010), and Bartling (2011) examine the effects of inequity-averse agents. Itoh (2004) and Bartling (2011) also consider agents who are status-seeking. Itoh (2004) shows that the principal is in general worse off if her agent cares more about the inequity between the principal and the agent. In contrast, in the presence of multiple agents, the principal could benefit from the agents' inequity aversion that concerns other agents by designing an appropriate interdependent contract that specifies wage payments as a function of all agents' performances. Englmaier and Wambach (2010) show that the compensation scheme converges to a linear sharing scheme as the concern for equity among agents becomes sufficiently important. Furthermore, the optimal contract may be team-based and overdetermined or incomplete. Bartling (2011) shows that team contracts can be optimal even when there is a positive correlation between the agents' performance measures. Kräkel (2016) analyzes peer effects in a multi-agent setting, where the comparisons with other agents influence each agent's motivation. He shows that, depending on the interplay of the peer effects, agents' efforts are either strategic complements or strategic substitutes. Besley and Ghatak (2005) examine motivated agents. They show that the matching of the mission preferences of principals and agents is important for organizational efficiency. Cassar (2016) models a situation where the principal's and the agent's project preferences are misaligned and analyzes what mission the principal optimally chooses in different contractual environments.

We finally add to the literature that considers adverse selection problems with social preferences. Delfgaauw and Dur (2007) examine a model with differently motivated agents who can apply for a job vacancy offered by a monopsonistic firm and show that the firm can screen agents by setting a threshold of motivation. Arce (2013) develops a model where social preferences establish a standard for effort and finds that the principal can screen between egoistic and social agents, but not between agents with different forms of social preferences. Non (2012) studies a model with possibly altruistic agents and principals and shows that an altruistic principal may find it optimal to signal her type and to screen for the agents' types.

II.2 Model

We next incorporate socially attentive preferences into a standard agency model.

Primitives. A principal needs to hire an agent. Both are risk-neutral. When working for the principal, the agent exerts effort $e \in [0, e^{\max}] \subseteq \mathbb{R}$, where $e^{\max}$ is positive and could be finite or infinite. The set of possible outcomes is $\{1, \ldots, n\}$, with $n \geq 2$. The agent's effort choice determines the probability distribution over the outcomes, where $p_i(e)$ denotes the probability that outcome $i \in \{1, \ldots, n\}$ realizes. The realized outcome affects the principal's return R_i and possibly also a third party, whose payoff is denoted by V_i. While the outcome is contractible, the agent's effort is non-contractible, i.e., there is moral hazard. A contract is thus a vector of outcome-contingent payments $(t_1, \ldots, t_n)$, where $t_i \in \mathbb{R}$ denotes the payment in case outcome $i \in \{1, \ldots, n\}$ realizes. The agent's liability is limited due to wealth or legal constraints, such that the payment cannot fall short of a threshold $\underline{t} \geq 0$.

Preferences. We deviate from the textbook moral-hazard model by allowing the agent to take into account how his decisions influence others. Formally, the agent puts a weight $\beta \in [0, 1]$ on the utilities of others.[7] We henceforth say that the agent is egoistic if he puts zero weight on others' utilities, $\beta = 0$, while the agent is socially attentive if he puts positive weight on them, $\beta > 0$. Unless explicitly stated differently, we suppose that the agent puts at least slightly more weight on his own utility than on that of others, $\beta < 1$. The agent's utility in case outcome $i \in \{1, \ldots, n\}$ realizes is hence

$$u_A = t_i - c(e) + \beta u_P + \beta u_T, \tag{1}$$

where $c(e)$ are the agent's effort costs, u_P is the principal's utility, and u_T is the third party's utility. The principal's utility equals the difference between the return she earns and the payment she makes, $u_P = R_i - t_i$, such that

$$E[u_P] = \sum_{i=1}^{n} p_i(e)(R_i - t_i). \tag{2}$$

[7] We thus use the weighted utilitarian approach, which is widely applied in many areas of economics and praised for its tractability, normative transparency, and axiomatic foundations (Balasubramanian, 2015). One could easily allow the agent to put different weights on the principal's utility and the third party's utility. Setting equal weights simplifies the exposition without much loss of generality, since different weights have the same effect as a variation of the third party's payoffs.

Since the third party's utility is $u_T = V_i$, her expected utility is

$$E[u_T] = \sum_{i=1}^{n} p_i(e)V_i. \tag{3}$$

Using the previous formulas, we can write the agent's expected utility as

$$E[u_A] = \sum_{i=1}^{n} p_i(e)\left(t_i + \beta(R_i - t_i + V_i)\right) - c(e). \tag{4}$$

Assumptions. We assume that c and p_i are twice continuously differentiable, $c'(e)$, $c''(e) > 0$ for $e > 0$, $c(0) = c'(0) = 0$, $\lim_{e \to e^{\max}} c'(e) = \infty$, and $\frac{p_i'(e)}{p_i(e)} < \frac{p_{i+1}'(e)}{p_{i+1}(e)}$ for all $i \in \{1, \ldots, n-1\}$.[8] Thus, the effort cost function is increasing, convex, satisfies limit conditions, and the outcomes are ordered according to their likelihood ratios. To guarantee the existence of a solution, we further suppose that $p_n'(e) >$ and $p_n''(e) \leq 0$, i.e., that the probability of outcome n to realize is increasing and concave in effort, and that effort is essential at least for the highest outcome, $p_n(0) = 0$.[9] Finally, the expected return $E[R_i] = \sum_{i=1}^{n} p_i(e)R_i$ is non-negative and increasing in effort and the third party's expected utility $E[u_T] = \sum_{i=1}^{n} p_i(e)V_i$ is non-negative and non-decreasing in effort.

Timing. First, the principal offers a contract to the agent, who then decides whether to accept or reject it. If the agent rejects, he receives a reservation utility of zero and the game ends. In case of acceptance, the agent then decides which effort to exert. Finally, the outcome is realized and the agent receives the contracted payment.

II.3 Analysis

II.3.1 Problem reduction

The principal's problem is to design a contract $\{t_1, \ldots, t_n\}$ that maximizes her expected payoff $E[u_P]$ subject to:

(i) the participation constraint $E[u_A] \geq 0$,

(ii) the incentive constraint that the implemented effort $\hat{e}$ satisfies $\hat{e} \in \operatorname{argmax} E[u_A]$,

(iii) the limited liability constraints $t_i \geq \underline{t}\ \forall i \in \{1, \ldots, n\}$.

Since the agent chooses his effort optimally and his liability is limited, it holds that

$$E\left[u_A | e^*\right] \geq E[u_A | e = 0] = \sum_{i=1}^{n} p_i(0)\left(t_i + \beta(R_i - t_i + V_i)\right) \geq 0. \tag{5}$$

[8] The monotone likelihood ratio property is notationally convenient, but not necessary for our results. We could weaken it by only requiring that there exists an outcome i for which the likelihood ratio is maximal. We thank an anonymous referee for highlighting this point.

[9] All results hold also when $p_n(0)$ is positive, but small.

Thus, the participation constraint is automatically satisfied, which is standard in agency models with limited liability.

We next show that the optimal contract has a simple binary structure. To be precise, the optimal contract is such that, if any outcome, only the outcome with the maximal likelihood ratio is rewarded. This insight generalizes Demougin and Fluet (1998), who have shown this for the case of standard (egoistic) preferences.

Proposition 1: *The principal optimally sets a contract $\left(t_1^*,\dots,t_n^*\right)$ which satisfies $t_i^* = \underline{t}$ for all $i \in \{1,\dots,n-1\}$ and $t_n^* \geq \underline{t}$.*

PROOF: Suppose, contrary to Proposition 1, there exists a contract $\{\tilde{t}_1,\dots,\tilde{t}_n\}$ with $\tilde{t}_j > \underline{t}$ for at least some $j < n$ which yields the principal a higher expected utility than any contract $\left(t_1^*,\dots,t_n^*\right)$ which satisfies $t_i^* = \underline{t}$ for all $i \in \{1,\dots,n-1\}$ and $t_n^* \geq \underline{t}$. The implemented effort level is denoted by $\tilde{e}$. Suppose first that $\tilde{e} = 0$. The contract $t_j = \underline{t}$ for all $j \in \{1,\dots,n\}$ causes the implementation of effort $e \geq \tilde{e} = 0$ for the minimal expected payment of $\underline{t}$, which by

$$E[u_P] = \sum_{i=1}^{n} p_i(e)(R_i - t_i) = E[R_i|e] - E[t_i|e]. \tag{6}$$

contradicts the claim.

Suppose next that $\tilde{e} > 0$. Since the contract $\{\tilde{t}_1,\dots,\tilde{t}_n\}$ implements effort $\tilde{e}$, it must hold that $\tilde{e} \in \operatorname{argmax} E[u_A|\{\tilde{t}_1,\dots,\tilde{t}_n\}]$, which particularly implies that the first-order condition of the agent's problem, which is necessary for an optimum, is satisfied:

$$\left.\frac{\partial E[u_A|\{\tilde{t}_1,\dots,\tilde{t}_n\}]}{\partial e}\right|_{\tilde{e}} = \sum_{i=1}^{n} p_i'(\tilde{e})\left(\tilde{t}_i + \beta(R_i - \tilde{t}_i + V_i)\right) - c'(\tilde{e}) = 0. \tag{7}$$

We next argue that the principal can improve by modifying the contract $\{\tilde{t}_1,\dots,\tilde{t}_n\}$.

The first modifications apply if $\tilde{t}_j > \underline{t}$ for some $j < n$ with $p_j'(\tilde{e}) > 0$. Let the principal pick some j for which this is true and set $t_j = \underline{t}$ and increase t_n by $\frac{p_j'(\tilde{e})}{p_n'(\tilde{e})}(\tilde{t}_j - \underline{t})$. By construction, this leaves the first-order condition unaffected, but changes the principal's expected payment by

$$-p_j(\tilde{e})(\tilde{t}_j - \underline{t}) + p_n(\tilde{e})\frac{p_j'(\tilde{e})}{p_n'(\tilde{e})}(\tilde{t}_j - \underline{t}) = p_j'(\tilde{e})\left(-\frac{p_j(\tilde{e})}{p_j'(\tilde{e})} + \frac{p_n(\tilde{e})}{p_n'(\tilde{e})}\right)(\tilde{t}_j - \underline{t}), \tag{8}$$

which is negative since outcome j has a lower likelihood ratio than outcome n. Repeat this modification for all payments $\tilde{t}_j > \underline{t}$ with $j < n$ and $p_j'(\tilde{e}) > 0$.

The final modifications concern payments $\tilde{t}_j > \underline{t}$ with $j < n$ and $p_j'(\tilde{e}) \leq 0$. The principal can improve by lowering t_j to $\underline{t}$ and lowering the payment t_n

Case (i): by $\frac{p_j'(\tilde{e})}{p_n'(\tilde{e})}(\tilde{t}_j - \underline{t})$ if t_n is then still not below $\underline{t}$ or
Case (ii): to $\underline{t}$ if the former modification causes t_n to fall below $\underline{t}$.

Repeat this modification for all payments $\tilde{t}_j > \underline{t}$ with $j < n$ and $p_j'(\tilde{e}) \leq 0$.

After the modifications, the contract has the structure $t_i = \underline{t}$ for all $i \in \{1,\dots,n-1\}$ and $t_n \geq \underline{t}$. This contract satisfies the limited liability constraints. Moreover, since $p_n''(e) \leq 0$, the agent's expected utility is concave in effort for any such contract such that the first-order condition (7) is necessary and sufficient. If Case (i) applied for all modifications, the modifications cause by construction that the principal implements the same effort $\tilde{e}$ for a lower expected payment, such that the contract $\{\tilde{t}_1, \ldots, \tilde{t}_n\}$ cannot be optimal. Otherwise, the principal's expected payment lowers to $\underline{t}$ and the implemented effort level increases, again implying that the contract $\{\tilde{t}_1, \ldots, \tilde{t}_n\}$ cannot be optimal. □

Proposition 1 shows that the case with more than two outcomes effectively reduces to the two-outcome case. Without loss of generality, we thus henceforth concentrate on the two-outcome case. We interpret outcome $i = 1$ as the unsuccessful outcome and outcome $i = 2$ as the successful outcome and abbreviate by writing $p(e)$ for $p_2(e)$.

II.3.2 Optimal contract

First-order approach. Since the agent's expected utility is concave in effort for all contracts that satisfy Proposition 1, the global incentive constraint $\hat{e} \in \operatorname{argmax} E[u_A]$ can be substituted by the local incentive constraint

$$\left.\frac{\partial E[u_A]}{\partial e}\right|_{\hat{e}} = p'(\hat{e})\left(\Delta t + \beta(\Delta R - \Delta t + \Delta V)\right) - c'(\hat{e}) = 0, \tag{9}$$

where $\Delta t := t_2 - t_1$ is the payment spread, $\Delta R := R_2 - R_1 > 0$, and $\Delta V := V_2 - V_1 \geq 0$.[10] Implicitly differentiating (9) yields that

$$\frac{\partial \hat{e}}{\partial \Delta t} = -\frac{p'(\hat{e})(1-\beta)}{p''(\hat{e})\left(\Delta t + \beta(\Delta R - \Delta t + \Delta V)\right) - c''(\hat{e})} > 0. \tag{10}$$

Thus, with an egoistic agent ($\beta = 0$), as well as with a socially attentive agent ($\beta > 0$), monetary incentives are effective, in the sense that the agent exerts more effort, the higher the monetary incentives Δt are.

Existence and structure of optimal contract. We next show that an optimal contract exists and characterize its structural properties.

Lemma 1: *There always exists an optimal contract $\left(t_1^*, t_2^*\right)$. The optimal contract satisfies $t_1^* = \underline{t}$ and $t_2^* \in [\underline{t}, \underline{t} + \Delta R)$.*

PROOF: By Proposition 1, $t_1^* = \underline{t}$. Suppose next that, contrary to our claim, $t_2^* \geq \underline{t} + \Delta R$. Then

$$E[u_P] = p(e^*)(R_2 - t_2^*) + (1 - p(e^*))(R_1 - t_1^*) \leq R_1 - \underline{t}. \tag{11}$$

[10] $\Delta R > 0$ follows since $E[R_i] = p(e)R_2 + (1 - p(e))R_1$ and $p(e)$ were assumed to be increasing in effort. $\Delta V \geq 0$ follows since $E[u_T] = p(e)V_2 + (1 - p(e))V_1$ was assumed to be non-decreasing in effort.

The principal can improve by setting the contract $\left(t_1 = \underline{t}, t_2 = \underline{t} + \Delta R/2\right)$, since

$$E[u_P] = p(e)(R_2 - \underline{t} - R_2/2 + R_1/2) + (1 - p(e))(R_1 - \underline{t}) > R_1 - \underline{t}, \tag{12}$$

where we used that the agent optimally chooses a positive effort level for this contract. Since $t_2^* \geq \underline{t}$ by the agent's limited liability, an optimal contract must thus satisfy $t_2^* \in [\underline{t}, \underline{t} + \Delta R)$.

It remains to show that a payment t_2 exists that maximizes the principal's objective function $E[u_P]$. Since the agent's effort choice, cf. (9), and thus also the principal's expected utility

$$E[u_P] = p(e)(R_2 - t_2) + (1 - p(e))(R_1 - t_1) \tag{13}$$

is continuous in t_2 and we can restrict the search for an optimal value of t_2 to a closed and bounded interval $[a, b] \supset [\underline{t}, \underline{t} + \Delta R)$, the Bolzano-Weierstrass Extreme Value Theorem applies, such that an optimal t_2 and thus an optimal contract must exist. □

The intuition for the structural properties of the optimal contract are as follows. First, in case the agent does not succeed, it is optimal to make only the minimal possible payment; thus $t_1^* = \underline{t}$. Second, it is never optimal for the principal to provide a payment spread that equals or exceeds her spread of returns ΔR. Because it is further not possible to offer a payment below the threshold $\underline{t}$, we must have that $t_2^* \in [\underline{t}, \underline{t} + \Delta R)$.

Since the agent's effort choice depends on the monetary incentives and his social attentiveness, it is convenient to write his choice as a function of these variables: $\hat{e} = e(\Delta t, \beta)$. The principal's problem can thus be stated as

$$\max_{t_1, t_2} E[u_P] = p(e(\Delta t, \beta))(R_2 - t_2) + (1 - p(e(\Delta t, \beta)))(R_1 - t_1) \text{ subject to } t_1, t_2 \geq \underline{t}. \tag{P1}$$

Using that $t_2 = \Delta t + t_1$ and that by Lemma 1 $t_1^* = \underline{t}$, we can rewrite the problem:

$$\max_{\Delta t} E[u_P] = p(e(\Delta t, \beta))(R_2 - \Delta t - \underline{t}) + (1 - p(e(\Delta t, \beta)))(R_1 - \underline{t}) \text{ subject to } \Delta t \geq 0. \tag{P1'}$$

The principal hence optimally sets the contract $\left(t_1^* = \underline{t}, t_2^* = \underline{t} + \Delta t^*\right)$, where Δt^* solves the problem (P1'). This problem is nontrivial since one can, in general, not express the agent's effort choice $\hat{e} = e(\Delta t, \beta)$ in closed form.

Optimal incentives. We now examine whether the principal optimally provides incentives, $\Delta t^* > 0$, or refuses to set incentives, $\Delta t^* = 0$. In the former case, the agent's remuneration is variable since $t_2^* > t_1^*$, i.e., dependent on the outcome, while in the latter case the remuneration is fixed since $t_2^* = t_1^*$, i.e., independent of the outcome.

Differentiating the principal's expected utility yields that

$$\frac{\partial E[u_P]}{\partial \Delta t} = p'(e(\Delta t, \beta)) \frac{\partial e(\Delta t, \beta)}{\partial \Delta t}(\Delta R - \Delta t) - p(e(\Delta t, \beta)). \tag{14}$$

Consider first an egoistic agent, $\beta = 0$. If the principal provides no monetary incentives,

$\Delta t = 0$, the agent invests zero effort. By (10) and (14) it hence holds that

$$\left.\frac{\partial E[u_P]}{\partial \Delta t}\right|_{\beta=0,\Delta t=0} = \frac{\big(p'(e(0,0))\big)^2}{c''(e(0,0))}\Delta R > 0. \tag{15}$$

The principal could thus improve by increasing Δt. Providing monetary incentives is hence optimal. Interestingly, this not necessarily true with a socially attentive agent.

Proposition 2: *If the agent is egoistic ($\beta = 0$) or his social attentiveness is sufficiently low (β is small), the principal optimally sets monetary incentives, $\Delta t^* > 0 \iff t_2^* > t_1^*$. In contrast, the principal sets no monetary incentives, $\Delta t^* = 0 \iff t_2^* = t_1^*$, if the agent's social attentiveness β is sufficiently high.*

PROOF: We first show that the principal optimally sets monetary incentives, $\Delta t^* > 0$, if the agent's social attentiveness β is low. Suppose, contrary to the claim, that the principal sets no monetary incentive, $\Delta t = 0$. If β is sufficiently low, then we see from (14) that

$$\left.\frac{\partial E[u_P]}{\partial \Delta t}\right|_{\Delta t=0} = p'(e(0,\beta)) \left.\frac{\partial e(\Delta t,\beta)}{\partial \Delta t}\right|_{\Delta t=0} \Delta R - p(e(0,\beta)) > 0, \tag{16}$$

because the effort $e(0,\beta)$ and thus the success probability $p(e(0,\beta))$ are small, i.e., approach zero as $\beta \to 0$, whereas $p'(e(0,\beta))$, $\left.\frac{\partial e(\Delta t,\beta)}{\partial \Delta t}\right|_{\Delta t=0}$, and ΔR are positive and do not approach zero. Accordingly, the principal can improve her expected utility by increasing Δt.

It remains to be shown that the principal optimally sets no monetary incentives, $\Delta t^* = 0$, if the agent's social attentiveness β is sufficiently high. By (10) and (14), it holds that

$$\frac{\partial E[u_P]}{\partial \Delta t} = \frac{\big(p'(e(\Delta t,\beta))\big)^2 (\Delta R - \Delta t)}{c''(e(\Delta t,\beta)) - p''(e(\Delta t,\beta))\big(\Delta t + \beta(R - \Delta t + \Delta V)\big)}(1-\beta) - p(e(\Delta t,\beta)). \tag{17}$$

Select some $\hat{\beta} \in (0,1)$ and let the agent's social attentiveness be such that $\beta \in [\hat{\beta},1)$. By (17),

$$\frac{\partial E[u_P]}{\partial \Delta t} \le \frac{\left(\sup\limits_{\beta\in[\hat{\beta},1),\Delta t\in[\underline{t},\underline{t}+\Delta R)} \big\{p'(e(\cdot))\big\}\right)^2 \Delta R}{\inf\limits_{\beta\in[\hat{\beta},1),\Delta t\in[\underline{t},\underline{t}+\Delta R)} \{c''(e(\cdot))\} - \sup\limits_{\beta\in[\hat{\beta},1),\Delta t\in[\underline{t},\underline{t}+\Delta R)} \big\{p''(e(\cdot))\big\}\big(\hat{\beta}(\Delta R + \Delta V)\big)}(1-\beta)$$
$$- \inf\limits_{\beta\in[\hat{\beta},1),\Delta t\in[\underline{t},\underline{t}+\Delta R)} \big\{p(e(\cdot))\big\} \tag{18}$$

holds for all $\beta \in [\hat{\beta},1)$ and $\Delta t \in [\underline{t}, \underline{t} + \Delta R)$. Note that by continuity, all suprema and infima exist (but they need not be attained). Furthermore, the fraction on the right-hand side of (18) is positive due to $p'(e) > 0$, $c''(e) > 0$, and $p''(e) \le 0$, but finite and independent of β. Therefore, and because $\inf\limits_{\beta\in[\hat{\beta},1),\Delta t\in[\underline{t},\underline{t}+\Delta R)} \big\{p(e(\cdot))\big\} = p(e(0,\hat{\beta})) > 0$, the right-hand side of (18) is negative if β is sufficiently large. It is negative if and only if β belongs to the nonempty interval $(\max\{\dot{\beta},\hat{\beta}\},1)$, where $\dot{\beta}$ is such that the right-hand side of (18) is zero. Hence, for all $\beta \in (\max\{\dot{\beta},\hat{\beta}\},1)$ and $\Delta t \in [\underline{t}, \underline{t} + \Delta R)$, the derivative $\partial E[u_P]/\partial \Delta t$ is negative, implying that the principal optimally sets Δt^* as low as possible, namely $\Delta t^* = 0$. □

The intuition why the optimally designed contract may include no monetary incentives is as follows. First, although monetary incentives are also effective with socially attentive preferences, $\partial\hat{e}/\partial\Delta t > 0$, such preferences limit the effectiveness of incentives, since a socially attentive agent reacts less strongly to incentives than an egoistic agent, see (10). Second, providing monetary incentives is costly for the principal, see (2). Accordingly, if the agent is sufficiently socially attentive, the principal's costs of providing monetary incentives are too high, in comparison to their benefit (in the form of a more motivated agent), to justify monetary incentives. The principal then optimally refuses to provide monetary incentives and prefers to remunerate the agent with a constant, outcome-independent payment. It is worth emphasizing that this is no limit result; see the proof or the example below.

Additional result with concavity. We can derive an additional result if the principal's problem is concave. The next proposition shows that there is a monotone negative relationship between the incentive power Δt^* and the agent's social attentiveness β.

Proposition 3: *Let $c''' \geq 0$ and $p''' \leq 0$ such that the principal's problem is concave. If $\Delta t^* > 0$, the power of monetary incentives is monotonically decreasing in the agent's social attentiveness β: $\partial\Delta t^*/\partial\beta < 0$.*

PROOF: If $c''' \geq 0$ and $p''' \leq 0$, we directly see from (17) that the derivative $\partial E[u_P]/\partial\Delta t$ is decreasing in Δt and β. Suppose that $\Delta t^* > 0$. Then Δt^* solves the first-order condition $\partial E[u_P]/\partial\Delta t = 0$. If β increases, Δt^* must be lowered to restore the first-order condition, so $\partial\Delta t^*/\partial\beta < 0$. □

II.3.3 Example

Suppose effort costs are quadratic, $c(e) = \alpha e^2$, where $\alpha > 0$, and effort is measured in units of success probability, such that $p(e) = e$ for all $e \leq 1$ and $p(e) = 1$, otherwise. Let $\Delta R + \Delta V < 2\alpha$ to guarantee an interior solution of e^*. Then $t_1^* = \underline{t}$ and

$$\Delta t^* = \begin{cases} \frac{\Delta R}{2} - \frac{\beta}{1-\beta} \times \frac{\Delta R + \Delta V}{2} & \text{for } \beta < \frac{\Delta R}{2\Delta R + \Delta V}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Thus, if the agent is sufficiently socially attentive, $\beta \geq \frac{\Delta R}{2\Delta R + \Delta V}$, the principal sets no monetary incentives. If, for instance, $\Delta R = \Delta V$, the threshold is $1/3$. In general, the threshold is between 0 and $1/2$. Figure 1 illustrates the example.

II.4 Adverse selection

In this section, we analyze the principal's problem when she does not know the agent's social preferences. We thus let the agent's type β be the agent's private information.

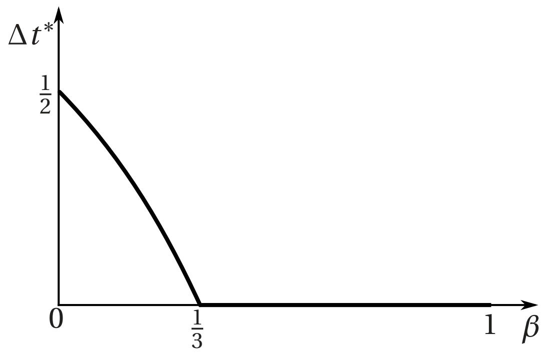

Figure 1: The optimal incentives Δt^* if $R_1 = V_1 = 1$ and $R_2 = V_2 = \alpha = 2$.

This generates a combined problem of moral hazard and adverse selection.[11] To formalize this, we let the agent's type β be drawn at an initial stage of the game from the cumulative distribution function F, where $F : \left[\underline{\beta}, \bar{\beta}\right] \to [0,1]$ with $0 \leq \underline{\beta} < \bar{\beta} \leq 1$ and the corresponding probability density function f. It is noteworthy that, with adverse selection, the principal can in general not implement the contract that is optimal without adverse selection (i.e., when she knows the agent's type).[12]

II.4.1 Optimal contract

The following result establishes that screening for the different types of agents is never optimal for the principal. The principal thus optimally designs a single contract, i.e., a pooling contract. This holds no matter how rich the set of types is and what shape the distribution function has. The idea of the proof is that any menu of contracts can be replaced by a single contract that makes the principal better off.[13]

Proposition 4: *Suppose the agent's type β is private information, i.e., not known to the principal. The principal then optimally offers a pooling contract.*

PROOF: See Appendix A.

This result is intuitive. If the principal would offer separate contracts for different types of agents, she would have to reward some types for the unsuccessful outcome. Simultaneously, to guarantee incentive compatibility, these types must receive relatively low rewards in case of success. Therefore, the incentives provided to these types are unnecessarily weak and the principal can improve by eliminating these contracts.

[11] For an introduction to these mixed agency models, see Laffont and Martimort (2001, Chapter 7).

[12] To see this, suppose the agent could be of two or more types, for which different contracts are optimal without adverse selection. The menu of contracts consisting of the contracts the principal would offer without adverse selection is not incentive-compatible, since $t_1^* = \underline{t}$ holds for all contracts and all types would select the contract with the highest payment t_2 from the menu.

[13] In Appendix A, we present a proof for a menu of binary contracts. We further demonstrate in Appendix B that it is not beneficial for the principal to construct a menu with more complicated contracts.

II.4.2 Design of the optimal pooling contract

The principal's problem when designing the optimal pooling contract (t_2^*, t_1^*) is to

$$\begin{aligned}\max_{t_1,t_2} E[u_P] &= \int_{\underline{\beta}}^{\bar{\beta}} E[u_P|\beta] f(\beta) d\beta \qquad \text{(P2)}\\ &= \int_{\underline{\beta}}^{\bar{\beta}} \left[p(e(\Delta t, \beta))(R_2 - t_2) + (1 - p(e(\Delta t, \beta)))(R_1 - t_1) \right] f(\beta) d\beta \\ &\text{subject to } \; t_1, t_2 \geq \underline{t}.\end{aligned}$$

Since setting negative incentives is never optimal (by essentially the same arguments as in the case of the principal knowing the agent's type), we have that $t_1^* = \underline{t}$, which allows us to write the principals problem as

$$\begin{aligned}\max_{\Delta t} E[u_P] &= \int_{\underline{\beta}}^{\bar{\beta}} \left[p(e(\Delta t, \beta))(R_2 - \Delta t - \underline{t}) + (1 - p(e(\Delta t, \beta)))(R_1 - \underline{t}) \right] f(\beta) d\beta \qquad \text{(P2')}\\ &\text{subject to } \; \Delta t \geq 0.\end{aligned}$$

Thus, in case the principal does not know the agent's type, she optimally sets the pooling contract $\left(t_1^* = \underline{t}, t_2^* = \underline{t} + \Delta t^*\right)$, where Δt^* solves (P2'). Note that, since it is never optimal to set incentives Δt^* that fall short of zero or exceed the return spread ΔR and the principal's objective function is continuous, an optimal contract always exists.

It is instructive to investigate the derivative of the principal's expected utility with respect to the incentives:

$$\begin{aligned}\frac{\partial E[u_P]}{\partial \Delta t} &= \int_{\underline{\beta}}^{\bar{\beta}} \frac{\partial E[u_P|\beta]}{\partial \Delta t} f(\beta) d\beta \qquad (20)\\ &= \int_{\underline{\beta}}^{\bar{\beta}} \left[p'(e(\Delta t, \beta)) \frac{\partial e(\Delta t, \beta)}{\partial \Delta t} (\Delta R - \Delta t) - p(e(\Delta t, \beta)) \right] f(\beta) d\beta.\end{aligned}$$

Hence, in order to determine the optimal incentive power, the principal takes the average effect of the incentives on her conditional expected utility into account. Due to this averaging effect, the following two results are immediate. First, if the agent is likely to have a rather low level of social attentiveness β, the principal optimally sets monetary incentives, $\Delta t^* > 0$, while she optimally sets no monetary incentives, $\Delta t^* = 0$, if the agent is likely to have a rather high level of social attentiveness.[14] Second, if the principal optimally provides incentives, the incentives are relatively weak. To be precise,

[14] The result follows directly from (20), since we know from Proposition 2 and its proof that the derivative of the conditional expected utility $E[u_P|\beta]$ with respect to the incentives Δt evaluated at $\Delta t = 0$ is positive when the agent's social attentiveness is low, while the derivative is negative for high levels of social attentiveness. Formally, at least for all β in the nonempty interval $(\max\{\dot{\beta}, \hat{\beta}\}, 1)$ and all $\Delta t \in [\underline{t}, \underline{t} + \Delta R)$, the derivative $\partial E[u_P|\beta]/\partial \Delta t$ is negative.

given the concavity of the principal's problem, the derivative $\partial E[u_P|\beta]/\partial\Delta t$ is decreasing in β and Δt, which together with (20) implies that the principal will set strictly lower incentives if the probability that the agent is egoistic is below one, compared to when the agent is egoistic for sure: $\Delta t^*|_{F(\beta=0)<1} < \Delta t^*|_{F(\beta=0)=1}$. More generally, for any pair of distribution functions F and $\tilde{F}$, where $\tilde{F}$ first-order stochastically dominates F, so that higher levels of social attentiveness are more likely under $\tilde{F}$ than under F,

$$\Delta t^*|_{\tilde{F}} \begin{cases} < \Delta t^*|_F & \text{if } \Delta t^*|_F > 0, \\ = \Delta t^*|_F & \text{if } \Delta t^*|_F = 0. \end{cases} \tag{21}$$

To summarize, adverse selection causes the principal to design a pooling contract where all types of agents experience either no or only relatively weak incentives and incentives are weaker if higher levels of social attentiveness are more likely.

II.4.3 Example

We return to our example with quadratic effort costs and effort measured in units of success probability. For concreteness, we further let the agent's type be uniformly distributed, $\beta \sim u[\underline{\beta},\bar{\beta}]$. The optimal incentives are then

$$\Delta t^* = \begin{cases} \frac{1-\bar{\beta}-\underline{\beta}}{2-\bar{\beta}-\underline{\beta}}\Delta R - \frac{\bar{\beta}+\underline{\beta}}{2(2-\bar{\beta}-\underline{\beta})}\Delta V & \text{for } \bar{\beta}+\underline{\beta} < \frac{2\Delta R}{2\Delta R+\Delta V}, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We directly see that in case of positive incentives, the power of incentives is decreasing in $\underline{\beta}$ as well as in $\bar{\beta}$. The maximal incentives are $\lim_{\underline{\beta},\bar{\beta}\to 0}\Delta t^* = \Delta R/2$, which is the same value as when the principal knows that the agent is egoistic for sure. Except for this limit case, the principal sets weaker incentives if she does not know the agent's type than when she knows that the agent is egoistic. For instance, if $\Delta R = \Delta V = 1$, the optimal incentives in case the principal knows that her agent is egoistic are $\Delta t^* = 1/2$, whereas they are $\Delta t^* = 0$ in case the principal does not know the agent's type and $\underline{\beta}+\bar{\beta} \geq 2/3$, which holds, for instance, if β is uniformly distributed in the unit interval. Interestingly, setting no incentives, $\Delta t^* = 0$, is also optimal if there is a one-in-three chance that the agent is egoistic, as estimated by Engel (2011), while there is a two-in-three chance that the agent is socially attentive and the attentiveness parameter β is uniformly distributed in the unit interval.

II.5 Conclusions

In standard agency models, the agent does not care how his decisions affect others. We relax this assumption and obtain two main results.

First, with socially attentive preferences, the principal optimally sets either no incentives or incentives that are rather weak. This finding is empirically relevant, since it provides an explanation for the puzzle why in practice many employees experience no financial incentives or rather weak incentives.

Second, the principal optimally provides a pooling contract in case she does not know the agent's social attentiveness. The principal thus does not screen via a menu of contracts, no matter how rich the set of possible attentiveness levels is and what shape the underlying distribution function has. The result is practically relevant, since it provides a theoretical foundation for the uniformity of observed contracts. We further show that the pooling contract provides either no or relatively weak incentives, compared to the benchmark with a surely egoistic agent.

Relaxing the assumption that agents do not care at all how their decisions influence others thus makes the theoretical model not only more realistic, but also generates predictions that fit the empirical findings.

III Contractible and Non-contractible Efforts with Socially Attentive Preferences

Joint work with Philipp Weinschenk

We follow Koch and Weinschenk (2021) in investigating the equilibrium utilities and efficiency when the effort of agents with social preferences is non-contractible. We compare these results with the case where effort is contractible. We show that with socially attentive preferences contractibility of effort does not generally cause the implementation of the efficient effort and may harm the generated surplus. This provides an efficiency argument for regulatory boundaries on the content of employment contracts and employers' control.

JEL Classification: D82, D91, M52.

Keywords: agency model, socially attentive preferences, incentives, contractibility.

III.1 Introduction

Some employers overstep regulatory boundaries and legal restrictions in order to control and surveil their employees' work (Brächer, 2021). Employers are likely keen to control their workers, because they expect these measures to increase the performance levels and output. There are indications that during the corona pandemic firms have extended the practice of surveilling their employees. Contributing factors here are the possibilities offered by monitoring work carried out in home office through applications that measure actual behavior of workers (Walker, 2021). Advice by standard agency literature could be interpreted as endorsing this overreaching behavior of employers, since one way of modeling control over an egoistic agent through the principal is making the agent's effort contractible.[15] In the standard literature, this leads to the efficient effort implementation and therefore an increased surplus, since it eliminates the inefficiency due to the egoistic agent's information rent. In Koch and Weinschenk (2021) we introduced the possibility of a of socially attentive agent, who cares about

[15] We follow the standard convention in the agency literature and talk about a male agent and a female principal.

how his actions affect the well-being of others.[16] We use this idea and extend the concept to investigate whether contractibility is still increasing the efficiency of the work relation between principal and agent. We can demonstrate that contractibility is not always efficiency-enhancing when the agent has socially attentive preferences.

In Koch and Weinschenk (2021), we study a model where an agent acts on behalf of a principal and the agent's non-contractible effort choice influences the probability distribution over outcomes. The outcome affects the principal, the agent, and possibly a third party. We augment the agent's preferences by allowing him to put weight on the utilities of others. We say that the agent is egoistic if he puts zero weight on others' utilities, while the agent is socially attentive if he puts a positive weight. One of the major results in Koch and Weinschenk (2021) is that, despite monetary incentives being effective when the agent is egoistic as well as when he is socially attentive, it might be optimal for the principal to refuse to provide monetary incentives if the agent is sufficiently socially attentive. The intuition for this result is that providing monetary incentives is costly for the principal. Since a socially attentive agent reacts less strongly to monetary incentives than an egoistic one, the principal's costs of providing monetary incentives may – in comparison to the benefit of monetary incentives, in the form of a more motivated agent – be too high to justify monetary incentives.

In order to be able to fully compare the cases of non-contractibility and contractibility with agents with socially attentive preferences, we first deepen the analysis of Koch and Weinschenk (2021) by investigating the effects of social preferences for the parties' utilities in equilibrium and the efficiency of the relationship with non-contractibility. We compare this outcome with the one a social planner would implement. In particular, we can show that the principal is better off if her agent is socially attentive rather than egoistic. Intuitively, despite the fact that monetary incentives are less effective in case the agent is socially attentive, this negative effect is overcompensated by the positive effect that for all potentially optimal contracts a socially attentive agent is more motivated to exert effort. Second, the principal implements an inefficiently low effort level irrespectively of whether the agent is egoistic or socially attentive. However, the implemented effort level approaches the efficient (i.e., surplus-maximizing) level as the agent's social attentiveness approaches its maximum.

We next analyze our model of socially attentive preferences when effort is contractible. For the benchmark of an egoistic agent and the absence of a third party – a scenario which is extensively studied in the existing literature – we obtain the stan-

[16] There is rich empirical evidence that many people take into account how their actions influence the well being of others. For instance, experiments on the dictator game – where one person determines how to share a certain endowment between themselves and a second person – show that a majority of persons do not behave selfishly. Andreoni and Miller (2002), for example, document that only a quarter of persons are selfish money-maximizers. In standard agency models, it is nonetheless assumed that agents do not care how their decisions influence others. This is a strong assumption, which we relax in Koch and Weinschenk (2021).

dard result that contractibility of effort leads to the implementation of the efficient effort level. Remarkably, this is, in general, no longer true if the agent is socially attentive or a third party is present. We further show that contractibility is not necessarily efficiency-enhancing. To be precise, the generated surplus could be lower when effort is contractible rather than non-contractible. The reason is that, while the principal implements an inefficiently low effort level when effort is non-contractible, she may implement an inefficiently high effort level when it is contractible. If the agent is sufficiently socially attentive, the over-implementation problem caused by contractibility is more severe than the under-implementation problem caused by non-contractibility. An important implication of this result is that regulatory boundaries on the methods employers can use to control their employees and legal restrictions on what can be specified in employment contracts could not only help employees, but also enhance efficiency.

One may presume that social attentiveness positively affects the generated surplus, since parties should then interact and cooperate in a more social way. Although this presumption is correct in case effort is non-contractible, it is wrong in case effort is contractible. Here, the generated surplus may be lower if the agent is socially attentive instead of egoistic. This is the case because the principal exploits the agent by demanding an excessively high effort level when effort is contractible and the agent's social attentiveness is high.

Relation to the literature. Since this paper builds on Koch and Weinschenk (2021), it is necessarily related to the literature that incorporates different forms of social preferences into agency models cited there. In addition, there has only little work been done with respect to the effects of contractibility of effort or the surveillance of agents with social preferences. To the best of our knowledge, only one other theoretical contribution is concerned with the contractibility of effort, if the agent has social preferences.[17] Moreover, there are only two experimental studies which focus on whether or not a principal will exploit her agent's social preferences, if she has information about his level of social preferences. On the theoretical side, Cassar and Armouti-Hansen (2020) analyze how firms can attract and screen potential future personnel by choosing a mission that is either aligned with the agent's preferences or with the principal's preferences. Cassar and Armouti-Hansen (2020) investigate what influence the contractual environment has on the principal's choice of mission. They compare four contractual environments by combining two two-dimensional characteristics. The agent's effort is either contractible or not and the agents level of intrinsic motivation is either symmetric information of both parties or asymmetric information, where only the agent knows his level of social attentiveness. They find that contractibility aligns the mission

[17] The literature on exploitative contracts highlights this issue as well. However, it focuses on (naive) agents making mistakes and the principal knowing about the propensity for mistakes of the agents. The modeling relation is therefore tenuous, cf. Köszegi (2014).

firmly with the companies goals of revenue-maximization. The mission does not need to serve as an incentive in these circumstances and does not need to be close to the preferred mission of the agent. The mission will also be closer to the principal's optimal setting if the agent's intrinsic motivation is not known to the principal. In this case, however, the principal makes the contract designed for agents with lower intrinsic motivation less attractive for agents with high intrinsic motivation, which reduces agents' information rent. Moreover, two experimental studies have highlighted, that a principal will exploit an agent with social preferences. Bigoni et al. (2021) find that principals use the information of agents' social preferences to implement contracts that are optimal for the firms. Vu (2019) confirms that employers will tailor the contracts to their agent's social preferences and to their own advantage, if they have information about the agent's social preferences. We enrich the theoretical landscape by providing a general but tractable model for the effects of contractibility of effort with socially attentive agents.

III.2 Benchmark moral hazard model

In order to allow the reader unfamiliar with Koch and Weinschenk (2021) to follow, we briefly recapitulate the necessary assumptions and results for our purpose in a benchmark moral hazard model and extend its analysis to include the parties' expected utilities and the efficiency of the relationship between an agent with socially attentive preferences and a principal.

Primitives. A principal needs to hire an agent. When working for the principal, the agent exerts effort $e \in [0, e^{\max}]$, where $e^{\max}$ is positive and could be finite or infinite. The agent's effort choice determines the probability distribution over outcomes. With probability $p(e)$ the agent yields a successful outcome that is associated with a high return $R > 0$ for the principal, while with probability $1 - p(e)$ the outcome is unsuccessful which causes a low return normalized to zero.[18] The outcome may also affect a third party, which experiences a payoff $V \geq 0$ if the agent succeeds and 0 otherwise. The case $V = 0$ captures the situation where a third party is absent or unaffected. In the moral hazard case the outcome is contractible but the agent's effort is not and a contract is a pair of payments $(w_0, w_R) \in \mathbb{R}^2$, w_0 in the unsuccessful case and w_R in case of success. The agent's wealth is normalized to zero, such that payments cannot be negative: $w_0, w_R \geq 0$.

Preferences. The principal and the agent are risk neutral. In Koch and Weinschenk (2021) we deviate from the textbook moral-hazard model by supposing that the agent might have socially attentive preferences. Formally, the agent puts a weight $\beta \in [0, 1]$

[18] We show in Koch and Weinschenk (2021) that one can focus on the two outcome case without loss of generality.

on the utilities of others.[19] We characterize the agent as egoistic if he puts zero weight on others' utilities, $\beta = 0$, while he is socially attentive if he puts positive weight, $\beta > 0$. Unless explicitly stated differently, we suppose that the agent puts at least slightly more weight on his own utility than on that of others, $\beta < 1$. The agent's expected utility is

$$E[u_A] = p(e)\,w_R + (1 - p(e))\,w_0 - c(e) + \beta E[u_{\neg A}], \tag{23}$$

where $p(e)\,w_R + (1 - p(e))\,w_0$ is the expected wage payment, $c(e)$ are the agent's effort costs, and $\beta E[u_{\neg A}]$ is the weighted sum of the other parties' utilities. The principal's expected utility equals the expected difference between the return she earns and the payment she makes to the agent:

$$E[u_P] = p(e)(R - w_R) + (1 - p(e))(0 - w_0). \tag{24}$$

The third party's expected utility is

$$E[u_T] = p(e)V. \tag{25}$$

By including (24) and (25) into (23), we can rewrite the agent's expected utility as

$$E[u_A] = p(e)\,w_R + (1 - p(e))\,w_0 - c(e) + \beta\big(p(e)(R - w_R + V) - (1 - p(e))\,w_0\big). \tag{26}$$

Assumptions. We impose the standard assumptions on the effort cost function c and the success function p: c and p are twice continuously differentiable, $c'(e)$, $c''(e) > 0$ for $e > 0$, $c(0) = c'(0) = 0$, $\lim_{e \to e^{\max}} c'(e) = \infty$, $p(0) = 0$, $p'(e) > 0$, and $p''(e) \leq 0$. Thus, the effort cost function is increasing and convex and the success function is increasing and weakly concave.

Timing. First, the principal offers a contract to the agent, who then decides whether to accept or reject it. If the agent rejects, all parties receive reservation utilities of zero and the game ends. In case of acceptance, the agent exerts effort. Finally, the outcome is realized and the agent receives the contracted payment.

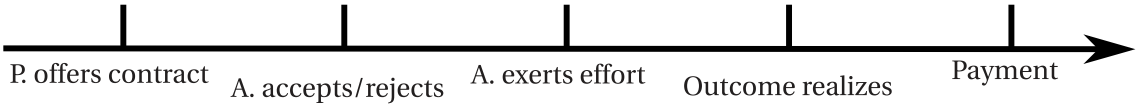

Figure 2: Timeline.

[19] As we discuss in Appendix D, allowing the agent to put different weights on the principal's and the third party's utilities complicates notation without providing significant new insights.

III.3 Analysis of the benchmark moral hazard model

This section provides a quick summary of the necessary results from Koch and Weinschenk (2021) for our investigation here. Differentiating the agent's expected utility with respect to effort yields the agent's effort choice for any given contract. i.e.

$$\frac{\partial E[u_A]}{\partial e} = p'(e)\big((w_R - w_0)(1-\beta) + \beta(R+V)\big) - c'(e). \tag{27}$$

We have to distinguish between two cases. First, if $(w_R - w_0)(1-\beta) + \beta(R+V) \leq 0$, the derivative $\partial E[u_A]/\partial e$ is non-positive for effort $e = 0$ and negative for all efforts $e > 0$. Accordingly, the agent optimally chooses to invest zero effort, $e^* = 0$. Second, if $(w_R - w_0)(1-\beta) + \beta(R+V) > 0$, the derivative $\partial E[u_A]/\partial e$ is positive for $e = 0$, such that the agent optimally chooses a positive effort level $e^* > 0$, where e^* solves the first-order condition $\partial E[u_A]/\partial e = 0$.[20]

The principal's problem consists of maximizing her expected utility subject to the agent's incentive constraint, the agent's participation constraint, and the limited liability constraints:

$$\max_{w_0, w_R, \hat{e}} E[u_P] \text{ subject to } \hat{e} \in \operatorname*{argmax}_e E[u_A],\ E[u_A] \geq 0,\ w_0 \geq 0, \text{ and } w_R \geq 0, \tag{P1}$$

where $\hat{e}$ denotes the effort level the principal seeks to implement. We can neglect the participation constraint in the principal's problem, since – as usual in such agency models – the agent's participation constraint $E[u_A] \geq 0$ is automatically satisfied for every contract (w_0, w_R) that satisfies the limited liability constraints.[21] We proceed in Koch and Weinschenk (2021) by proving the existence of an optimal contract. We restate the result here.

Proposition 5: *There always exists an optimal contract* $\left(w_0^*, w_R^*\right)$. *The optimal contract satisfies* $w_0^* = 0$ *and* $w_R^* \in [0, R)$.

Proposition 5 shows that an optimal contract always exists, whether the agent is egoistic or socially attentive. The intuition for the properties of the optimal contract are as follows. First, in case the agent fails, it is optimal to make only the minimal possible wage payment; thus $w_0^* = 0$.[22] Second, it is never optimal for the principal to

[20] Note that, due to $p'' \leq 0$ and $c'' > 0$, the second-order condition is satisfied: $\partial E[u_A]^2/\partial e^2 < 0$. Furthermore, because $\lim_{e \to e^{\max}} c'(e) = \infty$, the optimal effort e^* is lower than the maximal effort $e^{\max}$, such that we have an interior solution.

[21] To see this, note that the agent's expected utility is non-negative if $w_0, w_R \geq 0$ and he chooses to exert zero effort, see equation (26). Since the agent chooses the effort level to maximize his expected utility, his equilibrium expected utility must be non-negative as well. Therefore, the agent accepts any contract that satisfies the limited liability constraints.

[22] Having a payment of zero should be interpreted as the principal offering a payment equal to the minimal possible payment (which we normalized to zero, but which could be different in general).

provide a wage payment that equals or exceeds her return R and it is not possible to offer a negative wage payment; thus $w_R^* \in [0, R)$.

A useful implication of Proposition 5 is that, for all potentially optimal contracts, the agent's effort choice e^* is determined by the first-order condition[23]

$$\frac{\partial E[u_A]}{\partial e} = p'(e)\left(w_R(1-\beta) + \beta(R+V)\right) - c'(e) = 0. \tag{28}$$

Implicitly differentiating equation (28) yields that

$$\frac{\partial e^*}{\partial w_R} = -\frac{p'(e^*)(1-\beta)}{p''(e^*)\left(w_R(1-\beta)+\beta(R+V)\right) - c''(e^*)}, \tag{29}$$

which is positive. Thus, with an egoistic agent ($\beta = 0$), as well as with a socially attentive agent ($\beta > 0$), monetary incentives are effective, in the sense that the agent exerts more effort, the higher the monetary incentives are.

We further obtain that

$$\frac{\partial e^*}{\partial \beta} = -\frac{p'(e^*)(-w_R + R + V)}{p''(e^*)\left(w_R(1-\beta)+\beta(R+V)\right) - c''(e^*)}, \tag{30}$$

which is positive for all $w_R \in [0, R)$. Hence, for all potentially optimal contracts, the agent exerts more effort, the more socially attentive he is.

We also get that

$$\frac{\partial e^*}{\partial V} = \frac{\partial e^*}{\partial R} = -\frac{p'(e^*)\beta}{p''(e^*)\left(w_R(1-\beta)+\beta(R+V)\right) - c''(e^*)}, \tag{31}$$

which is positive if $\beta > 0$ and zero if $\beta = 0$. Therefore, while a socially attentive agent reacts positively towards a higher payoff of the third party or a higher return of the principal, an egoistic agent does not react. These comparative statics are summarized in Proposition 6

Proposition 6: *For all potentially optimal contracts, i.e., all contracts* (w_0, w_R) *satisfying* $w_0 = 0$ *and* $w_R \in [0, R)$*, the following holds:*

- *With an egoistic agent ($\beta = 0$), as well as with a socially attentive agent ($\beta > 0$), monetary incentives are effective, in the sense that the agent exerts more effort, the higher the monetary incentives provided by the principal are, $\partial e^*/\partial w_R > 0$.*
- *The agent exerts more effort, the more socially attentive he is, $\partial e^*/\partial \beta > 0$.*
- *While a socially attentive agent reacts positively towards a higher payoff of the third party, $\partial e^*/\partial V > 0$, an egoistic agent does not react, $\partial e^*/\partial V = 0$. The same holds true with respect to the principal's return R.*

[23] To derive (28), we used that the first-order condition not only applies if $(w_R - w_0)(1-\beta)+\beta(R+V) > 0$, but also if $(w_R - w_0)(1-\beta) + \beta(R+V) = 0$.

A last result from Koch and Weinschenk (2021) we need for the analysis below highlights that the principal does not set monetary incentives if the agent's social attentiveness is sufficiently high.[24] The intuition why the optimally designed contract may include no monetary incentives is as follows. First, socially attentive preferences limit the effectiveness of monetary incentives, in the sense that a socially attentive agent reacts less strongly to monetary incentives than an egoistic agent does, as can be seen in (29). Second, it is costly for the principal to provide monetary incentives, see (24). Consequently, the principal has to weigh off the costs of providing monetary incentives with the benefit of a more motivated agent. In case of a sufficiently socially attentive agent, the costs are to high in comparison to their positive effect on the agent's effort and the principal therefore optimally sets a constant, outcome-independent wage.

Proposition 7: *If the agent is egoistic ($\beta = 0$) or his social attentiveness is sufficiently low (β is small), the principal optimally sets monetary incentives, $w_R^* > w_0^*$. In contrast, the principal sets no monetary incentives, $w_R^* = w_0^*$, if the agent's social attentiveness β is sufficiently high.*

III.3.1 Equilibrium utilities

We are now in a position to further extend the analysis of Koch and Weinschenk (2021) by determining the parties' utilities for the optimal contract. We first show that the agent as well as the principal experience a positive rent, i.e., expected utilities that exceed their reservation utilities. This result holds independently of whether the agent is egoistic or socially attentive.

Proposition 8: *Under the optimal contract $\left(w_0^*, w_R^*\right)$, the agent as well as the principal experience a positive rent: $E\left[u_A | \left(w_0^*, w_R^*\right)\right],\ E\left[u_P | \left(w_0^*, w_R^*\right)\right] > 0$.*

Proof: Consider first the agent's expected utility. We know from before that $w_0^* = 0$, cf. Proposition 5, and that the agent optimally chooses the effort e^* that solves the first-order condition, cf. (28),

$$p'(e)\left(w_R^*(1-\beta) + \beta(R+V)\right) - c'(e) = 0. \tag{32}$$

By Proposition 7, either $w_R^* > w_0^*$ or $\beta > 0$ or both inequalities hold, such that e^* is positive. If the agent alternatively chose effort $e = 0$, his expected utility would be zero. Because e^* is the unique maximizer of the agent's expected utility and positive, the agent's expected utility must be positive when exerting the optimal effort e^*:

$$E\left[u_A | \left(w_0^*, w_R^*\right)\right] > 0. \tag{33}$$

[24] This is no limit result. The principal may optimally set no monetary incentives even if the agent puts a non-trivially higher weight on his own utility than on that of others. Cf. the proof in Koch and Weinschenk (2021, p. 596). It is further illustrated in the quadratic cost example considered in subsection III.3.4. There, the principal optimally sets no monetary incentives if the agent's social attentiveness β is at least as large as a threshold, which is – depending on the parameter constellation – at most 1/2.

Consider next the principal's expected utility. For the contract $(w_0 = 0, w_R = 0)$, the principal's expected utility is, by (24),

$$E[u_P|(w_0 = 0, w_R = 0)] = p(e(0,\beta))R. \tag{34}$$

First, if the agent is socially attentive, $\beta > 0$, he chooses a positive effort level when faced with the contract $(w_0 = 0, w_R = 0)$, such that the probability of a successful outcome is positive, $p(e(\cdot)) > 0$, and the principal experiences a positive expected utility

$$E\left[u_P|(w_0 = 0, w_R = 0), \beta > 0\right] > 0. \tag{35}$$

Thus, the principal's expected utility under the optimal contract $\left(w_0^*, w_R^*\right)$ must be positive as well:

$$E\left[u_P|\left(w_0^*, w_R^*\right), \beta > 0\right] > 0. \tag{36}$$

Second, if the agent is egoistic, $\beta = 0$, he chooses zero effort when faced with the contract $(w_0 = 0, w_R = 0)$, such that the principal experiences an expected utility of zero. However, in case $\beta = 0$, we know from Proposition 7 that it is strictly optimal for the principal to set monetary incentives, i.e., $w_R^* > w_0^*$. Accordingly, the principal's expected utility under the optimal contract $\left(w_0^*, w_R^*\right)$ must be positive in this case also:

$$E\left[u_P|\left(w_0^*, w_R^*\right), \beta = 0\right] > 0. \qquad \square$$

Intuitively, due to the non-contractibility of the agent's effort choice, the agent experiences an information rent which the principal cannot fully extract. However, the principal is still able to construct a contract that yields her a positive expected utility.

The next proposition shows that the principal benefits from a socially attentive agent.

Proposition 9: *The principal's expected utility is increasing in the social attentiveness of her agent,* $dE\left[u_P|\left(w_0^*, w_R^*\right)\right]/d\beta > 0$. *In particular, the principal's expected utility is higher if her agent is socially attentive rather than egoistic,* $E\left[u_P|\left(w_0^*, w_R^*\right), \beta > 0\right] > E\left[u_P|\left(w_0^*, w_R^*\right), \beta = 0\right]$.

Proof: By the Envelope Theorem and because for the optimal contract $w_0^* = 0$ and $w_R^* \in [0, R)$,

$$\frac{dE\left[u_P|\left(w_0^*, w_R^*\right)\right]}{d\beta} = p'(e(\cdot))(R - w_R^*) \times \frac{\partial e(\cdot)}{\partial \beta} > 0. \tag{37}$$

This particularly implies that

$$E\left[u_P|\left(w_0^*, w_R^*\right), \beta > 0\right] > E\left[u_P|\left(w_0^*, w_R^*\right), \beta = 0\right]. \qquad \square$$

This is intuitive: Despite the fact that monetary incentives are less effective for a socially more attentive agent, this negative effect is overcompensated by the positive effect that for all potentially optimal contracts a socially more attentive agent is more motivated to exert effort. The result is also practically relevant: if the principal has the choice between two agents that differ with respect to their social attentiveness, but are otherwise identical, the principal should seek to hire the more socially attentive agent.

III.3.2 Efficiency

We next compare the effort and surplus implemented by the principal to the efficient levels. For this purpose, consider a social planner, who seeks to maximize the expected surplus

$$E[s] = p(e)(R+V) - c(e). \tag{38}$$

We directly see that, while the effort is important for the surplus, the wage payments per se do not matter – they are simply transfers between the parties and cancel out. Maximizing over effort yields the first-order condition

$$\frac{\partial E[s]}{\partial e} = p'(e)(R+V) - c'(e) = 0. \tag{39}$$

Because the expected surplus is concave in effort, the effort level the planner seeks to implement solves (39). We refer to this effort level as the efficient effort level $e^{\text{efficient}}$. By formula (27), the planner sets monetary incentives of $w_R - w_0 = R + V$ to implement $e^{\text{efficient}}$. Since the principal only sets incentives of $w_R^* - w_0^* \in [0, R)$, cf. Proposition 5, we obtain the following result.

Proposition 10: *The principal implements an inefficiently low effort level:* $e^* < e^{\textit{efficient}}$.

Proposition 10 shows that the classical result, that a principal underprovides incentives to a wealth-constrained agent and thereby implements an inefficiently low effort level (cf. Laffont and Martimort 2001), also holds when the agent has socially attentive preferences. In our model, this result is due to two effects. First, the classical rent-extraction efficiency trade-off. Formally, the principal optimally provides incentives $w_R < R$ instead of $w_R = R$ in order to reduce the information rent she has to pay to her agent. Second, if $V > 0$, the principal also underprovides incentives because she ignores the positive effect of stronger incentives on the third party.

The following proposition shows, however, that the under-implementation of effort is a negligible issue if the agent's social attentiveness is large.

Proposition 11: *The effort level implemented by the principal approaches the efficient effort level as the agent's social attentiveness goes to* 1*:* $\lim_{\beta \to 1} e^* = e^{\textit{efficient}}$.

PROOF: By Propositions 5 and 7, the principal sets $w_0^* = w_R^* = 0$ if β is sufficiently close to 1. For this contract, the agent chooses the effort level e^* that solves

$$p'(e)\beta(R+V) - c'(e) = 0, \tag{40}$$

where we used the first-order condition (28). Comparing this equation to (39) yields that e^* approaches $e^{\text{efficient}}$ as β goes to 1. □

Since the expected surplus is concave in the implemented effort, Propositions 10 and 11 directly imply the following results.

Corollary 1: *The principal implements an effort level that causes an inefficiently low expected surplus:* $E[s|e^*] < E\left[s|e^{\textit{efficient}}\right]$. *Furthermore,* $\lim_{\beta\to 1} E[s|e^*] = E\left[s|e^{\textit{efficient}}\right]$.

III.3.3 Additional results with concavity

We can derive additional results if $c''' \geq 0$ and $p''' \leq 0$, such that the principal's problem is concave, $\partial^2 E[u_P]/\partial w_R^2 < 0$.[25] The first part of Proposition 12 shows that there is a monotone negative relationship between the power of monetary incentives $w_R^* - w_0^*$ and the agent's social attentiveness β. The second part of Proposition 12 shows that in the presence of a third party, $V > 0$, the equilibrium effort e^* is strictly increasing in the agent's social attentiveness β, but that the relationship is only weak in the absence of a third party, $V = 0$.

Proposition 12: *Let* $c''' \geq 0$ *and* $p''' \leq 0$ *such that the principal's problem is concave.*

- *If* $w_R^* > w_0^*$, *the power of monetary incentives is decreasing in the agent's social attentiveness* β: $\partial(w_R^* - w_0^*)/\partial\beta < 0$.
- *If* $V > 0$, *the implemented effort level* e^* *is increasing in* β: $de^*/d\beta > 0$. *If* $V = 0$, e^* *is constant in* β, *as long as* β *is sufficiently small such that* $w_R^* > w_0^*$, *and increasing in* β *otherwise.*

PROOF: Using (29) and the principal's expected utility with respect to w_R, i.e.

$$\frac{\partial E[u_P]}{\partial w_R} = p'(e(w_R,\beta))\frac{\partial e(w_R,\beta)}{\partial w_R}(R - w_R) - p(e(w_R,\beta)), \tag{41}$$

we directly see that when $c''' \geq 0$ and $p''' \leq 0$,

$$\frac{\partial E[u_P]}{\partial w_R} = \frac{\left(p'(\cdot)\right)^2(1-\beta)(R-w_R)}{c''(\cdot) - p''(\cdot)\left(w_R(1-\beta) + \beta(R+V)\right)} - p(\cdot) \tag{42}$$

is decreasing in w_R and β. Suppose that $w_R^* > w_0^*$. Then w_R^* solves the first-order condition $\partial E[u_P]/\partial w_R = 0$. If β increases, w_R^* must thus be lowered to restore the first-order condition, so $\partial w_R^*/\partial\beta < 0$. Since by Proposition 5 w_0^* is independent of β,

$$\frac{\partial(w_R^* - w_0^*)}{\partial\beta} = \frac{\partial w_R^*}{\partial\beta} < 0, \tag{43}$$

which proves the first part.

[25] Note that irrespectively of whether the agent is egoistic or socially attentive, $c''' \geq 0$ and $p''' \leq 0$ are sufficient conditions for concavity.

Regarding the relationship between e^* and β, we have to distinguish between two cases. First, consider the case where β is sufficiently large such that

$$\left.\frac{\partial E[u_P]}{\partial w_R}\right|_{w_R=0} \leq 0. \tag{44}$$

Because the principal's problem is concave – i.e., $\partial E[u_P]/\partial w_R$ is decreasing in w_R – the principal then optimally sets $w_R^* = 0$. If we increase β further, inequality (44) remains valid, so that the principal optimally keeps $w_R^* = 0$. This implies that

$$\frac{de^*}{d\beta} = \frac{\partial e^*}{\partial \beta} + \frac{\partial e^*}{\partial w_R^*} \times \frac{\partial w_R^*}{\partial \beta} = \frac{\partial e^*}{\partial \beta}, \tag{45}$$

which is positive by (30).

Second, consider the case where β is sufficiently small, such that

$$\left.\frac{\partial E[u_P]}{\partial w_R}\right|_{w_R=0} > 0. \tag{46}$$

The principal then optimally sets $w_R^* > 0$, where w_R^* solves the first-order condition $\partial E[u_P]/\partial w_R = 0$. Implicit differentiation of $\partial E[u_P]/\partial w_R = 0$, cf. (42), yields

$$\frac{\partial w_R^*}{\partial \beta} = -\frac{R - w_R}{1-\beta} + \frac{p(\cdot)p''(\cdot)V}{(p'(\cdot))^2(1-\beta) - p(\cdot)p''(\cdot)(1-\beta)}. \tag{47}$$

Together with (29) and (30) we get that

$$\frac{de^*}{d\beta} = \frac{\partial e^*}{\partial \beta} + \frac{\partial e^*}{\partial w_R^*} \times \frac{\partial w_R^*}{\partial \beta} \overset{\text{sign}}{=} V. \qquad \square$$

The intuition regarding the relationship between the equilibrium effort and the agent's level of social attentiveness is the following. First, in the presence of a third party, $V > 0$, the (positive) direct effect of a higher level of social attentiveness on the equilibrium effort $\partial e^*/\partial\beta$ dominates the (non-positive) indirect effect $\frac{\partial e^*}{\partial w_R^*} \times \frac{\partial w_R^*}{\partial \beta}$, such that the equilibrium effort is strictly increasing in the agent's level of social attentiveness. Second, in the absence of a third party, $V = 0$, the principal reacts towards a higher β by cutting back w_R^* (if possible) to an extent that keeps the implemented effort fixed, so that the direct and the indirect effect are equally strong. However, a cut of the monetary incentives w_R^* is only possible if w_R^* is positive, which holds true if the agent's social attentiveness is sufficiently small. If $w_R^* = 0$, the principal cannot reduce w_R^* further, so that the equilibrium effort is increasing in the agent's level of social attentiveness.

Propositions 7 and 12 show that with a socially attentive agent, the principal optimally either provides no monetary incentives or incentives that are rather weak. The first result provides a possible explanation for the puzzle that empirically only a fraction of employees experience monetary incentives.[26] The second result provides an explanation why monetary incentives are, if they exist, often weaker than predicted by standard agency models.[27]

Propositions 10 and 12 directly imply the following result.

Corollary 2: *Let $c''' \geq 0$ and $p''' \leq 0$, such that the principal's problem is concave. The expected surplus $E[s|e^*]$ is increasing in the agent's social attentiveness β if $V > 0$, while it is weakly increasing in β if $V = 0$.*

Social attentiveness, or a higher level of social attentiveness, is thus (at least weakly) beneficial for the surplus the principal and the agent generate. This holds true because social attentiveness increases the implemented effort level and thus mitigates the problem that an inefficiently low effort level is implemented.

III.3.4 Example

Suppose effort costs are quadratic, $c(e) = \alpha e^2$, where $\alpha > 0$, and effort is measured in units of success probability, such that $p(e) = e$ for all $e \leq 1$ and $p(e) = 1$, otherwise. Let $R + V < 2\alpha$ to guarantee an interior solution of e^*.

The principal optimally sets $w_0^* = 0$ and

$$w_R^* = \begin{cases} \frac{R}{2} - \frac{\beta}{1-\beta} \times \frac{R+V}{2} & \text{for } \beta < \frac{R}{2R+V}, \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

Thus, if the agent is sufficiently socially attentive, $\beta \geq \frac{R}{2R+V}$, the principal sets no monetary incentives, $w_R^* = w_0^*$. If, for example, the principal and the third party benefit from success to the same extent, such that $R = V$, the threshold is $1/3$. In general, the threshold is increasing in R, decreasing in V, and between 0 and $1/2$.

While the efficient effort is $e^{\text{efficient}} = \frac{R+V}{2\alpha}$, the implemented effort is only

$$e^* = \begin{cases} \frac{R+\beta V}{4\alpha} & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{\beta(R+V)}{2\alpha} & \text{otherwise.} \end{cases} \tag{49}$$

The principal's expected utility is

$$E\left[u_P|\left(w_0^*, w_R^*\right)\right] = \begin{cases} \frac{R+\beta V}{4\alpha}\left(\frac{R}{2} + \frac{\beta}{1-\beta} \times \frac{R+V}{2}\right) & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{\beta(R+V)}{2\alpha}R & \text{otherwise.} \end{cases} \tag{50}$$

[26] For example, Bell and Van Reenen (2013) find that 40% of all workers in the UK receive part of their annual wage in form of a bonus. Cf. also Lemieux et al. (2009), Bryson et al. (2012), and Gittleman and Pierce (2013).

[27] See, for example, Williamson (1985), Holmström and Milgrom (1990) or Che and Yoo (2001). Holmström and Milgrom (1990, p. 93) summarize that it is a "mystery to organizational observers, why there is so much less reliance on high-powered incentives than basic agency theory would suggest".

The agent's expected utility is

$$E\left[u_A|\left(w_0^*, w_R^*\right)\right] = \begin{cases} \frac{(R+\beta V)^2}{16\alpha} & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{(\beta(R+V))^2}{4\alpha} & \text{otherwise.} \end{cases} \tag{51}$$

The expected surplus generated is

$$E\left[s|e^*\right] = \begin{cases} \frac{R+\beta V}{4\alpha}(R+V) - \alpha\left(\frac{R+\beta V}{4\alpha}\right)^2 & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{\beta(R+V)}{2\alpha}(R+V) - \alpha\left(\frac{\beta(R+V)}{2\alpha}\right)^2 & \text{otherwise.} \end{cases} \tag{52}$$

Figures 3-5 illustrate the example for the values $R = V = 1$ and $\alpha = 2$.

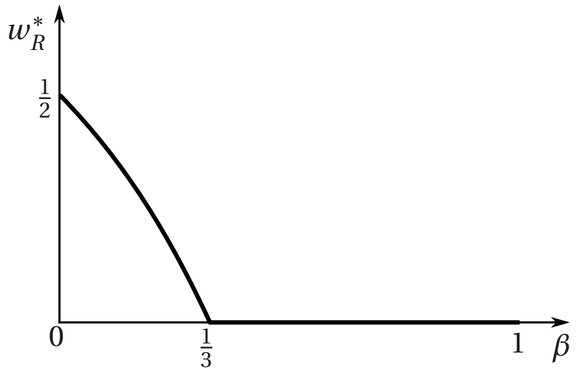

Figure 3: The wage w_R^*.

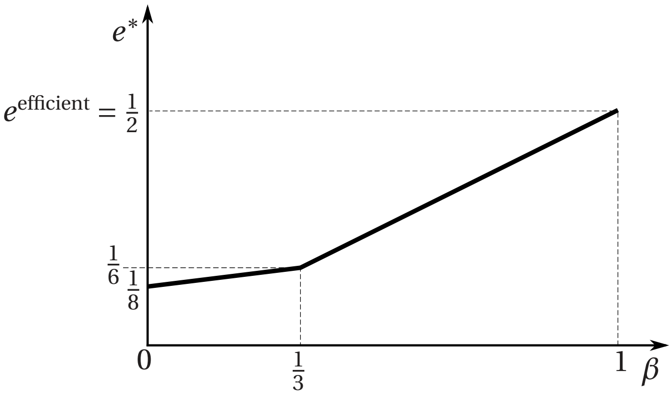

Figure 4: The effort e^*.

III.4 Contractible effort

We now analyze the model of a socially attentive agent in case his effort is contractible, i.e., there is no moral hazard. This allows us to compare the two constellations in the next step. With contractible effort, a contract consists of a wage payment w and an effort level e the agent has to exert. The following proposition characterizes the properties of the optimal contract; see Appendix C for the formal derivation.

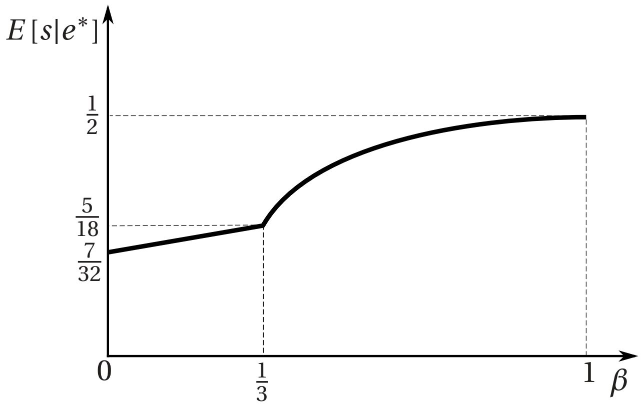

Figure 5: The surplus $E[s|e^*]$.

Proposition 13: *Suppose effort is contractible. The principal optimally implements the effort level*

$$e^{contractible} = \max\left\{e^{uncon}, \underline{e}\right\},$$

where $e^{uncon} := \text{argmax}_e \; p(e)R - c(e) + \beta p(e)V$ *and* $\underline{e} := \max\{e|c(e) - \beta p(e)(R+V) = 0\}$, *and sets the wage*

$$w^{contractible} = \frac{c\left(e^{contractible}\right) - \beta p\left(e^{contractible}\right)(R+V)}{1-\beta}.$$

Define the threshold $\bar{\beta} := \frac{c\left(e^{efficient}\right)}{p\left(e^{efficient}\right)(R+V)}$, *where the threshold satisfies* $\bar{\beta} \in (0,1)$.

1. *Case* $V > 0$. *The implemented effort is increasing in the agent's social attentiveness,* $\partial e^{contractible}/\partial\beta > 0$. *The principal implements an inefficiently low effort* $e^{contractible} < e^{efficient}$ *if* $\beta < \bar{\beta}$, *the efficient effort* $e^{contractible} = e^{efficient}$ *if* $\beta = \bar{\beta}$, *and an inefficiently high effort* $e^{contractible} > e^{efficient}$ *if* $\beta > \bar{\beta}$.
2. *Case* $V = 0$. *The implemented effort is weakly increasing in the agent's social attentiveness,* $\partial e^{contractible}/\partial\beta \geq 0$. *The principal implements the efficient effort* $e^{contractible} = e^{efficient}$ *if* $\beta \leq \bar{\beta}$, *while she implements an inefficiently high effort* $e^{contractible} > e^{efficient}$ *if* $\beta > \bar{\beta}$.

For the benchmark of an egoistic agent and the absence of a third party, $\beta = V = 0$, a scenario which is extensively studied by the existing literature, we obtain the standard result that contractibility of effort leads to the implementation of the efficient effort. Interestingly, as shown in Proposition 13, this is generally no longer true if the agent is socially attentive, $\beta > 0$, or a third party is present, $V > 0$.

First, if the agent's social attentiveness is rather low, $\beta < \bar{\beta}$, and a third party is present, $V > 0$, the principal and the agent bargain to a contract that maximizes the sum of their expected utilities, but they do not, or not fully, take into account how their contract affects the third party. Accordingly, an inefficiently low effort level is implemented.

Second, if the agent's social attentiveness is rather high, $\beta > \bar{\beta}$, the principal can

– and optimally does – exploit the agent by requiring an excessive effort level without having to compensate him with more than the minimal possible wage payment. Thus, while with non-contractibility there is the problem of under-implementation of effort, the opposite problem could arise with contractibility.

The above results have consequences for the generated surplus.

Proposition 14: *Suppose effort is contractible.*

1. *Case* $V > 0$. *The expected surplus is increasing in the agent's social attentiveness when* $\beta < \bar{\beta}$, $\partial E\left[s|e^{contractible}\right]/\partial\beta > 0$, *but decreasing in it when* $\beta > \bar{\beta}$, $\partial E\left[s|e^{contractible}\right]/\partial\beta < 0$. *In particular,* $E\left[s|e^{contractible}, \beta \neq \bar{\beta}\right] < E\left[s|e^{contractible}, \bar{\beta}\right]$ $= E\left[s|e^{efficient}\right]$.
2. *Case* $V = 0$. *The expected surplus is weakly decreasing in the agent's social attentiveness,* $\partial E\left[s|e^{contractible}\right]/\partial\beta \leq 0$. *In particular,* $E\left[s|e^{contractible}, \beta \leq \bar{\beta}\right] = E\left[s|e^{efficient}\right] > E\left[s|e^{contractible}, \beta > \bar{\beta}\right]$.

For both cases, $E\left[s|e^{contractible}, \beta = 0\right] > \lim_{\beta\to 1} E\left[s|e^{contractible}\right] = 0$.

PROOF: The results stated in the enumeration of Proposition 14 directly follow from Proposition 13 and the concavity of the expected surplus in the implemented effort level. The last result is readily obtained when plugging in the effort level the principal implements $e^{\text{contractible}}$ into the formula for the expected surplus (38). □

Proposition 14 reveals that social attentiveness (or a higher level of social attentiveness) could actually harm the surplus the principal and the agent generate. This result is possibly surprising, since one usually expects that socially attentive preferences have positive effects: social attentiveness should cause parties to interact and cooperate in a more social and better way, thus generating together a higher surplus, rather than a lower one. While this presumption is right with non-contractible effort, it is wrong with contractible effort. The reason is that, with contractible effort, social attentiveness – if sufficiently high – causes the principal to implement an excessive effort level, which harms the surplus. The following proposition shows that the principal benefits from the agent's social attentiveness with contractible effort also. However, only the principal experiences a rent in case effort is contractible.

Proposition 15: *Suppose effort is contractible. The principal's expected utility is positive,*

$$E\left[u_P|\left(w^{contractible}, e^{contractible}\right)\right] > 0,$$

while the agent's expected utility is zero,

$$E\left[u_A|\left(w^{contractible}, e^{contractible}\right)\right] = 0.$$

The principal's expected utility is increasing in the agent's social attentiveness,

$$dE\left[u_P|\left(w^{contractible}, e^{contractible}\right)\right]/d\beta > 0.$$

In particular, her expected utility is higher if the agent is socially attentive rather than egoistic, $E\left[u_P|\left(w^{contractible}, e^{contractible}\right), \beta > 0\right] > E\left[u_P|\left(w^{contractible}, e^{contractible}\right), \beta = 0\right]$.

PROOF: If effort is contractible, the principal could implement the same effort level and offer the same wage payment as the expected wage payment when effort is non-contractible. Since this would yield the principal a positive rent, cf. Proposition 8, the principal's rent under the optimal contract must be positive as well:

$$E\left[u_P|\left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] > 0. \tag{53}$$

By construction of the optimal contract, it is never optimal for the principal to leave any rent to the agent, such that the agent's expected utility is zero:

$$E\left[u_A|\left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] = 0. \tag{54}$$

It remains to be shown that the principal's expected utility is increasing in the agent's social attentiveness. To see this, consider the optimal contract the principal sets when her agent's social attentiveness is $\underline{\beta}$. If her agent's social attentiveness is higher, namely $\bar{\beta} > \underline{\beta}$, the principal could offer the same wage payment, but demand a higher effort, which results in a higher expected utility for her. Therefore, the principal's expected utility must be increasing in the agent's social attentiveness,

$$\frac{E\left[u_P|\left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right]}{d\beta} > 0. \qquad \square$$

In order to identify the effects of contractibility, we next compare the case with contractible effort to the case with non-contractible effort.

III.4.1 Comparison with the case of non-contractibility

The following proposition shows that the contractibility of effort has a positive effect on the effort level implemented by the principal.

Proposition 16: *The principal implements a higher effort level if effort is contractible than when it is non-contractible:* $e^{contractible} > e^*$.

PROOF: If effort is contractible, we know from Proposition 13 that the principal implements the effort level $e^{\text{contractible}} = \max\{e^{\text{uncon}}, \underline{e}\}$. Thus, $e^{\text{contractible}} \geq e^{\text{uncon}}$. Recognize that $e^{\text{uncon}} :=$ $\text{argmax}_e\ p(e)R - c(e) + \beta p(e)V$ or equivalently that e^{uncon} is the solution of the following condition:

$$p'(e)(R + \beta V) - c'(e) = 0. \tag{55}$$

Recall that in case effort is non-contractible, the equilibrium effort e^* solves (28), i.e., the following condition

$$p'(e)\left((1-\beta)w_R^* + \beta(R+V)\right) - c'(e) = 0. \tag{56}$$

Comparing (55) and (56), we directly see – using $w_R^* < R$, by Proposition 5 – that

$$R + \beta V > (1-\beta) w_R^* + \beta(R + V). \tag{57}$$

Due to the concavity of the success function, $p'' \leq 0$, and the convexity of the effort cost function, $c'' > 0$, this implies that $e^{\text{contractible}} \geq e^{\text{uncon}} > e^*$. □

Although the contractibility of effort causes a higher effort level to be implemented, this is not necessarily beneficial for the surplus.

Proposition 17: *If the agent's social attentiveness β is sufficiently low, the expected surplus the principal and the agent generate is higher when effort is contractible than when it is non-contractible: $E\left[s|e^{contractible}\right] > E\left[s|e^*\right]$. It is vice versa if the agent's social attentiveness β is sufficiently high: $E\left[s|e^{contractible}\right] < E\left[s|e^*\right]$.*

PROOF: Suppose first that the agent's social attentiveness β is sufficiently low, $\beta \leq \bar{\beta}$. Propositions 13 and 16 show that

$$e^* < e^{\text{contractible}} \leq e^{\text{efficient}}. \tag{58}$$

By the concavity of the expected surplus and since the effort $e^{\text{efficient}}$ maximizes the expected surplus, it thus holds that

$$E\left[s|e^{\text{contractible}}\right] > E\left[s|e^*\right]. \tag{59}$$

Suppose next that the agent's social attentiveness β is sufficiently high. When effort is contractible and $\beta > \bar{\beta}$, we know from Proposition 13 that $e^{\text{contractible}} > e^{\text{efficient}}$ and from equations (159)-(163) that then $e^{\text{contractible}} = \underline{e}$ and that $\underline{e}$ is increasing in β. Accordingly, for all $\beta > \bar{\beta}$,

$$E\left[s|e^{\text{contractible}}\right] < E\left[s|e^{\text{efficient}}\right] \tag{60}$$

and the difference between $E\left[s|e^{\text{efficient}}\right]$ and $E\left[s|e^{\text{contractible}}\right]$ is increasing in β,

$$\frac{\partial\left(E\left[s|e^{\text{efficient}}\right] - E\left[s|e^{\text{contractible}}\right]\right)}{\partial\beta} > 0. \tag{61}$$

When effort is non-contractible, we know from Corollary 1 that $E\left[s|e^*\right] < E\left[s|e^{\text{efficient}}\right]$ and that $\lim_{\beta\to 1} E\left[s|e^*\right] = E\left[s|e^{\text{efficient}}\right]$. Together with (60) and (61), this implies that

$$E\left[s|e^{\text{contractible}}\right] < E\left[s|e^*\right] \tag{62}$$

if the agent's social attentiveness β is sufficiently high. □

This result is interesting, because in the benchmark of an egoistic agent, the non-contractibility of effort leads to a loss of surplus. This insight is not robust, since with socially attentive preferences the surplus could actually be higher when effort is non-contractible than when it is contractible. Intuitively, with a sufficiently socially attentive agent, the under-implementation problem caused by non-contractibility is a less severe issue than the over-implementation problem caused by contractibility.

The next proposition describes how contractibility affects the parties' utilities.

Proposition 18: *The principal's expected utility is higher when effort is contractible than when it is non-contractible:*

$$E\left[u_P|\left(w^{contractible}, e^{contractible}\right)\right] > E\left[u_P|\left(w_0^*, w_R^*\right), e^*\right].$$

In contrast, the agent's expected utility is lower when effort is contractible than when it is non-contractible:

$$E\left[u_A|\left(w^{contractible}, e^{contractible}\right)\right] < E\left[u_A|\left(w_0^*, w_R^*\right), e^*\right].$$

PROOF: When effort is non-contractible, we know from Proposition 8 that the agent experiences a rent: $E\left[u_A|\left(w_0^*, w_R^*\right), e^*\right] > 0$. When effort is contractible, the principal could offer the agent the same wage payment as the expected wage payment the agent experiences when effort is non-contractible. That is, the wage $\tilde{w} = p(e^*)w_R^*$. With effort being contractible and $E\left[u_A|\left(w_0^*, w_R^*\right), e^*\right] > 0$, however, the principal can require an effort level $e^{\text{contractible}}$ that exceeds e^*, which improves her expected utility. Under the optimal contract, the principal must thus experience a higher expected utility when effort is contractible than when it is non-contractible:

$$E\left[u_P|\left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] > E\left[u_P|\left(w_0^*, w_R^*\right), e^*\right]. \tag{63}$$

From Propositions 8 and 15 it directly follows that

$$E\left[u_A|\left(w_0^*, w_R^*\right), e^*\right] > E\left[u_A|\left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] = 0. \qquad \square$$

The principal always benefits from effort being contractible, since this allows her to perfectly control the agent's effort without having to care about incentive compatibility. This is in contrast to the agent, who suffers from contractibility, because the principal then does not leave the agent any rent.

The results stated in Propositions 17 and 18 have important practical implications: If technically possible and legally permissible, a principal would always use methods that make her agent's effort observable and contractible. Such methods are controversially debated, since they often interfere with the agent's privacy. Our results contribute to this debate by showing that regulatory boundaries on the methods principals can use to control their agents, as well as legal restrictions on what can be specified in contracts, could protect the agents from being exploited by their principals. Additionally, such regulatory boundaries and legal restrictions need not only help agents, but could also improve efficiency, i.e., raise the generated surplus.

III.4.2 Example

As in subsection III.3.4, suppose $c(e) = \alpha e^2$, with $\alpha > 0$, $p(e) = e$ for all $e \leq 1$ and $p(e) = 1$, otherwise. Furthermore, let $R + V > 2\alpha$.

The principal optimally sets

$$w^{\text{contractible}} = \begin{cases} \frac{(R+\beta V)^2 - 2\beta(R+\beta V)(R+V)}{4\alpha(1-\beta)} & \text{for } \beta < \frac{R}{2R+V}, \\ 0 & \text{otherwise.} \end{cases} \tag{64}$$

The optimally implemented effort level is

$$e^{\text{contractible}} = \begin{cases} \frac{R+\beta V}{2\alpha} & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{\beta(R+V)}{\alpha} & \text{otherwise.} \end{cases} \tag{65}$$

The agent's expected utility is $E\left[u_A | \left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] = 0$. The principal's expected utility is $E\left[u_P | \left(w^{\text{contractible}}, e^{\text{contractible}}\right)\right] = e^{\text{contractible}} R - w^{\text{contractible}}$. The expected surplus is

$$E\left[s | e^{\text{contractible}}\right] = \begin{cases} \frac{R+\beta V}{2\alpha}(R+V) - \alpha\left(\frac{R+\beta V}{2\alpha}\right)^2 & \text{for } \beta < \frac{R}{2R+V}, \\ \frac{\beta(R+V)}{\alpha}(R+V) - \alpha\left(\frac{\beta(R+V)}{\alpha}\right)^2 & \text{otherwise.} \end{cases} \tag{66}$$

Observe that the expected surplus is higher with contractible effort than with non-contractible effort, $E\left[s | e^{\text{contractible}}\right] > E\left[s | e^*\right]$, if $\beta < 2/3$, the same for $\beta = 2/3$, but lower if $\beta > 2/3$. Figures 6-8 illustrate the example for the values $R = V = 1$ and $\alpha = 2$.

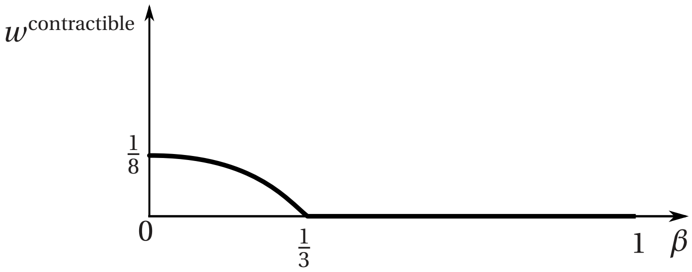

Figure 6: The wage $w^{\text{contractible}}$.

III.4.3 The problem with unlimited liability

We next briefly examine the scenario where the agent is not wealth constrained and his liability is unlimited. The principal is then free to set arbitrary wage payments. The principal's problem is to maximize her expected utility subject to the agent's participation and incentive constraints. In the optimum, the agent's participation con-

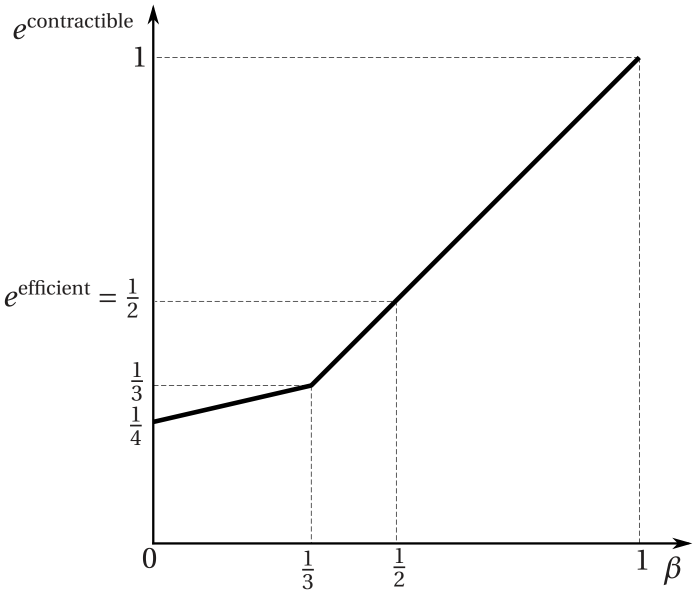

Figure 7: The effort $e^{\text{contractible}}$.

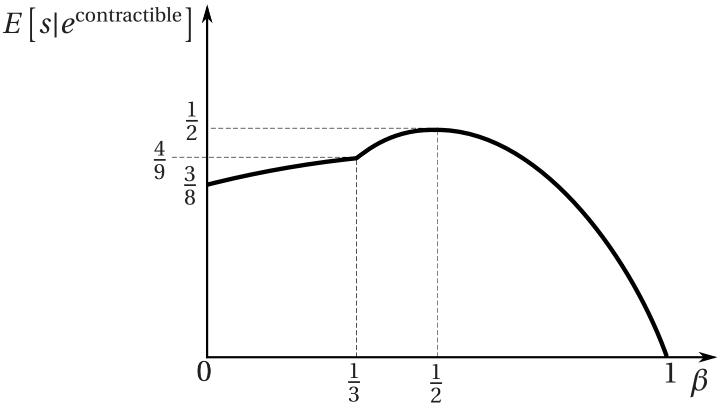

Figure 8: The surplus $E\left[s|e^{\text{contractible}}\right]$.

straint has to bind,[28] such that the agent's expected utility equals his reservation utility $E[u_A] = 0$. Substituting the binding participation constraint into the principal's objective function allows us to write the problem as

$$\max_{e} \; p(e)R - c(e) + \beta p(e)V. \tag{67}$$

The principal thus optimally implements the effort level e^{uncon}, which solves

$$p'(e)(R + \beta V) - c'(e) = 0. \tag{68}$$

We directly see that, in the presence of a third party, $V > 0$, e^{uncon} falls short of $e^{\text{efficient}}$. In contrast, in the absence of a third party, $V = 0$, we have $e^{\text{uncon}} = e^{\text{efficient}}$. Thus, the principal either implements an inefficiently low effort or the efficient effort.

III.5 Conclusions

In Koch and Weinschenk (2021) we established that with socially attentive preferences, the principal may optimally refuse to provide monetary incentives although monetary incentives are effective. We have illustrated here the consequences for the parties' utilities and the efficiency aspects of their relationship with socially attentive preferences in an agent. Irrespective of whether the agent is egoistic or socially attentive, both parties' expected utilities exceed their reservation utilities and the principal implements an inefficiently low effort level. However, the implemented effort level approaches the efficient level as the agent's social attentiveness goes to one. The principal's expected utility is increasing in the social attentiveness of her agent. In particular, the principal's expected utility is higher if her agent is socially attentive rather than egoistic.

We also find novel insights when the agent's effort is contractible. If effort is contractible the agent's expected utility lowers to the reservation utility, while the principal's utility and the implemented effort level increase. For the standard case in the literature (egoistic agent and absence of the third party), the principal implements an inefficiently low effort level when effort is not contractible, while the efficient effort is implemented when effort is contractible. Contractibility thus improves efficiency, i.e., causes that the principal and the agent achieve the maximal possible surplus. While also with socially attentive preferences the principal implements an inefficiently low effort when effort is not contractible, contractibility of effort does, in general, not cause the implementation of the efficient effort level. In particular, the principal requires the agent to exert an excessively high effort level if the agent's social attentiveness is rather high. The surplus the principal and the agent generate could in fact be lower when ef-

[28] Otherwise, i.e., if $E[u_A] > 0$, the principal can lower w_0 and w_R by some $\varepsilon > 0$, which lowers her wage costs, while leaving the implemented effort level unchanged.

fort is contractible rather than not contractible. Non-contractibility of effort thus does not only protect the agent's information rent, but could also protect the agent from being exploited and thereby raise the generated surplus. A practical implication of this result is that restrictions on the methods employers can use to control their employees could not only help the employees, but also enhance efficiency.

IV Endogenous Socially Attentive Preferences

Joint work with Philipp Weinschenk

We extend the analysis in Koch and Weinschenk (2021) in two regards in this paper. First, we show that the principal benefits from having a socially-attentive agent. This renders investments into increasing the level of social preferences attractive. We investigate the efficiency of such an investment if the agent's preferences are endogenous. We furthermore explore the case where both contractual partners have social preferences and the consequences for the contractual structure.

JEL Classification: D82, D91, M52.

Keywords: agency model, socially attentive preferences, incentives.

IV.1 Introduction

Motivating employees to exert effort is one of the main objectives of managerial leadership. The most prominent avenue for motivating selfish agents in standard contract theory is via a bonus for successful work, i.e. providing monetary incentives. Lately, other modeling perspectives have highlighted that agents might not only be interested in their monetary payoff, but also care about the purpose of their work and its wider impact on society. Companies may thus have a new opportunity to motivate their agents, for example if they can raise the sensitivity of their agent with regard to the potentially good work he is doing for the company, but also for society as a whole. If successfully implemented, such measures may increase the agent's effort provision. In Koch and Weinschenk (2021) we contribute to this strain of the theoretical debate by introducing socially attentive agents.

In Koch and Weinschenk (2021) an agent acts on behalf of a principal and the agent's non-contractible effort choice influences the probability distribution over outcomes. The outcome affects the principal, the agent, and possibly a third party. We allow the agent to have socially attentive preferences, i.e. he takes others' well-being into consideration and assigns a positive weight to their utilities. One important result of the analysis of this model is that despite being effective, irrespective of whether the agent is egoistic or socially attentive, it might be optimal for the principal to not provide monetary incentives to a sufficiently socially attentive agent. The intuition for this result is that providing monetary incentives is costly for the principal. Since a socially attentive agent reacts less strongly to monetary incentives than an egoistic one, the principal's costs of providing monetary incentives may – in comparison to the benefit of monetary incentives, in the form of a more motivated agent – be too high to justify monetary incentives.

In this paper, we investigate an extension to Koch and Weinschenk (2021) where the agent's preferences are endogenous.[29] We formalize this by letting the principal undertake investments that influence her agent's social attentiveness.[30] She could, for example, increase the agent's social attentiveness by highlighting the positive mission the firm has or the positive impact the agent's work has on society. Engaging in and maintaining this motivational strategy is costly to the principal and she therefore must decide on the optimal investment level. We find that while it is not always optimal for the principal to motivate her agent via monetary incentives, it is always optimal to motivate him via investments into his social attentiveness. We further find that the principal's investment is, in general, not efficient and determine the factors that cause the inefficiency.

We document the robustness of the results, by allowing that not only the agent, but also the principal is socially attentive, i.e., puts positive weight on others' utilities. We show that all our results stay qualitatively valid. Quantitatively, however, a socially attentive principal provides weakly higher monetary incentives than an egoistic principal. Moreover, a socially attentive principal's investment into the agent's preferences are closer to the social optimum than that of an egoistic principal. In a short extension, we allow that the agent's effort does not necessarily benefit the third party and that the agent may have unsocial preferences – defined as the agent putting negative weight on others' utilities. Interestingly, if the third party's expected utility is negatively related to the agent's effort choice – a scenario which is, for example, plausible in criminal organizations – the principal may be better off from having an egoistic or unsocial agent rather than a socially attentive agent.

Relation to the literature. Since our paper is an extension of Koch and Weinschenk (2021), it is naturally related to the literature cited there, which incorporates social preferences into agency models. In addition, this paper is related to the labor economics literature concerned with non-monetary motivation. Cassar and Meier (2018) provide an extensive overview over non-monetary components of job satisfaction and their in-

[29] It is well known – see, for example, Tversky and Kahneman (1981) or the modern salience theory introduced by Bordalo et al. (2013) – that preferences are not something fixed, but something that can be influenced by framing or reference points effects. For an overview of the literature on salience theories in markets, see Herweg et al. (2018). In practice, agents' preferences towards their work can be influenced, for example, via emphasizing the importance of the firms' mission or by underlining the value of the agent's work. The field of industrial and organizational psychology has identified an array of additional factors that influence employees' motivation. For an overview see, Steers et al. (2004).

[30] We follow the standard convention in the agency literature and talk about a male agent and a female principal.

fluence on recompensation schemes. Bandiera et al. (2005) present evidence that social preferences influence workers' productivity. They compare the productivities under piece rates with that under relative incentives (where individual effort imposes a negative externality on others). Bandiera et al. (2009) show that the social connections in firms influences the behavior of workers and managers. Ashraf and Bandiera (2017) emphasize that performing tasks that benefit others can increase altruistic capital. A higher level of altruistic capital improves the marginal utility of effort dedicated to improving others' welfare in the future. This can lead to a virtuous cycle of steadily accumulating altruistic capital. Here, policies that highlight the importance of the task of the agent to society or increase the value of the positive impact of the task, can elevate the agent's effort provision. We agree with Ashraf and Bandiera (2017) that preferences are not fixed and can be influenced. Our approach differs from theirs since the principal can invest into the agent's social preferences directly. Cassar (2019) highlights that the motivation of the principal is decisive for the effect of the agent's social preferences on a third party. She finds, experimentally, that selfish principals exploit agents' motivation to work hard in order to benefit a charity, by reducing monetary incentives. In contrast, a principal with strong social preferences does not reduce monetary incentives in her setup. In the same vain, Briscese et al. (2021) find that firms use Corporate Social Responsibility to reduce wages of motivated agents. Our model can account for these observations. The principal will trade off investing in stronger social preferences of their agent with monetary incentives. A higher level of social attentiveness of the principal reduces this effect.

We also contribute to the literature that explores the relationship between organizational missions and organizational performance. Most firms nowadays publish their missions and these often include commitments to social issues, such as protecting the environment, encouraging diversity, and supporting the community (Bartkus and Glassman, 2008). Bart and Baetz (1998) examine the relationship between mission statements and organizational performance, using a sample of Canadian organizations, and show that specific characteristics of mission statements are selectively associated with higher levels of organizational performance. Building on a sample of US and Canadian organizations, Bart et al. (2001) show that mission statements are positively associated with financial performance. They also emphasize that, in order to be successful, mission statements must be rational, contain sound content, have organizational alignment, and bring sufficient behavioral change in the desired direction. Their research thus shows that effective mission statements must be accompanied by an array of (possibly costly) measures by the organization. These findings are confirmed by Bartkus and Glassman (2008), who show that mission statements that are merely the result of institutional pressures cause symbolic statements that are not related to the actual behaviors of organizations. Williams (2008) uses textual and content analysis methods to investigate the mission statements of Fortune 1000 firms. She shows that higher-performing firms include aspects of philosophies, targeted markets, strategies for survival, public image, team work, safety, and concern for employees significantly more often than lower-performing firms. Kahn (2021) contributes that missions do not only serve to attract suitable employees (cf. Non et al., 2021), but that they also increase the effort of employed personnel over all tasks, while financial incentives only improve the effort related to the incentivized task.

Our paper contributes to this literature by showing that it is always optimal for principals (i) to invest into measures that increase the agent's social attentiveness and (ii) given that the agent is socially attentive, to invest into measures that improve the organization's corporate social responsibility, i.e., the alignment of the society's and the principal's interests.

IV.2 Benchmark moral hazard model

We extend the model developed in Koch and Weinschenk (2021) in two major directions in this paper, endogenous preferences and a socially attentive principal. In order to establish the benchmark moral hazard model, we first recapitulate Koch and Weinschenk (2021) to the necessary extent for our purpose here.

Primitives. A principal needs to hire an agent. When working for the principal, the agent exerts effort $e \in [0, e^{\max}]$, where $e^{\max}$ is positive and could be finite or infinite. The agent's effort choice determines the probability distribution over outcomes. With probability $p(e)$ the agent yields a successful outcome that is associated with a high return $R > 0$ for the principal, while with probability $1 - p(e)$ the outcome is unsuccessful which causes a low return normalized to zero. The outcome may also affect a third party, which experiences a payoff $V \geq 0$ if the agent succeeds and 0 otherwise. The case $V = 0$ captures the situation where a third party is absent or unaffected. While the outcome is contractible, the agent's effort is non-contractible, i.e., there is moral hazard. A contract is hence a pair $(w_0, w_R) \in \mathbb{R}^2$, where w_0 is the payment to the agent if he yields the unsuccessful outcome and thereby generates the return 0, while w_R is the payment if he yields the successful outcome and generates return R. The agent's wealth is normalized to zero, such that payments cannot be negative: $w_0, w_R \geq 0$.

Preferences. The principal and the agent are risk neutral. We follow Koch and Weinschenk (2021) by supposing that the agent might have socially attentive preferences. Formally, the agent puts a weight $\beta \in [0, 1]$ on the utilities of others. We henceforth say that the agent is egoistic if he puts zero weight on others' utilities, $\beta = 0$, while the agent is socially attentive if he puts positive weight, $\beta > 0$. Unless explicitly stated differently, we suppose that the agent puts at least slightly more weight on his own utility than on that of others, $\beta < 1$. While we initially take β as given, we later consider the possibility that the principal can take measures to influence β. The agent's expected utility is

$$E[u_A] = p(e)\,w_R + (1 - p(e))\,w_0 - c(e) + \beta E[u_{\neg A}], \tag{69}$$

where $p(e)\,w_R + (1 - p(e))\,w_0$ is the expected wage payment, $c(e)$ are the agent's effort costs, and $\beta E[u_{\neg A}]$ is the weighted sum of the other parties' utilities. The principal's expected utility equals the expected difference between the return she earns and the payment she makes to the agent:

$$E[u_P] = p(e)(R - w_R) + (1 - p(e))(0 - w_0). \tag{70}$$

In (70) the principal maximizes the expected profit, an assumption we maintain for the main analysis of the paper. In section IV.5, we discuss the case where the principal is also socially attentive. The third party's expected utility is

$$E[u_T] = p(e)V. \tag{71}$$

By including (70) and (71) into (69), we can rewrite the agent's expected utility as

$$E[u_A] = p(e)\,w_R + (1 - p(e))\,w_0 - c(e) + \beta\big(p(e)(R - w_R + V) - (1 - p(e))\,w_0\big). \tag{72}$$

Assumptions. We impose the standard assumptions on the effort cost function c and the success function p: c and p are twice continuously differentiable, $c'(e)$, $c''(e) > 0$ for $e > 0$, $c(0) = c'(0) = 0$, $\lim_{e \to e^{\max}} c'(e) = \infty$, $p(0) = 0$, $p'(e) > 0$, and $p''(e) \leq 0$. Thus, the effort cost function is increasing and convex and the success function is increasing and weakly concave.

Timing. First, the principal offers a contract to the agent, who then decides whether to accept or reject it. If the agent rejects, all parties receive reservation utilities of zero and the game ends. In case of acceptance, the agent exerts effort. Finally, the outcome is realized and the agent receives the contracted payment. Figure 9 summarizes the timing of the contracting game.

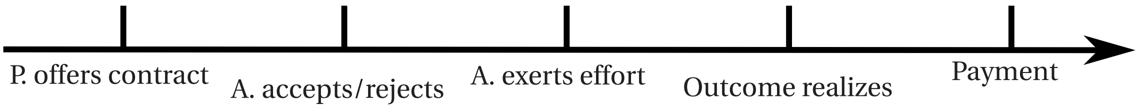

Figure 9: Timeline.

IV.3 Analysis of the benchmark moral hazard model

We follow the argument regarding the optimal contract developed in Koch and Weinschenk (2021). Differentiating the agent's expected utility with respect to effort yields the agent's effort choice for any contract, i.e.

$$\frac{\partial E[u_A]}{\partial e} = p'(e)\big((w_R - w_0)(1-\beta) + \beta(R+V)\big) - c'(e). \tag{73}$$

The optimal effort for any contract is, for standard arguments, $e^* \geq 0$. The principal's problem is to maximize her expected utility subject to the agent's incentive constraint, his participation constraint, and the limited liability constraints. Koch and Weinschenk (2021) prove that under these conditions the following holds.

Proposition 19: *There always exists an optimal contract $\left(w_0^*, w_R^*\right)$. The optimal contract satisfies $w_0^* = 0$ and $w_R^* \in [0, R)$.*

The intuition for the properties of the optimal contract in Proposition 19 are as follows. First, in case the agent does not succeed, it is optimal to make only the minimal possible wage payment; thus $w_0^* = 0$.[31] Second, it is never optimal for the principal to provide a wage payment that equals or exceeds her return R and it is not possible to offer a negative wage payment; thus $w_R^* \in [0, R)$.

A useful implication of Proposition 19 is that, for all potentially optimal contracts, the agent's effort choice e^* is determined by the first-order condition[32]

$$\frac{\partial E[u_A]}{\partial e} = p'(e)\left(w_R(1-\beta) + \beta(R+V)\right) - c'(e) = 0. \tag{74}$$

Implicitly differentiating equation (74) yields that

$$\frac{\partial e^*}{\partial w_R} = -\frac{p'(e^*)(1-\beta)}{p''(e^*)\left(w_R(1-\beta) + \beta(R+V)\right) - c''(e^*)}, \tag{75}$$

which is positive. Thus, with an egoistic agent ($\beta = 0$), as well as with a socially attentive agent ($\beta > 0$), monetary incentives are effective, in the sense that the agent exerts more effort, the higher the monetary incentives are.

We further obtain that

$$\frac{\partial e^*}{\partial \beta} = -\frac{p'(e^*)(-w_R + R + V)}{p''(e^*)\left(w_R(1-\beta) + \beta(R+V)\right) - c''(e^*)}, \tag{76}$$

which is positive for all $w_R \in [0, R)$. Hence, for all potentially optimal contracts, the agent exerts more effort, the more socially attentive he is.

[31] Having a payment of zero should be interpreted as the principal offering a payment equal to the minimal possible payment (which we normalized to zero, but which could be different in general).

[32] To derive (74), we used that the first-order condition not only applies if $(w_R - w_0)(1-\beta) + \beta(R+V) > 0$, but also if $(w_R - w_0)(1-\beta) + \beta(R+V) = 0$.

We also get that

$$\frac{\partial e^*}{\partial V} = \frac{\partial e^*}{\partial R} = -\frac{p'(e^*)\beta}{p''(e^*)\big(w_R(1-\beta)+\beta(R+V)\big)-c''(e^*)}, \tag{77}$$

which is positive if $\beta > 0$ and zero if $\beta = 0$. Therefore, while a socially attentive agent reacts positively towards a higher payoff of the third party or a higher return of the principal, an egoistic agent does not react.

The second and last result from Koch and Weinschenk (2021) necessary for our purposes here is that in the optimal contract, the principal does not provide monetary incentives to a sufficiently socially attentive agent. The intuition for this is that, first, socially attentive preferences limit the effectiveness of monetary incentives, in the sense that a socially attentive agent reacts less strongly to monetary incentives than an egoistic agent does, as can be seen in (75). Second, providing monetary incentives is costly for the principal, see (70). Accordingly, if the agent is sufficiently socially attentive, the principal's costs of providing monetary incentives are too high, in comparison to their benefit (a more motivated agent), to justify monetary incentives. Proposition 20 summarizes.

Proposition 20: *If the agent is egoistic ($\beta = 0$) or his social attentiveness is sufficiently low (β is small), the principal optimally sets monetary incentives, $w_R^* > w_0^*$. In contrast, the principal sets no monetary incentives, $w_R^* = w_0^*$, if the agent's social attentiveness β is sufficiently high.*

IV.4 Endogenous preferences

We now extend our analysis from Koch and Weinschenk (2021) by supposing that the principal can influence the strength of the agent's social attentiveness β. In practice, this could be achieved by increasing an agent's identification with the company he is working for or by emphasizing the importance of the company's mission, i.e., the relevance of the agent's task. We capture this formally by supposing that the principal can elevate the agent's social attentiveness by investing into it.

Denoting the initial level of social attentiveness by β_0, the principal's problem is to select a $\beta \in [\beta_0, 1]$, which causes costs $\chi(\beta)$. By investing the amount $\chi(\beta)$, the principal thus achieves that her agent has social attentiveness β. We let $\beta_0 < 1$, such that there is scope for strengthening the agent's social attentiveness. We assume that χ is twice continuously differentiable, $\chi'(\beta)$, $\chi''(\beta) > 0$ for $\beta > \beta_0$, $\chi(\beta_0) = \chi'(\beta_0) = 0$, and $\lim_{\beta \to 1} \chi'(\beta) = \infty$.[33]

The principal's problem is to find the optimal investment $\chi(\beta)$, or equivalently the

[33] Observe that the assumptions we impose on the function χ are essentially the same as the ones we imposed before on the effort-costs function c.

optimal β. Using that under the optimal contract $w_0 = 0$, her problem is

$$\max_{\beta} E[u_P] = p(e(w_R, \beta))(R - w_R) - \chi(\beta). \tag{78}$$

The level of β the principal optimally sets is denoted by β^*, where β^* solves the first-order condition

$$\frac{\partial E[u_P]}{\partial \beta} = p'(e(w_R, \beta))(R - w_R)\frac{\partial e(w_R, \beta)}{\partial \beta} - \chi'(\beta) = 0. \tag{79}$$

We directly see that for all potentially optimal contracts $\beta^* > \beta_0$ due to $w_R \in [0, R)$, i.e. it is always optimal for the principal to make a positive investment $\chi(\beta^*) > 0$ into the agent's social attentiveness. We next investigate whether the principal's investment in the agent's social attentiveness is efficient and therefore compare it to the level of investment a social planner would choose.

Proposition 21: *The principal always makes a positive investment $\chi(\beta^*)$ to increase the agent's social attentiveness, where β^* solves the first-order condition* (79)*. The principal's investment is lower [equal, higher] than that of a planner if and only if $V > [=, <]\bar{V}$, where $\bar{V} := \frac{\beta^*}{1-\beta^*}(R - w_R)$.*

PROOF: A planner seeks to maximize the expected surplus, i.e.

$$\max_{\beta} E[s] = p(e(w_R, \beta))(R + V) - c(e(w_R, \beta)) - \chi(\beta). \tag{80}$$

The planner's optimal β, denoted by $\beta^{\text{efficient}}$, solves the first-order condition

$$\frac{\partial E[s]}{\partial \beta} = \left[p'(e(\cdot))(R + V) - c'(e(\cdot))\right]\frac{\partial e(\cdot)}{\partial \beta} - \chi'(\beta) = 0. \tag{81}$$

Solving the first-order condition of the agent's maximization problem (74) for $c'(e(\cdot))$ yields

$$c'(e(\cdot)) = p'(e(\cdot))\left((1 - \beta)w_R + \beta(R + V)\right). \tag{82}$$

Using (82), we can rewrite (81) as

$$\frac{\partial E[s]}{\partial \beta} = p'(e(\cdot))\left[(1 - \beta)(R + V - w_R)\right]\frac{\partial e(\cdot)}{\partial \beta} - \chi'(\beta) = 0. \tag{83}$$

By the concavity of the expected surplus $E[s]$, $\beta^{\text{efficient}} > [=, <]\beta^*$ if and only if

$$\left.\frac{\partial E[s]}{\partial \beta}\right|_{\beta^*} = p'(e(\cdot))\left[(1 - \beta^*)(R + V - w_R)\right]\frac{\partial e(\cdot)}{\partial \beta} - \chi'(\beta^*) > [=, <]0. \tag{84}$$

Subtracting (79) from (84) yields, that $\beta^{\text{efficient}} > [=, <]\beta^*$, if and only if

$$(1 - \beta^*)(R + V - w_R) > [=, <]R - w_R \iff V > [=, <]\frac{\beta^*}{1 - \beta^*}(R - w_R). \qquad \square$$

It is interesting to realize that while it is not always optimal for the principal to motivate the agent via monetary incentives (cf. Proposition 20), it is always optimal to motivate him via investments into his social attentiveness (cf. Proposition 21).

There are two reasons why the principal chooses a different β and thus a different investment $\chi(\beta)$ than the planner. First, the principal ignores the negative effect on the agent's effort costs caused by stronger social attentiveness and the associated higher effort level. Second, the principal does not take into account the positive effect a higher level of social attentiveness has on the third party. If the third party's payoff V is large, the second effect dominates the first effect and the principal consequently underinvests, $\chi(\beta^*) < \chi(\beta^{\text{efficient}})$. In contrast, if V is small, the first effect dominates and the principal overinvests, $\chi(\beta^*) > \chi(\beta^{\text{efficient}})$.

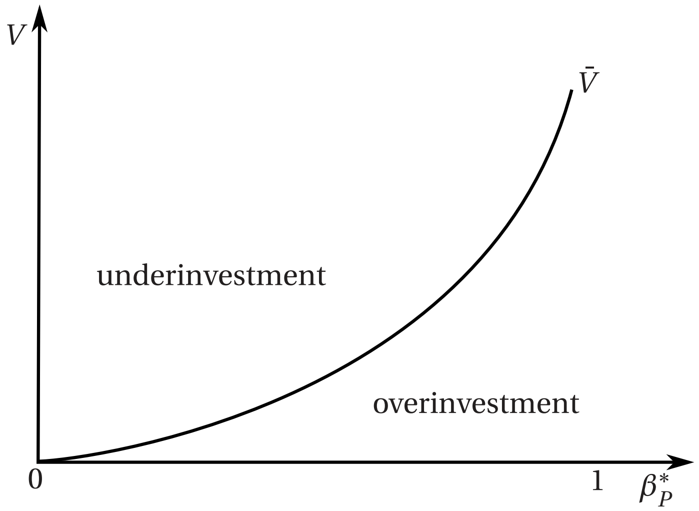

Figure 10: Schematic illustration of the critical value $\bar{V}$ and the areas of underinvestment, $\chi(\beta_P^*) < \chi(\beta_S^*) \Longleftrightarrow \beta_P^* < \beta_S^*$, and overinvestment, $\chi(\beta_P^*) > \chi(\beta_S^*) \Longleftrightarrow \beta_P^* > \beta_S^*$.

Figure 10 depicts a schematic illustration. For combinations of V and β_P^* above the line $\bar{V}$, $\beta_P^* < \beta_S^*$ and the principal under invests. She over invests for combinations under the line, where $\beta_P^* > \beta_S^*$. Summarizing, we find that it is likely that the principal will not invest in a socially optimal fashion, if she can manipulate the agent's social preferences with her investment.

We obtain an additional result if $c''' \geq 0$ and $p''' \leq 0$, such that the principal's problem is concave.

Proposition 22: *Let $c''' \geq 0$ and $p''' \leq 0$ such that the principal's problem is concave. If $w_R^* > w_0^*$, the power of monetary incentives is decreasing in the agent's social attentiveness β: $\partial(w_R^* - w_0^*)/\partial\beta < 0$.*

PROOF: Using (75) and the principal's expected utility differentiated with respect to w_R, i.e.

$$\frac{\partial E[u_P]}{\partial w_R} = p'(e(w_R,\beta))\frac{\partial e(w_R,\beta)}{\partial w_R}(R - w_R) - p(e(w_R,\beta)), \tag{85}$$

we directly see that when $c''' \geq 0$ and $p''' \leq 0$,

$$\frac{\partial E[u_P]}{\partial w_R} = \frac{\left(p'(\cdot)\right)^2 (1-\beta)(R-w_R)}{c''(\cdot) - p''(\cdot)\left(w_R(1-\beta) + \beta(R+V)\right)} - p(\cdot) \tag{86}$$

is decreasing in w_R and β. Suppose that $w_R^* > w_0^*$. Then w_R^* solves the first-order condition $\partial E[u_P]/\partial w_R = 0$. If β increases, w_R^* must thus be lowered to restore the first-order condition, so $\partial w_R^*/\partial \beta < 0$. Since by Proposition 19 w_0^* is independent of β,

$$\frac{\partial(w_R^* - w_0^*)}{\partial \beta} = \frac{\partial w_R^*}{\partial \beta} < 0. \qquad \square$$

Proposition 22 implies that the more the principal invests into the agent's social attentiveness and the higher the agent's social attentiveness therefore is, the less monetary incentives she optimally provides. Monetary incentives and the investments in the agent's attentiveness are thus substitutes.

Remark 1. The principal might also be able to take measures that ensure that a successful outcome is not only valuable for herself, but also for the society as a whole. In practice, firms could take an array of measures to improve their corporate social responsibility, for example by ensuring fair working conditions in supplying firms or investments in projects that reduce the environmental damages caused by their operations. This could formally be captured by assuming that the principal can increase the payoff V experienced by the third party for costs $\rho(V)$, where the function ρ satisfies the same assumptions as the function χ specified before. One can show that, in order to motivate her agent, the principal optimally invests to increase V whenever the agent is socially attentive.

Remark 2. It is also possible that the design of the contract itself influences the agent's social attentiveness. Formally, β is then a function of the contract (w_0, w_R). The design of the contract has then a direct as well as an indirect effect on the agent's effort choice.

IV.5 Socially attentive principal

We now consider the case where the principal and the agent both have socially attentive preferences for the net utility of the other and the third party, respectively.[34] We denote the strength of the principal's social attentiveness by $\lambda \in [0,1)$. The principal's expected utility is

$$E[u_P] = p(e)\,(R + \lambda V - (1-\lambda)\,w_R) - (1-p(e))(1-\lambda)\,w_0 - \lambda c(e). \tag{87}$$

[34] If we instead suppose that the principal and the agent both have socially attentive preferences regarding the gross expected utility of each other and the third party, then amplification effects arise. While this feature is unappealing, the qualitative effects stay unchanged.

It is straightforward to verify that (for essentially the same arguments) the previously derived results stay valid.[35] In particular, it is optimal not to reward an unsuccessful outcome, such that $w_0^* = 0$, and the principal optimally sets no monetary incentive, $w_R^* = w_0^*$, if the agent's social attentiveness β is sufficiently high. Interestingly, the principal provides weakly stronger incentives the more socially attentive she is.

Proposition 23: *Suppose the principal's problem is concave. For all $\bar{\lambda}$ and $\underline{\lambda}$, with $\bar{\lambda} > \underline{\lambda}$, it holds that $w_0^* = 0$ and that $w_R^*|_{\bar{\lambda}} \geq w_R^*|_{\underline{\lambda}}$. Moreover, if $w_R^*|_{\underline{\lambda}} > w_0^*$, then $w_R^*|_{\bar{\lambda}} > w_R^*|_{\underline{\lambda}}$.*

Proof: Taking into account that the agent chooses effort $e^* = e(w_R, \beta)$ and using that $w_0^* = 0$, we can write (87) as

$$E[u_P|\lambda] = p(e(w_R,\beta))\,(R + \lambda V - (1-\lambda)\,w_R) - \lambda c(e(w_R,\beta)). \tag{88}$$

Suppose now that, contrary to our claim, $w_R^*|_{\bar{\lambda}} < w_R^*|_{\underline{\lambda}}$. Then $w_R^*|_{\underline{\lambda}} > 0$ must hold true. Due to the concavity of the principal's problem, $w_R^*|_{\underline{\lambda}}$ must hence solve the first-order condition

$$\frac{\partial E\left[u_P|\underline{\lambda}\right]}{\partial w_R} = p'(e(\cdot))\frac{\partial e(\cdot)}{\partial w_R}\left(R + \underline{\lambda}V - (1-\underline{\lambda})\,w_R\right) - (1-\underline{\lambda})\,p(e(\cdot)) - \underline{\lambda}c'(e(\cdot))\frac{\partial e(\cdot)}{\partial w_R} = 0. \tag{89}$$

Plugging (82) into (89), we obtain

$$\begin{aligned}\frac{\partial E\left[u_P|\underline{\lambda}\right]}{\partial w_R} = p'(e(\cdot))\frac{\partial e(\cdot)}{\partial w_R}\left(R + \underline{\lambda}V - (1-\underline{\lambda})\,w_R\right) - (1-\underline{\lambda})\,p(e(\cdot))& \\ - \underline{\lambda}\frac{\partial e(\cdot)}{\partial w_R}p'(e(\cdot))\left((1-\beta)\,w_R + \beta(R+V)\right) = 0,&\end{aligned} \tag{90}$$

which can be rewritten as

$$R = \frac{\frac{(1-\underline{\lambda})\,p(e(\cdot))}{p'(e(\cdot))\frac{\partial e(\cdot)}{\partial w_R}} - (1-\beta)\,\underline{\lambda}V}{1-\beta\underline{\lambda}} + w_R. \tag{91}$$

Now consider the case with the higher level of social attentiveness $\bar{\lambda}$. Differentiating the principal's expected utility with respect to w_R and plugging (82) and (91) in yields

$$\left.\frac{\partial E\left[u_P|\bar{\lambda}\right]}{\partial w_R}\right|_{w_R^*|_{\underline{\lambda}}} = p'(e(\cdot))\frac{\partial e(\cdot)}{\partial w_R}(1-\beta)\underbrace{\left(\bar{\lambda} - \underline{\lambda}\frac{1-\beta\bar{\lambda}}{1-\beta\underline{\lambda}}\right)}_{>0}V - p(e(\cdot))\underbrace{\left((1-\bar{\lambda}) - (1-\underline{\lambda})\frac{1-\beta\bar{\lambda}}{1-\beta\underline{\lambda}}\right)}_{<0}, \tag{92}$$

which is positive. Hence, $w_R^*|_{\bar{\lambda}} < w_R^*|_{\underline{\lambda}}$, as supposed before, could never be optimal for a principal with a social attentiveness of $\bar{\lambda}$. Accordingly, $w_R^*|_{\bar{\lambda}} \geq w_R^*|_{\underline{\lambda}}$ must be true.

It remains to show that if $w_R^*|_{\underline{\lambda}} > w_0^* = 0$, then $w_R^*|_{\bar{\lambda}} > w_R^*|_{\underline{\lambda}}$. Note that if $w_R^*|_{\underline{\lambda}} > 0$, we must also have that $w_R^*|_{\bar{\lambda}} > 0$, because we know from before that $w_R^*|_{\bar{\lambda}} \geq w_R^*|_{\underline{\lambda}}$. Then $w_R^*|_{\underline{\lambda}}$ solves

[35] There is one noteworthy exception: a socially attentive principal optimally sets $w_R^* \in [0, R+V)$, such that the monetary incentives $w_R^* - w_0^*$ are possibly larger than the return R.

the first-order condition $\frac{\partial E[u_P|\underline{\lambda}]}{\partial w_R} = 0$, while $w_R^*|_{\bar{\lambda}}$ solves $\frac{\partial E[u_P|\bar{\lambda}]}{\partial w_R} = 0$. Equation (92) implies that $w_R^*|_{\bar{\lambda}} > w_R^*|_{\underline{\lambda}}$. □

These results are intuitive. If the principal's social attentiveness is relatively high, payments to the agent are relatively little painful for her, which is why she is willing to provide relatively high monetary incentives, if she provides any incentives.

Additionally, note that a socially attentive principal will make an investment decision with respect to β that is closer to the social optimum. A socially attentive principal's investment problem is

$$E[u_P|\lambda] = p(e(w_R, \beta))\,(R + \lambda V - (1-\lambda)\,w_R) - \lambda c(e(w_R, \beta)) - \chi(\beta), \tag{93}$$

where the assumptions on the function χ are the same as in section IV.4. Taking the first order condition of (93) with respect to β yields

$$\frac{\partial E[u_P|\lambda]}{\partial \beta} = \left[p'(e(\cdot))\,(R + \lambda V - (1-\lambda)\,w_R) - \lambda c'(e(\cdot))\right]\frac{\partial e(\cdot)}{\partial \beta} - \chi'(\beta) = 0. \tag{94}$$

Comparing (94) with (79) and (81) respectively, we see that a socially attentive principal optimally sets a level of social preferences, denoted β_λ, that is between the β^* sought by an egoistic principal and the $\beta^{\text{efficient}}$ a social planner implements. We find that for $\lambda \to 0$, $\beta_\lambda = \beta^*$ and for $\lambda \to 1$, $\beta_\lambda = \beta^{\text{efficient}}$. The socially attentive principal therefore still under- or overinvests, but the deviation from the socially optimal investment level decreases with the principal's level of social attentiveness.

IV.6 Unsocial agent and negative effects

Up until now, we assumed that the agent puts a non-negative weight on others' utilities and that the third party experiences a non-negative utility if the agent succeeds. We next describe the consequences when this is no longer true.

The case $\beta < 0$. An agent with this characteristic puts a negative weight on others' utilities, i.e., is unsocial. For all potentially optimal contracts, the agent is less willing to exert effort than in case her preferences are characterized by $\beta \geq 0$. For the principal, it is nonetheless optimal to incentivize the agent, i.e., set $w_R^* > w_0^*$, if $R + \beta V > 0$. As we can see from (73), the agent can then be motivated to exert positive effort for some $w_R < R$. This is not possible if $R + \beta V \leq 0$, such that the principal cannot do better than providing no monetary incentives $w_R^* = w_0^*$.

The case $V < 0$. A successful outcome for the principal is then negative for the third party. This is, for example, a plausible scenario for a criminal organization, which has a negative impact on the society, or a heavily polluting industry harming the environment. It is useful to rewrite (73) as

$$\frac{\partial E[u_A]}{\partial e} = p'(e)\left(w_R + \beta(R + V - w_R)\right) - c'(e), \tag{95}$$

where we used that $w_0^* = 0$, i.e., that the principal does not reward failure. We directly see from (95) that, in case $R + V - w_R < 0$, a constellation which has to be satisfied if $R + V < 0$, a more socially attentive agent (i.e., an agent with a higher β) is willing to exert less effort. Accordingly, the principal is worse off, the more socially attentive her agent is. This is a plausible result: A criminal organization, for example, may want to hire an agent who is unscrupulous and does not care about the negative impact the organization's actions have on the society.

IV.7 Conclusions

In standard agency models, the agent does not care how his decisions affect the well-being of others. This is a rather strong assumption, which we relaxed in Koch and Weinschenk (2021). Considering socially attentive preferences in agency models is not only important due to normative and positive reasons, but also yields new and interesting insights. We extended the analysis in this paper to include endogenous preferences and the possibility of a socially attentive principal.

With endogenous preferences we find that while it is not always optimal for the principal to motivate the agent via monetary incentives, it is always optimal to motivate him via investments that raise his social attentiveness. This underlines the practical relevance of measures which improve employees' identification with their employers' mission and the sense of purpose employees experience about their work. Furthermore, the principal's investment may be higher or lower than that of a planner. If the principal's problem is concave, monetary incentives and investments in the agent's attentiveness are substitutes.

Considering a socially attentive principal demonstrates that she, maybe unsurprisingly, provides relatively higher monetary incentives to her agent. Furthermore, a socially attentive principal makes more efficient investment decisions into her agent's social preferences compared to an egoistic principal.

V Economic Modeling of Social Norms

We review the theoretical economic literature on social norms. The goal of this contribution is to improve accessibility of economic models on social norms for all behavioral researchers in the social sciences. We first distinguish social norms from other theoretical social motivators for human behavior, namely social preferences and identity economics. In the main part of the paper we compile the contributions of economic social norm analysis. While the insights on social norms for economists have been numerous, social science research in general has learned little new. We highlight future research avenues by introducing lines of contact between theoretical and empirical economic work on social norms.

JEL Classification: B29, D91, Z13.

Keywords: social norms, economic theory, behavioral economics, evolutionary economics.

V.1 Introduction

Over the last forty years behavioral economists have intensified their engagement with psychological and social motivations for people's behavior (Tversky and Kahneman 1981; Gneezy and Rusticini, 2000; Thaler, 2008). The adherence to the paradigm of revealed preferences in neoclassical economics had made economic investigations into the motivational factors of people's behavior dispensable. Observing people's choices was sufficient to construct the house of neoclassical economic research. There had always been unorthodox voices in economics but it was not until the contradictions between theoretical prognosis and empirical observation of people's actual behavior were exposed (e.g. Dawes and Thaler, 1988) that more economists considered these forms of motivation as important to understanding economic issues. Behavioral economics has since made many strides into economic research. This review is designed to give an overview over the theoretical models of one identified social source of motivation for human behavior, namely social norms. As Kenneth Arrow (1994) pointed out succinctly, economists, despite their claim to employ methodological individualism as research agenda, have, at least implicitly, always dealt with social categories to explain individual human behavior.

> ”I do conclude that social variables, not attached to particular individuals, are essential in studying the economy or any other social system and that, in par-

> ticular, knowledge and technical information have an irremovably social component, of increasing importance over time." (Arrow, 1994, p. 8)

This review sheds light on the question of how economists think about the social variable called a social norm. The ultimate goal of this review is to give all behavioral social scientists interested in social norms an overview over the topics and theoretical approaches devised in economics to tackle the issue of social norms.[36] This will, hopefully, lead to increased communication, critique, and cooperation between the different branches of the social sciences. This review, naturally, cannot deal with all technical aspects and details of the presented models. The focus is instead on the modeling decisions regarding social norms and what the model contributes to our understanding about social norms. In order to make the review accessible to all behavioral scientists, the different models are organized according to the main issue with regard to social norms. This allows a quick overview over the topics in which economists have introduced and dealt with the consequences of social norms in their models. The selection of contributions reflects particularly illustrative examples of the modeling choices in each category.

We highlight which aspects of social norms have been opened to economic modeling and research. How do economists incorporate social norms into their theoretical models? How has the understanding of social norms as a concept of societal coexistence improved and which aspects of economic understanding have been broadened by employing a social norm approach? By choosing to focus primarily on the formal models proposed by economists, we threaten to leave out important aspects of modern economic research with regard to social norms, namely experimental economics and field studies. Since these were the initial drivers of behavioral economic research, we would be remiss to ignore their contribution in this review. In order to keep the focus of this review, we will make use of empirical findings to buttress theoretical models and to check their prognoses where possible. We also discuss theoretical models that inform, ground, and motivate empirical research agendas.

The paper is organized as follows. In section V.2 a working definition of social norms in economics is developed. We discuss the similarities as well as the contentious points to arrive at our working definition (V.2.1). We continue in subsection V.2.2 with differentiating social norm analysis from two other proposed social motivators for human behavior: social preferences and identity economics. We find that social norms are essential to give meaning to social preference analysis. Social norm analysis thus guides social preference analysis. With regard to identity economics, we find that social norm analysis is mainly focused on investigating the effect of one pertinent so-

[36] In light of this goal, we decided to keep the mathematical notation of the original papers in this review and will explain the respective variables for each model separately, instead of streamlining notation in this review. This should allow interested researchers improved access to orient themselves quickly when they consult the original work.

cial norm, whereas identity economics proposes to model the complexity of several sources of pressure on behavior at the same time. In practical analysis, however, the differences between the two approaches are reduced, since the identity analysis is often only focused on one pertinent aspect of a persons identity.

Section V.3 organizes the review of economic modeling of social norms in eleven categories. We start with reputation models in subsection V.3.1. A concern for reputation which people are supposed to have is one of the main modeling assumptions in the literature on social norms. The subsection on signaling (V.3.2) highlights the complexity of social interactions and how norms influence human behavior, but can also be used to pretend to belong to a certain group of people. The next category (V.3.3) describes threshold models, where the social norm prescribes a certain level of the condoned behavior. Together with reputation concerns, this is a central building block in theoretical models on social norms, as well as for empirical research. Subsection V.3.4 highlights the interaction of several norms at the same time. In subsection V.3.5 we focus on how economists model the interrelations of social norms and the society they are relevant in. Norms have different strengths and effects in different societies. The subsections on matching (V.3.6) and evolutionary game theory (V.3.7) illustrate models that focus on establishing how a social norm came about and how it can continue to exist. It will illustrate, that social norms are the outcome of many small social interactions. In a similar intellectual vein, but proposing a different explanation for the existence and change of social norms, the subsection on norm entrepreneurs (V.3.8) presents models that posit that influential actors can design and change norms in society. The role of history as well as the difficult prerequisites to be successful in the endeavor of norm entrepreneurship are presented here. In subsection V.3.9 the effect of norms on voluntary contribution as well as the possible crowding in and crowding out of voluntary contributions because of policy interventions are presented. The role of a norm to work hard for the decision of whether or not to look for job is discussed in subsection V.3.10. Finally the subsection on social distance (V.3.11) highlights that subcultural social norms can have strong effects on people if the benefits of interactions depend on whether or not one deals with a stranger or a socially close person.

In section V.4 we establish that most of the empirical literature on social norms employs a particularly simple model of social norms to motivate and ground their research agendas (V.4.1). In addition, we will highlight possible avenues for future economic research with regard to social norms by identifying points where empirical and theoretical approaches to social norms might find common interests in subsection V.4.2. Section V.5 concludes.

V.2 Preliminaries

Before starting our review, we develop and present our working definition of a social norm for this review in subsection V.2.1. It has been selected to be as inclusive of economic models as possible. In subsection V.2.2 we differentiate the social norm approach from two close theoretical competitors, social preferences and identity.

V.2.1 A definition of social norms

This subsection will not attempt to give a full overview over the definitions of social norms in the economic literature.[37] Rather, we will try to establish common traits that are found in the theoretical economic literature on norms, which will lead to our working definition that will guide the reader's understanding in the following sections of the paper.

Historically, economists have attributed the existence of social norms to different functions they fulfill in societies. One capability of social norms is that they can be simple solutions to equilibrium selection problems or can serve as default solutions to social dilemmas. In these cases, social norms can guide appropriate behavior and improve the social efficiency of the outcome (Ullmann-Margalit, [1977] 2015). Moreover, social norms are seen to guide behavior outside of equilibrium (Sugden, [1986] 2005); Bicchieri, 2006), thus giving orientation in difficult situations. It is also argued that social norms can be welfare enhancing, for example, by reducing the free rider problematic in the provision of a public good (Festré, 2010). However, the effect of a social norm need not always be beneficial for individuals or the society (Fehr and Gächter, 2000b; Alpman 2013; Abbink et al., 2017). In fact, one impulse for the modern engagement in economics with social norms was precisely to explain the existence of inefficient social norms (Akerlof, 1980).[38] Inefficient social norms are difficult to grasp from an economic perspective, where optimizing behavior should, in theory, erode them.

When it comes to defining social norms, there is a surprising level of agreement in the literature. Abstracting from issues of formulation, the consensus is that a social norm is a shared agreement by a large portion of the relevant group of people about the permissible, forbidden and obligatory actions and behaviors in a given situation. Individuals try to conform to these norms and disapprove of norm transgressions (Ostrom, 2000; Bicchieri 2006; Festré, 2010; Abbink et al., 2017; Krupka and Weber, 2013; Harris et al., 2015; d'Adda et al., 2020). For economic purposes this means that in ad-

[37] For an introduction cf. e.g. Elster (1989), Bicchieri (2006), Festré (2010), Eriksson (2015), Fehr and Schurtenberger (2018), or Kliemt (2020).

[38] For a recent empirical contribution concerned with a social norm hindering efficiency see Castro and Czura (2021) on health management.

dition to material payoffs people are assumed to care about social payoffs that depend on the adequate behavior with respect to the social norms. For modeling purposes it is usually only one social norm that is under investigation. This social norm prescribes certain behavior and proscribes other behavior. The individual, who for our purpose is assumed to be able to make free decisions, has to consider the social norms specifications when choosing their action. It is an important feature of social norms that they focus on adequate actions or behaviors instead of the outcomes that these actions generate. Norms, therefore, put constraints on the action set of an agent by pre- and proscribing certain actions (Elster, 1989; Bicchieri and Chavez, 2010; Krupka and Weber, 2013). The individual follows the social norm because conformity with the social norm may engender improved social utility, as other members of society condone and reward the selected action. At least, conformity should not diminish social utility, while norm transgression will lead to a loss in social utility for the individual, e.g. because they are punished (Krupka et al., 2017).

Furthermore, it is not disputed that social norms permeate society at all levels. However, not every social norm influences behavior all the time. Rather, social norms depend on the social situation to be applicable. Nevertheless, multiple social norms may still demand conformity at the same time, sometimes with contradictory prescriptions for behavior. In these cases, salience, framing, and focus become important factors to determine which behavior or action is adequate in a situation (Kahnemann, 1992; Kallgren et al., 2000; Bicchieri and Chavez, 2010; Farrow et al., 2017). This has lead recent empirical research to expend considerable effort to elicit the social norms of their test subjects that are applicable in the situation they present their test subjects with, before actually starting their investigation. (Vostroknutov, 2020).

One big point of contention in the economic literature on social norms is the question whether external sanctions are constitutive for social norms or not. This would exclude internalized social norms, where feelings of guilt, shame or pride would influence behavior. A majority of economists sustain that norms are maintained by credible sanctions and punishment in case of not complying with the action or behavior prescribed by the norm.[39] Experimental economics provides ample evidence that people punish norm transgression. Typically, these punishments are costly to enact for the punisher, which raises the economic question of why people do it. Punishment in a social context can range from a raised eyebrow to physical punishment, ostracism, exile, or worse (Fehr and Gächter, 2000a, 2000b, 2002; Fehr and Fischbacher, 2004; Hoff et al., 2011; Abbink et al., 2017). For example, Farrow et al. (2017) insist on external sanctions as hallmark for the working of a social norm. They refer to internal control of behavior as personal norms. Alpmann (2013) also emphasizes that a social norm depends on the observation of behavior by others, whereas internalized norms, that

[39] Cf. e.g. Voss (2001).

trigger emotions like guilt or shame, should be referred to as moral norms.

However, the view that external punishment is constitutive for social norms is not without opposition. For instance, evolutionary game theorists uphold that several support mechanisms can be thought of to maintain social norms which do not require external punishment. For example, failing to coordinate actions with others in simple coordination games reduces the payoff automatically without requiring additional punishment. Social norms as signals of accepted behavior and reference points may help orient an individual without direct threat of external punishment. The argument is not that external punishment is not an important factor in maintaining social norms, but that other factors exist and may be able to maintain social norms without external punishment. The important characteristic from an evolutionary economics' perspective that maintains norms is not external punishment, but that there is positive feedback between norm prescribed behavior and individual actions (Young, 2015). Another reason why evolutionary economists embrace the possibility of maintenance of social norms free of external punishment is that their models employ some form of learning mechanism for norm transmission from one generation to the next. These norms are often assumed to be internalized. This broader view of social norms is moreover buttressed by a sociologically inspired perspective, which indicates that social norms can be and often are internalized, mainly through socialization processes. From this position emotions of guilt, shame or pride can be enough to maintain norm adherence. Internalized norms are learned behavior of how one ought to behave in a society, irrespectively of whether there is an observer who might administer punishment (Elster 1989; Gintis, 2003; Burke and Young, 2011). These emotions[40] have been interpreted as internal sanctions (Biel and Thøgersen, 2007), which would deliver new support to the argument that sanctions are constitutive for social norms but not necessarily administered from the outside. There has been no agreement on the issue of external punishment as being constitutive for a norm to be considered social.

Before we proceed to define social norms for this review, there is an important distinction between two types of norms, especially in empirical research, we will frequently make reference to. A social norm can either be descriptive, i.e. describing how most others behave and encouraging to behave accordingly, or injunctive, i.e. pointing out what one ought to do, i.e. an absolute rule (Cialdini et al., 1990; Goldstein et al., 2008; Burks and Krupka, 2012; Bicchieri et al., 2021).

This paper will embrace the broader definition of social norms, including internalized norms, for two reasons. First, it allows for most of the evolutionary game theory literature on social norms to be easily included. This strain of literature has contributed some of the most intriguing and important insights into the generation and

[40] Interestingly, the economics and law literature considers the origin of social norms to lie in emotions of esteem, shame and guilt, cf. McAdams and Rasmusen (2007).

characteristics of social norms from an economic perspective. Second, the sociologically inspired argument that an internalized norm can still be a social norm because it has to be learned in a society to be internalized, either directly through socialization by parents or exchange with peers or indirectly through opaque societal influences, is convincing. Such a norm, of what one ought to do, is decidedly produced in a social environment and hence is a social construct. Adequate behavior according to these internalized norms is determined by societal exchange and not easily changed at a whim of the individual. These norms influence behavior and, from a theoretical point of view, both aspects of conformity are interesting, whether transgressions are punished externally, or whether internal emotions that reference expectations of others about adequate behavior are the deciding factors that affect conduct and choice of action.

A social norm for the purpose of this review is thus an externally generated rule of behavior shared in a group of people. It is sustained with external sanctions and emotions of guilt and shame. Norms reduce people's choice sets. They may want to choose a different action due to selfish reasons but have to weigh up the pros and cons of such a norm transgression (Festré, 2010). In the following, usage of the word norm will refer to this definition of social norms unless explicitly stated otherwise.

V.2.2 Delimitations of social norms

Here, we discuss the relations between social norm analysis and two close competitors to explain prosocial behavior: social preferences and identity economics.[41] We start with social preferences. While some researchers use the two concepts social norms and social preferences rather interchangeably[42], we will try and highlight where differences lie. Social preferences explain the existence of pro-social behavior because people do not only care about their own payoff but also about the payoff of others in an interaction. During the last twenty years, behavioral models of social preferences have come to prominence in economic analysis. The most prominent modelizations are Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Charness and Rabin (2002).[43] All three approaches aim at explaining experimental findings that run counter to selfish payoff maximization behavior, especially in ultimatum and dictator games where equal sharing of the payoff occurs relatively often. For example, Fehr and Schmidt are

[41] Other candidates for comparison might be reciprocity models, e.g. Falk and Fischbacher (2006), which presume that people reward kind and punish unkind behavior, models of guilt aversion, e.g. Hauge (2016), where one tries to avoid disappointing others' expectations or moral believes, or lying aversion, e.g. Chen et al. (2008), where the difference between one's promise and one's action can be cause for disutility. All of these approaches, however, can be interpreted as special subtypes of social preference analysis and are therefore not considered in more detail.

[42] E.g. Fershtman et al. (2012), p.132, unscrupulously refer to social preferences of inequity aversion as a norm of inequity aversion.

[43] For an introduction to these three models cf. Dhami (2016), chapter 6. For a critique cf. Binmore (2006) and Binmore and Shaked (2010).

concerned with inequity aversion. The player is concerned with their own material payoff but also cares, to some extent, about the payoff of another player relative to their own, i.e. whether it is higher or lower. This concern enters additively into the utility function in addition to one's own payoff.

The main difference between models of social preferences and models of social norms to explain prosocial behavior lies in the source of the prosocial behavior. With social preferences, the player is concerned with their own payoff and, in some form or another, the payoff of another player. The later aspect influences their behavior and can explain why the individual did not necessarily choose the action that would have maximized their own payoff. In models of social norms, on the other hand, the main reference for the players is whether or not they comply with the prescriptive behavior or avoid the proscribed behavior implied by the social norm that is salient in the situation. In case of social preferences, the change in utility of a player is due to the added value of the payoff of the other player, while the utility increase or decrease in case of social norms is whether or not the players adhered to the social norm. The actions taken by others or the payoffs of others are only of secondary concern in social norm specifications, if at all. Vice versa, models of social preferences are not concerned with individuals acting according to a perceived or desired social standard of behavior (Krueger et al., 2008; Krupka et al., 2017; d'Adda et al., 2020; Chang et al., 2019; Vostroknutov, 2020).[44]

There is a discussion, especially in empirical research, about the applicability of the two concepts. Kimbrough and Vostroknutov (2016), for example, favor a social norm analysis over a social preference analysis. They claim that calculating payoffs is not the reason for prosocial behavior. Instead people are supposed to have a propensity to follow social norms. This desire to follow norms leads people to behave in such a way that social preferences may seem like the best explanation for their actions. Kimbrough and Vostroknutov (2016) continue with the claim that all social preference models implicitly assume the existence and functioning of social norms, e.g. social preferences for fairness require a social norm that stipulates what a fair share is in the specific situation. Fehr and Schurtenberger (2018) concur insofar as they argue that social preferences are the intrinsic motives of people, but in order for behavior to be considered as prosocial, it needs to be contextualized by a social norm. Whether models of social preferences or models of social norms are better equipped to explain the experimental data is not conclusively resolved. For example, Gächter et al. (2013) find that in their setup with peer effects the Fehr and Schmidt inequity aversion model is better suited to explain the data than a social norm approach. On the contrary, Chang et al. (2019) report that in a politically framed dictator game their social norms model explains the

[44] Models of guilt or lying aversion, on the other hand, depend on the beliefs about adequate behavior of the interacting parties. In contrast, social norm analysis depends on the beliefs of everybody, regardless of whether they interact with each other, cf. Krupka et al. (2017).

data better than a social preferences model à la Charness and Rabin (2002).

As a last difference between social norm and social preference approaches discussed in the literature is context sensitivity and framing. Wilson (2010) claims that social preference approaches are not flexible enough to account for a change in circumstances. Social preferences that explained behavior well in one set of conditions may be inadequate in another, since in order for them to be applicable, they depend on social norms to first define the situation. In order to know which social preference concept is suited to analyze the observed behavior, it has to be determined first whether the situation calls for reciprocal, fair or equitable behavior, for example. Only after this deeper meaning of the context is understood, can the respective models be applied. Social preference models which do not account for this underlying social prerequisite can thus sometimes serendipitously explain behavior, i.e. when coincidentally the correct model of social preferences is used to analyze an adequate social interaction, or they cannot if the model does not coincide with the salient social norm. Social norm approaches, on the other hand, can be applied to different circumstances and can explain why people sometimes act as if they have prosocial preferences and sometimes act selfishly (Fershtman et al., 2012; Chang et al., 2019; Vostroknutov, 2020), exactly because they are focused on the specific norm that guides behavior in this situation. Social norm approaches, however, have to be careful not to become arbitrary and explain behavior by making ad hoc reference to some social norm that is supposedly explaining the observed behavior, because this could explain any behavior as compliant with a social norm.[45] As alluded to above, in reaction to this problem, norm sensitization and norm elicitation procedures have gained traction as a necessary additional step in empirical work. For theoretical work, this implies that argumentative recourse to results of other social sciences on social norms has to be taken in order to ground the social norm approach (Postlewaite, 2011; Rege, 2004). As a summary, social norms focus on actions taken, not the utility outcomes these actions generate. They can be viewed as the social background for social preferences that lent these preferences sense in the first place. Additionally, the social norm approach is supposedly more flexible in being applied to different situations because different norms can apply, which can explain behavior that appears inconsistent from a social preference perspective. With preferences being understood to be more rigid in economics, they may fail to be sufficiently sensible to the context.

The second major strain of research that is closely related to social norms is identity economics. This idea was lastly advanced by Akerlof and Kranton (2000, 2005). The basic idea is that people gain utility if their actions are in line with the prescriptions and proscriptions their identity implies. Identity can then serve as a non-monetary source of motivation and may explain why some people choose actions or behave in a way that

[45] Cf. Cole et al. (1992); Postlewaite (1998).

is inconceivable for people with a different identity. The reason for such an action, that may be self-detrimental from an outside perspective, is to maintain one's self-image and to avoid negative emotions or cognitive dissonance.[46] Identity economics aspires to give room to the fact that an identity is a complex system that prescribes adequate behavior. This is in contrast to social norm analysis, where usually only the impact of a single salient norm is under investigation as reason for behavior that is not compatible with selfish utility maximization.

From a modeling perspective, identity economic research extends the utility function with the introduction of social categories. A person cares about their material payoff and would like to choose the action that gives them the best payoff. Moreover, each person belongs to a number of social categories and has a conception of the categories everybody else belongs to. To some extent, people may have discretion over the social categories they belong to but most are determined through socialization and origin. Each social category prescribes appropriate behavior for specific situations. The identity of an individual is determined by the categories they are assigned to and the (exogenous) social status that these categories confer. Furthermore, a person's self-image depends on taking actions as close as possible to the prescribed norm of the action in this social category. This may create an internal conflict. Deviating from the norm to pursue material interests leads to cognitive dissonance. Additionally, the actions of others constitute an externality on the individual. If others exhibit behavior that confirms their assigned social category, this confirms and improves the self-image of the individual. If they do not follow the appropriate behavior according to their social category, especially when they belong to the same category as the individual under investigation, the latter may loose some identity, since the categories become blurry. In summary, a person's total utility consists of two parts, namely their material payoff and the utility due to identity concerns. Total utility depends on the person's own actions and the actions of others. Unfortunately, this complex description of the human condition is not kept up stringently in the technical modelization and application. There, it is usually only one social category and its respective prescribed behavior that is addressed.[47] This reduction in modeling complexity gives away one of the model's assets.

[46] The psychological theory of cognitive dissonance was introduced and adapted to economics by Akerlof and Dickens (1982). According to them, the basic premise of cognitive dissonance is that people dislike having contradictory descriptions of themselves or their character. They prefer to have a consistent, positive self-image and are willing to adapt or manipulate contradictory information about themselves or their environment to ensure a consistent belief system. Konow (2000) postulated that not complying with social norms can cause cognitive dissonance, since the individual has to bear the psychological costs of not having done what one ought to do. Spiekermann and Weiss (2016) present an interesting cognitive-dissonance model, where people try to avoid having to follow a norm by exploiting the fact that information about the applicability of the norm is noisy. Since the primary concern of their paper is modeling a strategy of norm avoidance, its relevance for the modeling of social norms is limited.

[47] This reduction in complexity to only one dimension can also be seen in applied research building on the ideas of Akerlof and Kranton (2000, 2005). See e.g. Goette et al. (2006) for group membership in the Swiss army, Benjamin et al. (2010), who prime subjects on an aspect of their racial identity, and find

The strength of the identity approach is highlighting the complexity of the interplay of social categories and the possible sources for behavior modification, combined with framing and situational effects.

Although this short summary cannot do Akerlof and Kranton (2000, 2005) and their subtle and sophisticated social, moral, and psychological thoughts justice, it is sufficient to compare it to social norm analysis. If employed with only one dimension of the identity in focus, the formal differences with social norm analysis are greatly reduced. It boils down to the conflict between the prescribed ideal behavior of the social category and an individual's preferred but deviating action. Nevertheless, since others' actions also affect the individuals material payoff, identity economics is still adding an externality aspect that social norm analysis is lacking. Recently, the increased output of articles with reference to the identity model may illustrate that the concept has advantages in comparison to social preference and social norm analyses for explaining human behavior.

In conclusion, while social norm analysis may provide the necessary context to render social preference analysis tenable, identity analysis promises to incorporate the interplay of different social norms of different social categories. Depending on the social interaction under investigation, all three forms of analysis have their merits, and their relations should be kept in mind. In the following, we will turn our attention to the modeling of social norms in the economic literature.

V.3 Economic modeling of social norms

In the following, we will illustrate and categorize the theoretical modeling of social norms. The organization is based on the main aspect with respect to social norms presented in the respective papers. This categorization attempt is necessarily imperfect. For example, concern for reputation when considering norm transgression will be important for models outside of subsection V.3.1 and signaling of type will be an important aspect outside of subsection V.3.2, e.g. in subsection V.3.8 on norm entrepreneurs. Despite these overlaps, we find our categorization useful to highlight which aspects relevant to economists have so far been considered and investigated with the help of the concept of social norms and how the concept has been adapted in line with the research interest. We start our review with the work of Akerlof (1980), which is still referred to in the literature as a cornerstone of theoretical economic engagement with social norms.

that the salience of identity influences economic preferences, Barr et al. (2018) on differences on the acceptance of discrimination based on different identities, Chang et al. (2019) with US political identity as either democrat or republican, Dahl et al. (2020) and the educational opportunities of immigrant girls, or Grossman and Helpman (2021) on group benefits of trade policy.

V.3.1 Norms and reputation

Akerlof (1980) proposes the idea that people may care about the reputation they enjoy among their peers. Therefore, it might be costly to go against the socially expected behavior, even if an action that goes against this socially sanctioned behavior would leave the individual economically better off. The individual's utility therefore depends on their material and reputational payoff. The lever for conforming to the socially sanctioned behavior is that a violator of the norm[48] is collectively punished by society. According to Akerlof (1980), this can explain the continued existence of inefficient norms.[49] In Akerlof's (1980) model, an established social norm will continue to exist if transgression leads to a sufficient loss in reputation and the costs of disobedience are sufficiently high. Akerlof (1980) deviates from standard economic modeling of his time by making part of the utility non-individualistic, i.e. dependent on what others do and expect.

Technically, Akerlof (1980) extends an Arrow-Debreu general equilibrium model[50] by making people's utility partially dependent on their reputation, which depends on a social norm. A share of society μ believes in a certain norm. In addition to consumption goods, people care about their reputation, i.e. how they are perceived in their community. Moreover, the utility of the share μ of society that believes in the norm also depends on whether their actions are in line with the norm. The utility function takes the general form:

$$U = U(G, R, A, d^C, \epsilon), \tag{96}$$

where G is a vector for consumption goods, R represents reputation, A is a dummy variable for whether the individual obeys or disobeys the norm, d^C is a dummy variable representing the belief or disbelief in the norm, and ϵ are personal tastes. Personal tastes measure how much value the laborer puts on reputation, relative to income. The effect of reputation on an individual's utility depends on whether they obey the norm and on the size of the part of the community, who believe in the norm, μ. With a larger μ, disobeying the norm leads to a larger loss in reputation,

$$R = R(A, \mu). \tag{97}$$

Furthermore, disobeying the norm weakens the believe in the norm in society. Over

[48] Akerlof (1980) refers to social customs. The concept has been applied to norms and is well within our definition of social norms.

[49] The puzzle of the persistence of inefficient norms had been touched upon before, cf. e.g. Arrow (1972).

[50] Cf. Mas-Colell et al. (1995), chapter 19.

time, behavior may erode or strengthen the norm according to

$$\dot{\mu} = g(\mu, x), \tag{98}$$

where x is the fraction of society that actually obeys the norm, compared to μ, the share who believe in the norm. If $\mu > x$, then $g < 0$; if $\mu < x$, then $g > 0$. In plain words, if the number of believers is smaller than the number of practitioners of the norm (i.e. some believers disobey and carry the penalty for doing so), the norm erodes, while it is strengthened in the case where non-believers follow the established rule. Akerlof (1980) continues to analyze his extended Walrasian model for the short run (with a fixed μ) and the long run (where μ can change). In both cases, he finds that the norm may prevent some exchanges that would be carried out in the standard model, either because the people actually believe in the norm and want to adhere to its prescriptions, or because people, who do not believe in the code, may not carry out trades they find economically attractive, due to the fear of loss of reputation. Norms may therefore constrain economic exchange. In the short run, there are two possible equilibria, one where almost everybody and one where almost no one believes and obeys the norm.

Akerlof (1980) illustrates his ideas with an example economy. A fair wage norm keeps outsiders with lower wage demands than insiders from becoming employed if their successful employment requires some training that the employed insiders have to give to outsiders and withhold. The norm is then to not help newly employed people in order to keep wages for insiders high. This may explain involuntary unemployment and thus constitutes an inefficient norm. We highlight important aspects of the modeling to help improve the reader's grasp on the ideas and to discern how exactly a norm enters the model and how it is employed for economic purposes. There are two types of agents with different utility functions: Laborers, each with an initial endowment of one unit of labor, and capitalists, each endowed with one unit of capital. The ratio of capital to labor is unity. Laborers want to consume capital. Their utility function is

$$U = a_L + b_L K + c_L \epsilon R - d^R d^C \bar{C}, \tag{99}$$

where $a_L \in (-\infty, \infty)$, $b_L \in (0, \infty)$, $c_L \in (0, \infty)$ are parameters of the general equilibrium model and the subscript L signifies labor. d^R is a dummy variable (zero if the laborer obeys the norm, and one if they disobey), d^C is a dummy variable (zero if the laborer does not believe in the norm, and one if they believe), K is the laborer's final allocation of capital, R is the laborer's reputation, and $\bar{C}$ is the loss of utility if a believer breaks the norm. Consequently, a person who does not break the norm does not lose utility. If a believer breaks the norm ($d^R = d^C = 1$), they lose utility $\bar{C}$. A non-believer who breaks the norm ($d^R = 1$, $d^C = 0$) does not lose utility due to $-d^R d^C \bar{C}$. However, everybody loses utility due to $c_L \epsilon R$, because Akerlof (1980) assumes a particular reputation function of the following form

$$R = -d^R \mu \bar{R}, \tag{100}$$

where $\bar{R}$ is a positive constant. Therefore, every person who breaks the norm ($d^R = 1$) loses utility in form of loss of reputation. How severely this affects them, depends on the proportion μ in society, who believe in the norm, i.e. the strength of the norm, cf. (100). In this illustrative model, Akerlof (1980) assumes that tastes ϵ are uniformly distributed. On the other side of the market, the capitalists wish to consume labor. Their utility function is

$$U = a_K + b_K L + c_K R, \tag{101}$$

where $a_K \in (-\infty, \infty)$, $b_K \in (0, \infty)$, $c_K \in (0, \infty)$ are the parameters of the general equilibrium model and the subscript K signifies capital. L is the final allocation of labor and R is reputation again. The norm in this economy is that trade of capital for labor should only occur at a certain exchange rate $\bar{\omega} > 1$ (with capital as numeraire, $\bar{\omega}$ is a wage level). Note that the fair wage norm is relevant for laborers and capitalists alike. The last piece of the model is the change in the belief in the norm, which takes the particular form

$$\dot{\mu} = \beta(x - \mu), \tag{102}$$

where β is a positive constant and x is the portion of people who obey the code. For $x > \mu$, the portion of believers increases; and for $x < \mu$, it decreases.

Akerlofs (1980) assumptions concerning the market structure do not deviate from a standard Arrow-Debreu specification, except for his addition of reputation concerns with respect to norm compliance in the utility functions. Since $\bar{\omega}$ represents a fair wage for laborers, supply for labor at $\bar{\omega}$ will often exceed demand and, therefore, some agents will be rationed. In the short-run, μ is fixed. Rewriting the two utility functions with respect to the wage ω, notional (unconstrained) demand for labor is found by maximizing the capitalists' utility function with respect to ω. There exists a threshold wage ω_{thresh} at which capitalists are indifferent between trading at $\bar{\omega}$ and upholding the norm, and trading at ω that solves the capitalists' maximization problem, but breaks the norm and reduces reputation. If ω is above the threshold, the capitalists will prefer to trade at $\bar{\omega}$. For values below the threshold, the capitalists will prefer to break the code and trade at ω. Similarly, laborers prefer to honor the norm and trade at $\bar{\omega}$, if their utility of doing so is larger than their utility of trading at ω, breaking the norm and incurring the damage to their reputation. The laborers' inequality can be solved for the laborers' tastes ϵ. There exists a critical value ϵ_{thresh}. For ϵ above this threshold, laborers obey the fair wage norm. For lower levels of tastes, the laborers break the norm. These are the relevant modeling choices for our purpose here. In the long-run equilibrium, μ, the portion of agents who believe in the norm, must equal the fraction who obey the norm for the norm to continue to exist. Akerlof (1980) proves that such long-run equilibria, where a norm does not erode, may exist in his setup. According to him, this is an explanation why involuntary unemployment in a market setting can occur.

Akerlof (1980) models several important features that will guide theoretical research with regards to norms in the following. First, he highlights the strength of a norm as important for its impact on economic decisions. The strength of the norm is measured by the portion μ of society who believe in the norm. Second, he assumes that people have a taste for a reputation as norm abiding individuals. There are positive externalities of norm following: The more people believe and follow the norm, the more efficient are the sanctions in the form of loss of reputation (Festré, 2010). Third, Akerlof (1980), at least in his illustrative formulation, additively plugs the taste for reputation and the possible punishment for violating the norm into the agent's utility function, thus setting these concerns for the norm on equal footing to the agent's concern for consumption. Fourth, he allows the agents to have different types with different tastes for norm adherence, ϵ. Fifth, Akerlof (1980) is concerned with changes in the strength of the norm over time and whether economically inefficient norms will be eroded or whether they can maintain themselves. Sixth, he shows that, under certain circumstances, many different levels of adherence to the norm are possible equilibria of the game. He thus highlights the fact that different levels of norm adherence may exist in different societies. He adds verbally that different norms may exist and be effective at the same time in one society, but does not include this consideration into his model. Finally, Akerlof (1980) is mainly concerned with the effects an existing social norm can have on individual decision making. The origin of the norm is not relevant. This will be a recurring feature of many social norm analyses outside the evolutionary game theory literature. The focus will be on illustrating the effects of an existing social norm, without concern for why this norm came about.

One fair critique leveled at Akerlof's (1980) formulation is his use of dummy variables. The smallest transgression in his models leads to losing all reputation. One can argue that there should be a range instead of a binary option (Festré, 2010). Romer (1984) argues that the punishment should usually fit the crime. The strength of the infraction of the prescribed behavior should influence the severity of the punishment. Romer (1984) thus argues for the introduction of a measurement for the degree of the transgression and a respective degree of punishment, instead of a complete loss of reputation. Romer (1984) adds that even with norm breaking at the margin, inefficient social norms may still exist.

We next look at some of the development the model of Akerlof (1980) underwent. Corneo (1995) is interested in why voluntary union membership exists, since unions without mandatory membership face a free rider problem. He follows Akerlof's (1980) model and considers union membership to be a norm for laborers.[51] His modification

[51] For earlier contributions of applying Akerlof's (1980) model to norms for union membership cf.

of the laborers' utility function consists of the ability of employers to resist union formation by offering a financial bonus δ to workers who do not unionize and therefore break the norm. Furthermore, the direct negative effect for a believer in the norm, to breaking the norm ($\bar{C}$ in equation (99)), is replaced with the costs of union membership. This, however, reverses the effect and makes norm breaking less painful. Moreover, there is now only one dummy variable d. Either a laborer joins the union ($d = 1$) or they do not. If they do not and break the norm, they receive the bonus and have no costs of union membership. Additionally, Corneo (1995) does not analyze the effects in a general equilibrium model but models a bargaining game in one firm. Corneo (1995) employs Akerlof's (1980) norm ideas in a non-overlapping generation model with an infinitely lived economy, where laborers live one period. The technically simple analysis reveals (with some additional assumptions) that a long-run equilibrium (the number of believers and practitioners of the norm are equal in each period) with a high union membership and few free riders can exist, with wages above the workers' reservation utility. However, a stable long-run equilibrium also exists, where no unionization takes place, thus replicating Akerlof's (1980) result of several equilibria. Which equilibrium obtains, depends on the starting value of belief in the norm at time $t = 0$, i.e. μ_0. If it exceeds a critical value, the strong union membership equilibrium follows and vice versa. This makes the idea explicit that different norms can result from different starting circumstances, which was merely verbally stated in Akerlof (1980). The firm's opposition to unionization increases the threshold and thus makes unionization less likely. Interestingly, higher union bargaining power in this model weakens the norm of union membership, since a stronger union increases the gains from wage negotiations, which increases employers' opposition, which reduces union membership. In summary, the sophistication in Corneo's (1995) application is reduced compared to Akerlof (1980). There are less interacting variables, the belief in the norm by the worker is not important anymore, since the originally negative direct effect of norm transgression is actually another bonus of saving on union membership, and instead of a general equilibrium background, the equilibrium analysis is partial and the main concern is only the worker's side of the market. These simplifications allow for a streamlined application, but at the expense of some of the interactive flavor of Akerlof (1980).[52]

A second influential early contribution to the economic literature on social norms is a model by Bernheim (1994). He follows Akerlof (1980) insofar as he assumes, too, that people value reputation in addition to consumption. He, however, introduces a continuous reduction in reputation.[53] This punishment for norm transgression is en-

Booth (1985) and Naylor (1989).

[52] Some later contributions, cf. e.g. Lai et al. (2003), have reduced the complexity of Akerlof's (1980) model further.

[53] Bernheim (1994) denominates his concept 'status', but it is theoretically equivalent to Akerlof's (1980) reputation.

dogenously produced in the model. Agents have different optimal consumption levels according to their type. An agent's reputation depends on their type. Since the type cannot be observed, however, it has to be inferred from the agent's actions. Without concern for reputation, the agent would choose their optimal consumption level, but their choice of consumption influences their perceived reputation. When reputation is relatively important compared to consumption, many of the heterogeneous types with different optimal consumption decisions conform to a strict social norm of consumption. Only extreme types do not conform. Formally, the utility of an agent is

$$U(x,t,\phi) = g(x-t) + \lambda \int_T h(b)\phi(b,x)db, \tag{103}$$

where x is the consumption decision, and t is the type of the agent or, alternatively, their optimal consumption level. The function g is twice continuously differentiable, strictly concave, symmetric, and reaches its maximum at $x = t$, i.e. if the actual consumption choice equals the optimal consumption level for this type. The parameter λ measures how important reputation is for the agent. Without going into full detail, the integral is the contribution of the inferred reputation to the agent's utility. The set of types is T, $h(b)$ is a function that measures the reputation an agent of inferred type b has, with the adequate mathematical characteristics, and $\phi(b,x)$ is the inference function, which assigns a probability to each possible inference b, given an observed action x. This influential model has contributed three key aspects to the modeling of social norms. First, the effect of reputation is not discontinuously changing with a transgression, but the reputation function $h(b)$ is assumed to be a continuous function. Second, a norm sensitivity parameter λ of how important reputation is for the agent is introduced. Third, the interaction of reputation concern and action choice is modeled as a signaling problem. The punishment for transgression leads to high levels of conformism in Bernheim's (1994) model. The norm of conformism is therefore not assumed but rather the result of the added reputation concerns.

The reputation models of Akerlof (1980) and Bernheim (1994) continue to be important references for the social norm literature. They have introduced important features that are mainstays of research. Contributions employing their concepts have not always lived up to their standard but have demonstrated the flexibility of the approach. A norm in Akerlof (1980) prescribes which actions are norm compliant and which are not. The norm is assumed to exist. For Bernheim (1994) a norm is a threshold of behavior, that the individual aims to live up to. It is endogenously determined. Both assume that people have an intrinsic desire to have a good reputation. This is how they explain behavior that is not compatible with selfish utility maximization. While Akerlof (1980) advises, that there are several possible equilibria, i.e. several possible levels of norm strength, Bernheim (1994) illustrates that if reputation is sufficiently important, one strict norm will exist. Both suggest that in order to maintain a norm, external punishments are necessary and equate the strength of the norm with the size of the external punishment.

V.3.2 Signaling

In the reputation model above social norms reduced the utility of agents who did not follow the prescribed behavior. Another aspect of social norms is that they can be used by individuals to display their type, i.e. their social affiliation with a specific social subgroup. Signaling is, however, not always a simple procedure in a complex information structure when transmission is noisy. We present the model by Bénabou and Tirole (2006) which deals with several interacting parameters relevant for social norm analysis. We discuss the implications of this model and other signaling approaches in the following.

Bénabou and Tirole (2006) aim at combining three motivators for people's behavior in one model. First, people respond positively to material incentives, e.g. in the work place, but crowding out is also observed, e.g. especially with voluntary contributions.[54] Second, people want to be seen as good members of society by adhering to social norms that prescribe prosocial behavior, i.e. people have reputation concerns. Finally, people want to avoid cognitive dissonance and want to maintain a positive self-image. These three factors, extrinsic material incentives, reputation concerns, and intrinsic social image concerns, interact in Bénabou and Tirole (2006). Actions may be signals to prove one's norm conformity, but monetary incentives and social image concerns may interfere with clear messaging.

The agents in the model can decide to participate in contributing to a positive social action, e.g. helping to finance a public good. Their participation level a can either be selected from a discrete (e.g. contributing or not) or a continuous (e.g. time used for the positive social action) choice set $A \subset \mathbb{R}$. The costs for a are $C(a)$. Additionally, a entails a material reward, ya, with $y \lesseqgtr 0$, allowing to model subsidies or taxes on the action. Agents differ with respect to how much they intrinsically value supporting the public good, v_a, and their intrinsic valuation for money or consumer goods, v_y. Summarizing, an agent who contributes a will benefit according to

$$(v_a + v_y y)a - C(a) \tag{104}$$

The type of an agent $\mathbf{v} \equiv (v_a, v_y) \in \mathbb{R}^2$ is private information and drawn from a continuous distribution with density $f(\mathbf{v})$ and mean $(\bar{v}_a, \bar{v}_y)$. The reputation of an agent

[54] Crowding out may occur when initially intrinsically motivated actions carried out for no financial gain are instead rewarded with extrinsic monetary incentives. This may reduce the provision of the intrinsically motivated actions, due to reduced intrinsic motivation. Classical examples are donating blood and picking children up from kindergarten on time, cf. e.g. Bénabou and Tirole (2003).

depends linearly on the posterior expectations of the type $\mathbf{v}$. The value of reputation for selecting a and receiving y is given by

$$R(a, y) \equiv x[\gamma_a E(v_a|a, y) - \gamma_y E(v_y|a, y)] \tag{105}$$

with $\gamma_a \geq 0$ and $\gamma_y \geq 0$. γ_a illustrates that people want to be seen as working for the common good and γ_y represents that people want to avoid being seen as greedy. The variable $x > 0$ measures the visibility or salience of action a, e.g. the likelihood of being observed or the duration of public records of the action chosen. With the definitions of the agents' reputation concerns $\mu_a \equiv x\gamma_a$ and $\mu_y \equiv x\gamma_y$, Bénabou and Tirole (2006) have their main preference and information specification in place. Agents' reputation concern $\boldsymbol{\mu}$ can be either the same for all agents and thus common knowledge or private information of heterogeneous agents. An agent of type $\mathbf{v} \equiv (v_a, v_y)$ and reputation concerns $\boldsymbol{\mu} \equiv (\mu_a, \mu_y)$ has to solve

$$\max_{a \in A} \left[(v_a + v_y y)a - C(a) + \mu_a E(v_a|a, y) - \mu_y E(v_y|a, y) \right]. \tag{106}$$

That is, an agent values their contribution to the public good as well as their material payoff from their contribution to the public good. Furthermore, they care about being seen as good members of society and avoiding being seen as greedy or materially motivated. Assuming a well-behaved decision problem and a continuous choice set, rearranging the agent's first-order condition with respect to a leads to

$$C'(a) = v_a + v_y y + \mu_a \frac{\partial E(v_a|a, y)}{\partial a} - \mu_y \frac{\partial E(v_y|a, y)}{\partial a}. \tag{107}$$

From (107) one can see that knowing a is enough to infer the sum of the agent's motivational factors: intrinsic, extrinsic and reputational. Since these are individual characteristics, the problem becomes extracting the values of v_a and v_y from the signal a. Choosing a, the agent reveals their sum of valuations to be $v_a + v_y y = C'(a) - \mu_a \frac{\partial E(v_a|a,y)}{\partial a} + \mu_y \frac{\partial E(v_y|a,y)}{\partial a}$. Assuming a normal distribution of the valuations, the expected values of the valuations can be determined. There are several interesting interactions happening. For example, a higher incentive or reward for contributing y does indeed improve information about v_y but reduces information about v_a. The material effect of higher incentives is accompanied by reputational concerns. A higher y increases the utility from contributing for all agents directly but also increases the value of v_y and decreases the value of v_a. A higher y will thus attract some new, relatively greedy contributors and also drive off some agents with high values of v_a. The net effect can be ambiguous. If agents are heterogeneous with regard to the importance of reputation for them, i.e. have different $\boldsymbol{\mu}$, the noisiness of the signal a is increased and inference becomes more difficult. Finally, better observability, i.e. a larger x, increases the noisiness of a further, since their public image becomes more important to agents, which increases conformity.

Bénabou and Tirole (2006) continue their analysis by showing that material rewards lead observers to conclude that the intrinsic motivation to contribute was not as important for choosing a. They next highlight crowding out effects due to the introduction of material incentives or legal sanctions. The former reduce reputational motivation, the latter social motivation. If there is little probability weight on greedy agents in the distribution, the reputation concerns refer to avoiding being seen as one of the bad ones. Contrarily, if only few genuinely socially oriented agents exist, one aims at imitating them. Finally, they analyze the effects of non-material rewards and punishments and find that public shaming or praising might not be an effective measure to better behavior in a society because observers will, for example, interpret good behavior as motivated by concerns about reputation and less because the person acted out of intrinsic motivation. In relation to norms, Bénabou and Tirole (2006) illustrate that social norms can endogenously emerge from the inferences agents make from the observed behavior of others. Equilibria are shown to exist with no participation, full participation or partial participation in contributing to the good. These equilibria are interpreted as social norms of contributing. The paper thus also contributes to the notion that several social norms are possible. Which one obtains is to some extent arbitrary. One main contribution of the work of Bénabou and Tirole (2006) is the spiritedness to allow for many possible types of agents with heterogeneous preferences for reputation, intrinsic motivation and self-interest. This makes for a very vivid model. Furthermore, the mechanic that people observe what others do and infer what others' motives might have been, is an important step for models of social norms. Norms are here seen as the outcome of human interaction under conditions of imperfect information and noisy signals. This illustrates that social interactions are not always as clearly structured as, for example, in Akerlof (1980), where everybody knew what action the norm prescribed.[55]

Bénabou and Tirole continue developing this line of work as they seek to model the interplay of values, laws, and norms.[56] In Bénabou and Tirole (2011), they present a simplified version of the general analysis above. Norms with social punishments (awarding either honor or stigma) develop because agents observe others' behavior

[55] The model by Bénabou and Tirole (2006) has been fruitfully employed in the empirical literature. Besley et al. (2019) use it with regard to tax evasion, where the prosocial norm consists of paying one's taxes. People do not want to be seen as evading taxes. The authors exploit a tax change in the UK in the 1990s and find that the reduced level of tax compliance that followed the change of the law, can be explained by a negative shock to the intrinsic norm of following the law. Bursztyn and Jensen (2017) give an overview of field experiments regarding reputation concerns and orient their empirical investigation with the model of Bénabou and Tirole (2006) as theoretical foundation.

[56] A very interesting recent direction is the investigation of narratives, how they disperse in a society and how they might influence behavior and stabilize norms, cf. Bénabou et al. (2018). Since the added contribution with relation to the modeling of norms is rather tenuous, we do not include this literature in this review.

and infer their motives. The preferences in society regarding which behavior is condoned and which is not, are only imperfectly known by the heterogeneous agents, adding a second element of uncertainty. The main advancement is the role played by law. A principal, who is better informed about the distribution of preferences in society, has to decide whether to punish or reward certain behaviors in society and to what extend. Agents' actions now signal their types to the principal and the principal signals the preferences of society back. The principal can use incentives to give an additional signal. For example, a lower subsidy for an action can credibly send the message that this is a prosocial action a that everybody, except for the worst kind of members of society, takes and therefore does not require strong incentives. Those who do not choose a in this case will suffer great reputational losses, due to stigma.

Continuing this line of thinking, Ali and Bénabou (2020) are mainly concerned with the adequate amount of privacy and publicity of information in a society. However, the basic model they employ is a simplified version of Bénabou and Tirole (2006). Again, people are concerned about their reputation and how others will behave with respect to them if they know what action they chose. They can either choose one that is condoned or one that is proscribed by a norm. Here, norms play an important role on a secondary plain for determining sanctions and rewards, which are the lever through which the model works. The model posits a single principal and a continuum of small agents, who face a binary choice to contribute to a public good or not. Every agent profits in three ways from contributing to the public good. First, they experience an intrinsic payoff from giving, second, they profit from the provision of the public good, and third, they enjoy benefits from contributing because they gain reputation. How much an agent values each form of utility is their private information. The principal can choose the degree of publicity, i.e. the variable of observability x in Bénabou and Tirole (2006), of each agent's contribution. The problem the principal faces is the following: More publicity will improve the contributions to the public good in an economical way. However, with more publicity, reputation concerns may become more important for the agents' contribution level than the actual utility gain the public good gives them. Additionally, higher publicity and thus more reputation concerns make the agents' behave in a more conformist manner, making it difficult for the principal to infer the agents' preferences for the public good from their behavior (in case the principal has only imperfect information about how much the agents care about their reputation or the public good). The framework developed in Bénabou and Tirole (2006) allows Ali and Bénabou (2020) to rather elegantly deal with multidimensional signaling and higher order beliefs of heterogeneous agents. The signaling of norms and preferences can then be employed to devise policies to improve behavior and consequently social welfare. The last two contributions thus illustrate the interaction of social norms with formal institutions of society.

There are other signaling approaches with reference to social norms. We present two of them. Sliwka (2007) models a simple signaling game with three types of agents, who work for a principal. Two types never change their behavior: one is selfish and only concerned with their own payoff; one is fair, who cares to some extent about the principal's payoff. The third group conform to the behavior of the group of agents (selfish or fair), which they think constitutes a majority in the company. These conformists thus follow the majority behavior, i.e. the social norm they perceive to be relevant, irrespective of the norms' content. The principal has to decide whether to trust or to control the work of the agent. It is beneficial to control a selfish agent, to ensure they work, but to trust a fair agent. An agent's type is private information, but the principal receives a (possibly noisy) signal about the share of fair types. The principal may now find it in their interest to send a signal of trust by not controlling the agent, communicating that the majority of agents in the firm are of the fair type. If the signal is credible, a conformist agent may behave like a fair agent. Sliwka (2007) shows that for a certain cut-off value for the portion of fair agents, a separating equilibrium for this game may exist, where the principal trusts the agent, if the signal indicates a high share of fair agents, and controls if the signal indicates a low share of agents. For example, a principal can signal trust by raising the base wage the agent receives for sure instead of increasing the incentives based on performance. The higher cost of raising the fixed wage can serve as a credible signal that a majority of the agents is, in fact, of the fair type, since these costs might be too high for a principal, who received a signal indicating the majority of agents being selfish. The contract design thus serves as signal about the beliefs of the principal about the composition of the work force and the norms at work, which are important to the conformist part of the employed staff. As such, there are only two possible norms in this model that are relevant for the conformists: behaving like a selfish agent or behaving like a fair agent. The model highlights that trust and distrust can reduce the uncertainty about a prevalent social norm in a population.

Bursztyn et al. (2020a) build a simple model with two types of agents, who have to take an action in front of an audience. There are two possible states of the world, each corresponding to a different distribution of the types. Each type has a preferred action, which they would choose in private or if they had no concern for the audience's opinion of them. However, each type cares about what each audience member believes about them. If the agent can make an audience member believe that they are of the same type as the audience member, the agent gets a positive payoff. The agents know their own type but they do not know the distribution of types in the audience. For a large enough audience, it can be optimal for an agent to take the action that is less beneficial for them, since the audience's approval of their actions overcompensates their loss in utility from not taking their type-specific action. An informative public signal about the distribution of the types in society changes the beliefs of the agents about the distribution of types in the audience. The informative signal also changes the audience's posterior belief about the agent's type. With an informative signal indicating one type being prevalent, an agent is (weakly) more likely to take an action that is preferred by that type, and (weakly) less likely to choose the other type's preferred action, compared to the situation without a signal. For the audience, a public signal indicating one type's prevalence means that an agent who chooses the action preferred by the relatively less prevalent type, is certainly of the relatively less prevalent type. However, agents who choose the action that is associated with the type whose prevalence the signal indicates, might be of the less prevalent type, and only their reputation concerns made them take the action associated with the other type. Thus, it is difficult for the audience to differentiate between agents who are genuinely of the same type as they are and imposters who conform to the perceived majority behavior, i.e. the norm. The contribution of Bursztyn et al. (2020a) explicitly models the tension when uncertainty about the relevant social norm exists. Additionally, they try to explain the rapid change of a norm. If a very strong public signal, like a landslide victory in an election for example, reveals that certain positions are popular in a society, beliefs can be updated rather quickly and people may try to conform to this revealed norm.

Signaling models can explain the existence of social norms due to reputation concerns. Additionally, external sanctions can be used, as in Ali and Bénabou (2020) for example, by an informed party to signal information. Moreover, signaling models highlight that social norms may be the outcome of many small social interactions of people carried out under imperfect information. The norms that obtain are the possible equilibria of the signaling game. The complexity with several informational imperfections can be handsomely modeled in the contributions. However, the setup is mainly concerned with detailing the ways to signal certain information. This information is usually used to signal norms or prevalent forms of behavior, which are conducive to behavioral change. Nevertheless, the models are applicable beyond social norms. Ellingsen and Johannesson (2008), for example, provide a theoretically close model to Bénabou and Tirole (2006), with esteem as the coveted utility enhancing characteristic. While close in their modeling decisions, they do not interpret the desire for social esteem as a social norm and distinguish themselves from social norm research. This highlights that the first priority of these models is addressing signaling, and that social norms enter only on a secondary plane. It could therefore be argued that these models are only concerned with the informational structures of the problems and not its contents, which would make them neutral with respect to social norms. Such an argument, however, cannot convince. As long as the informational structure investigated can be logically interpreted as representing social norms, the models contribute to their understanding. In this case, e.g. the interplay of social norms with other possible motivating factors, the possibilities of influencing behavior, making use of perceived norms in society to explain herd behavior, and to communicate credibly about societal dispositions.

V.3.3 Norms as a threshold

Norms as a threshold of behavior are a basic design choice that is used as a building block in many models. Threshold models are often employed in empirical research, due to the fact that they are particularly easy to design and allow to easily connect observed behavior to theory. We present a purely theoretical contribution in Huck et al. (2012) and add two examples of empirical contributions to this form of modeling.

Huck et al. (2012) consider social norms and their effect on team production. They find that the same social norm may have different consequences under different incentive schemes. It can either improve effort provision or diminish it, or have no effect at all. In their view, social norms develop out of a desire or pressure towards social efficiency. A social norm discourages behavior that causes a negative externality on others and inspires behavior that excites positive externalities.

In the model, each agent in a team of n agents chooses effort $x_i \geq 0$. All the efforts together constitute an effort profile $\vec{x} = (x_i, ..., x_n) \in R_n^+$. An agent's utility is the sum of their material, $(u_i(\cdot))$, and their social, $(v_i(\cdot))$, payoff

$$U_i(\vec{x}) = u_i(\vec{x}) + v(\vec{x}, \hat{x}^i). \tag{108}$$

The functions are twice differentiable. To define social norms, the authors first introduce social ideals. The subset of effort profiles that cannot be Pareto improved upon by a different profile describes the possible social ideals in the group of agents. Each agent has a specific social ideal of what constitutes the best course of action for the group, $\hat{x}^i$. Adhering to the social norm in this model means choosing the ideal effort level prescribed in $\hat{x}^i$.[57] The function $v_i(\cdot)$ is designed to capture two externalities that the effort choice of agent i causes. On the one hand, the externality that i's deviation from the socially ideal behavior has on the material utility of all other agents if they adhere to the norm and choose effort according to the social ideal of i, $\hat{x}^i_{-i}$. This externality is modeled as the sum of the effect for each of the other agents, j,

$$\psi_i(x_i, \hat{x}^i_{-i}) = \sum_{j \neq i} \left[u_j(x_i, \hat{x}^i_{-i}) - u_j(\hat{x}^i) \right]. \tag{109}$$

If agent i acts according to their social ideal, there is no externality, $\psi_i(\hat{x}^i) = 0$. The second externality describes the material effects on i if i acts according to their social ideal, for any action profile of the other agents

$$\psi_{-i}(\hat{x}^i_i, x_{-i}) = u_i(\hat{x}^i_i, x_{-i}) - u_i(\hat{x}^i), \tag{110}$$

[57] With a slight abuse of notation, the conforming action chosen by agent i will also be denominated by $\hat{x}^i$.

which is also zero at the social ideal $\psi_{-i}(\hat{x}^i_{-i}) = 0$. Both externalities can have positive as well as negative effects depending on who shirks or is overly zealous. Summarizing, the social utility function $v(\cdot)$ is a function of the two externalities,

$$v(\vec{x}, \hat{x}^i) = g^i \left[\psi_i(x_i, \hat{x}^i_{-i}), \psi_{-i}(\hat{x}^i_i, x_{-i}) \right], \tag{111}$$

with g^i twice differentiable and non-decreasing in the externality i has on the other players, i.e. the first argument. That means, if i exerts more effort than the norm prescribes, they improve their social utility by increasing the positive externality their higher effort has on others. The second argument represents the social utility effect on the agent due to the externality of the effort provision of others. The closer their effort is to the norm, the higher might be the social utility gain of higher own effort. Differentiating equation (108) with respect to the agent's effort, one can see that a change in effort has a direct effect on the material payoff and an indirect effect on the social utility of the agent, mediated by the consequences of the agent's effort choice for the interactions with others. The costs of higher effort may then be compensated by increased social utility

Huck et al. (2012) illustrate the consequences of their model for two different payment schemes with simple functional forms. With team pay, the social norm produces a positive externality. Higher effort of one agent benefits all others. Consequently, the exerted effort, compared to a benchmark of egoistic agents, is higher. Agents can work harder than their social ideal prescribes, but higher financial incentives may crowd out social norm incentives, potentially lowering efforts. With team pay, a firm thus always wants to hire norm sensitive agents. With relative performance pay, i.e. competition between the agents, one agent's effort may affect another agent's payoff negatively. Agents who care about the above social norm will then provide less effort. The firm may increase profits by employing selfish agents, who do not care about the norm. Thus, the form of the incentive contract influences the sign of the externalities and has therefore consequences for a firm's hiring processes and the composition of a firm's workforce. This can explain why the matching of firms and employees is usually not assumed to be random. Rather, specific types of agents want to work for a specific type of firm, which wants to hire exactly this type of agent. Moreover, Huck et al. (2012) identify that social norms can lead to multiple equilibria, but do not venture into equilibrium selection debates. Multiple equilibria may complicate the contract design further, as it is not a priori obvious that the norm will guide behavior to an attractive equilibrium. With regards to social norms, Huck et al. (2012) produce one of the more involved threshold models. The social norm consists in stipulating a certain level of effort that has to be exerted. The norm prescribes efficient social behavior, since fulfilling the ideal levels of effort will lead to a Pareto efficient outcome for the team. The different remuneration schemes highlight that norms are context sensitive. Their environment determines the sign and size of their effect.

A further paper that posits a social norm as a threshold is Kessler and Leider (2012). Their empirically focused research asks why contracts often do not prescribe expected behavior of the parties in detail but are oftentimes incomplete. They argue that agreements made on the spot generate norms specific to this contractual relationship, which guide behavior towards mutually, i.e. socially, beneficial actions. To orient their experimental design, they sketch a simple model of social norms. In this setup an individual gains utility from their monetary payoff. Moreover, the agents are norm sensitive, i.e. whenever they contravene the prescription of a social norm, they lose utility. Importantly, the relevant social norm for the parties is induced by the contract they agree on. This induced norm takes here the simplest form of a prescribed level of action $\hat{x}$. Individual i's utility function, if they do not adhere to the norm, i.e. if their input is lower than the norm, $x_i < \hat{x}$, is

$$U_i(x_i, x_j; \hat{x}) = \pi_i(x_i, x_j) - \gamma_i g(\hat{x} - x_i), \tag{112}$$

where π is the material payoff from the interaction, given the inputs of the contracting parties i and j. The increasing function g represents the disutility from disobedience, and γ_i indicates a level of individual norm sensitivity. With $\gamma = 0$, the individual is selfish; with $\gamma \to \infty$, the individual always obeys the norm. For $x_i \geq \hat{x}$, the individual's utility is just $U_i(x_i, x_j; \hat{x}) = \pi_i(x_i, x_j)$, i.e. if a partner does more than was agreed on, they only get the material payoff of the interaction. Cheating and providing less than agreed is therefore less attractive to norm sensitive players. Kessler and Leider's (2012) experimental results on handshake agreements to take the action corresponding to the first best solution suggest that contracts can indeed induce norms that significantly alter behavior towards being more social, especially in games with strategic complements. They interpret this finding as a possible reason why incomplete contracts are surprisingly efficient and commonplace. An induced high norm increases efficiency of the trade without having to rely on costly enforcement mechanisms. Social norms are here produced by the social interaction, giving the parties to the trade a guideline to follow.[58]

In the last threshold model, Abbott el al. (2013) look at motives for recycling. Their

[58] Bartling and Schmidt (2015) provide experimental evidence on contracts as reference points in a renegotiation game. Social norms, in their setup, can help explain why some sellers do not charge a mark up in the renegotiation stage of the game. However, they can identify this behavior only if the initial price stipulated in the initial contract was low. They close by emphasizing that different behavioral mechanisms could be at work and that future experimental designs must aim at disentangling their different effects. Iyer and Schoar (2015) report on their field studies with respect to hold-ups, contract renegotiations and social norms. They find that norms of fairness and reputation concerns can help abide by incomplete contracts, out of fear that wanting to renegotiate could be seen as an attempt to extract more of the surplus. However, this fear of external sanctions can also deter efficient contract renegotiations.

ultimate goal is to develop a theoretical model to guide their empirical analysis of British household recycling data. They identify three motives for recycling: warm glow[59], social norms of reputation, and environmental concern. In the following, we will focus on the modeling of the social norm motive. A certain level of recycling activity $\bar{R}$ is the norm. Positive peer approval p is obtained if household i's recycling activity R_i is fulfilling the prescriptions of the norm. Peer approval is modeled as $p = \phi(R_i - \bar{R})$, with $\phi'(z) > 0$ for $z < 0$ and $\phi'(z) \geq 0$ for $z \geq 0$, where $z = R_i - \bar{R}$. Abbot el al. (2013) assume that below the norm prescribed level peer approval rises at an increasing rate, $\phi''(z) > 0$, while above the norm level it rises at a decreasing rate, $\phi'' \leq 0$. This formulation of the norm is then added into the utility maximization problem of the household.[60]

The modelization of the norm in Kessler and Leider (2012) and Abbot et al. (2013) is particularly simple. The norm exists as an exogenously given factor. Its impact on the maximization problem of norm sensitive individuals is not surprising. Kessler and Leider (2012) additionally model a norm sensitivity parameter. The effect of the function that translates deviations from norm behavior into disutility can thus be further modified according to how norm sensitive the player is. As a building block in more involved models, this threshold specification will be recurrently referred to in the following. We will return to it in subsection V.4.1, where theoretical contributions in empirical work will be compiled.

V.3.4 Multiple norms

So far most of the models under consideration concentrated on the effect of a single norm which affected a specific form of behavior. However, people are subject to many different social norms, which may prescribe different behavior in a situation. This problem is tackled in the model of Fischer and Huddart (2008). Additionally, personal values may interfere with injunctive and descriptive norms which work at the same time. This is described in d'Adda et al. (2020).

Fischer and Huddart (2008) approach social norms with a multiple-task agency model.[61] A continuum of agents work for a principal. The agent can carry out two actions, one which benefits the principal's payoff, called desirable action $a_i \geq 0$, and one which is detrimental or costly to the principal, called undesirable action $u_i \geq 0$. A

[59] Cf. Andreoni (1990). Warm glow describes that people may derive a positive utility from performing a task without concern for material payoffs or the outcome.

[60] Abbot et al. find an effect of the social norm and the general environmental concerns motive in their data, but cannot establish a relation between recycling behavior and warm glow, which might have to do with the particular characterization and operationalization of warm glow as time spent on the recycling activity. In contrast, in a field study by Viscusi et al. (2011), warm glow and social norms both have a positive influence on recycling behavior.

[61] Cf. Holmstrom and Milgrom (1991), for the central reference for this class of models.

contract specifies a fixed wage w_i and a performance dependent wage b_i. An agent's performance dependent remuneration depends on an informative report r_i the principal receives about the level of actions taken by the agent. The report is stochastic with mean $h(a_i + u_i)$, such that the principal cannot distinguish between the two actions. However, both actions are positive for the agent because they increase the performance measure. Additionally, a norm parameter exists for each action. Technically, a norm parameter N_{a_i} influences the total and marginal costs of taking action a_i. This norm parameter is composed of the agent's personal values and the social norms they adhere to,

$$N_{a_i} \equiv (1 - \alpha_i) A_i + \alpha_i S_a, \tag{113}$$

where $\alpha > 0$ indicates the weight agent i attaches to S_a, the average level of action a taken in the organization, i.e. the social work norm concerning action a, and, correspondingly, to A_i, the agent's personal value. In summary, given a contract specifying a fixed wage w_i and a bonus dependent on performance b_i, agent i maximizes

$$z(a_i, u_i) \equiv w_i + b_i h(a_i + u_i) - f(a_i - N_{a_i}) - f(u_i + N_{u_i}), \tag{114}$$

where $f(\cdot)$ is the cost function for the respective actions a_i and u_i, with $f(\cdot)$ continuous, $f' > 0$ and $f'' > 0$. N_{u_i} is the norm parameter for the undesirable action and constructed similarly to N_{a_i} above, with a personal norm of the agent with regard to the undesirable action, U_i, and a social norm for the average level of the undesirable action in the company S_u. It is important to note that Fischer and Huddart (2008) model the norm parameter in such a way that a higher value of the norm parameter always benefits the principal. A higher norm parameter of a desirable action reduces the marginal costs of this action and therefore encourages the agent to dedicate a higher level of effort to this action. A higher level of the norm parameter for the undesirable action, on the other hand, increases the marginal costs for that action, and therefore this action becomes less attractive to the agent. This reflects the idea that a norm concerning perceived positive behavior encourages more of the action, whereas a norm concerning perceived negative behavior implies sanctions or punishment for such an action. The social norms in this company are therefore designed to encourage desired behavior and to prevent undesired behavior.

Employing a first-order approach, Fischer and Huddart (2008) find that, for any contract and the norms S_a and S_u in the company, there always exists a unique post-contracting equilibrium of desirable and undesirable actions. For interior solutions, a higher bonus b has a direct effect on effort, since the agent increases effort for both the desirable as well as the undesirable action. Moreover, there is an indirect effect of increased material incentives. An induced higher level of a_i strengthens the social norm concerning the desirable action, which leads to more effort provision for the desirable action, since its costs are reduced. However, the same is true for the undesirable action, but here the social norm is weakened, which reduces the costs and leads to higher effort provision. Depending on the parameters, this can dampen the direct effect of a higher b.

The analysis carried out also leads to insights regarding the composition and size of an organization if agents differ in their sensitivity towards norms. It might be beneficial to split a firm to maintain social norms by putting similar agents together. However, this idea has to be contrasted with possible synergy effects of having different agents carry out different tasks. Furthermore, if agents only differ with regard to the norm parameter for the desirable action, a contract exists that only agents for which the desirable norm is important (high α) accept, i.e. agents self-select due to their objectives and the principal's being aligned. If the agents only differ with respect to the importance they attribute to the norm for the undesirable action, such self-selection is not possible. Additionally, if the social norm prescribes a higher effort level than the private value, $S_a > A_i$, the desirable action is increasing in α, and the undesirable action is decreasing. The opposite holds true for $S_a < A_i$. In plain words, if the agent is relatively norm sensitive and the norm prescribes a higher level than his personal values, the agent directs more effort to the desirable action and less to the undesirable. However, if the norm is weaker than the personal values of the relatively norm sensitive agent, the agent reduces the effort directed to the desirable action and increases the effort for the undesirable action. Similarly, but not illustrated here, such effects on the actions exist for the sensitivity parameter, the social norm and the personal value of the undesirable action.

Fisher and Huddart (2008) show that there are important interaction effects if several norms are relevant at the same time. Norms are interpreted as efficiency enhancing social constructs in this model. Despite modeling two social norms, the effects of the norms have the same direction. Stronger norms are beneficial for the principal here. Norms are cost modifying devices, that encourage behavior beneficial for the company and discourage harmful behavior. This stylized representation might be a bit too optimistic, as norms that encourage behavior with adverse consequences may also exist.[62] The model in Fisher and Huddart (2008) is, admittedly, open to such an extension, which would improve the theoretical landscape on the critical issue of multiple interactive norms.

[62] Ichino and Maggi (2000) are interested in different work norms inside the same company. They analyze how accepted shirking in the work place is in a nationwide operating Italian bank, i.e. what the average accepted shirking level in different regions is. They find that there are different shirking levels with differences especially striking between the north and the south of Italy. They explain the difference by the prevailing regional social norms and the cultural background of the employees. Being born in the south increases the propensity to shirk wherever one works, and working in the south increases the propensity to shirk, wherever one is born. This highlights the interplay of norms one internalized and norms one is surrounded by currently, and also hints at locality as an important factor for the differences of social norms (cf. Young, 2015).

In the second model in this subsection, d'Adda et al. (2020) build on the work on norms by Bicchieri (2006). Individual behavior depends on material incentives as well as subjective values, the expectations of what others ought to do, and what one expects others to be doing, i.e. injunctive and descriptive norms, respectively. Their most important contribution to the discussion of norms is the concept of partial norms. An ideal norm is the same for everybody and everybody agrees what the threshold or adequate action for norm compliance is. A partial norm, however, allows for some interval or set of values that all constitute norm compliance, although these actions might be quite different. A player in the model of d'Adda et al. has to donate a certain amount and tries to minimize a quadratic loss function, where the norm is modeled as the target, i.e. as a threshold,

$$W = x + \frac{(N-x)^2}{2\theta}. \tag{115}$$

The player cares about reducing their donation x but is also norm sensitive. $\theta > 0$ is a parameter that indicates how strongly the tradeoff between material and normative motives affects the player. The larger θ, the closer the player comes to selfish behavior, and the lower θ, the more important norm following becomes. Their norm sensitivity N is composed of the following parts. They have a private view of what one ought to do, their subjective value r. r is a random variable, which means that different players have different subjective value realizations. Next, $E(r)$ is what the player expects everybody else's r to be, i.e. an injunctive norm. The player's variance with respect to what they believe others believe to be the right thing to do is $V(r)$. Additionally, the player has expectations about what others actually do $E(x)$, i.e. a descriptive norm. These parts of the player's utility function that model this norm setup can be summarized as $N = r + \alpha[E(r) - r] + \beta[E(x) - r]$, where $\alpha, \beta > 0$ and $\alpha + \beta < 1$. α captures the relative weight of the injunctive norm, β the relative weight of the descriptive norm, and $1 - \alpha - \beta$ the relative weight of personal values. The authors follow this up with a common trade-off problem: weighing off material gains against intrinsic values and social motivations. The only heterogeneity between the players is their personal value r. With the help of simplifying assumptions, plugging N into the loss function and solving a system of linear equations reveals that the donation x will be

$$x = r + (\alpha + \beta)[E(r) - r] - \frac{\theta}{1-\beta}. \tag{116}$$

The donation will be larger the larger the player's subjective value and the higher they expect other player's normative disposition to be. The effect of α has the same sign as $[E(r) - r]$, but β's effect depends on $[E(r) - r]$ as well as the size of θ. The effect of a higher uncertainty about the norm, i.e. a higher $V(r)$, on the amount donated can be positive or negative. α can be interpreted as indicating how far society is from an ideal norm, i.e. if everybody had the same value system $E(r) = r$. Then everybody would fully conform. With differences in r the injunctive norm is weakened, indicated by a smaller α. Additionally, a higher $V(r)$ changes N because players put more value on their personal value r than their expectation of what others ought to do, $E(r)$. If $r > E(r)$, N increases and the player will donate more. If $r < E(r)$, N decreases and they will donate less. A higher $V(r)$ will also lead to a larger variance of actual donations $V(x)$. Uncertainty about the norm can thus allow people to behave closer to their own personal values, offering them greater liberty to act selfishly, which may entail higher contributions, since subjective values can be understood as ”selfish” here.

d'Adda et al. (2020) model the interplay between subjective personal values and injunctive social norms. Unfortunately, the relations between the descriptive and the injunctive norm are underdeveloped. Nevertheless, they find that the beliefs about prevalent norms are sensitive to information about others' behavior. A partial norm, with some variance concerning the correct level of action in order to comply with the norm, reduces the conformity effect of the norm.[63]

The two models with multiple norms highlight that complicated interaction and interdependence between social norms are to be expected, if several are applicable at the same time. Additionally, uncertainty about which norm is relevant or how incongruent information should be treated, refers people back to decide for themselves how to act, maybe leaving them without orientation. There is a need for more models and research concerned with the interaction of several social norms that are relevant simultaneously. According to Görges and Nosenzo (2020), in their review of the experimental literature of social norms in the labor market, the interaction of different norms is still little understood, as is the question which norms do apply and when, and what happens when different groups with different sets of norms have to interact (cf. subsection V.4.2 below).

V.3.5 Norms and society

This subsection incorporates two papers that analyze social norms as fundamental building blocks of society. In Alpmann (2013), the need of human beings to interact with other human beings is analyzed. Norms restrict the possibilities of acting only with respect to one's own interest if one wants to be part of society. In Michaeli and Spiro (2015), constitutional aspects of societies and how they relate to freedom of speech are in focus. The way societies punish social norm transgressions has far reaching consequences in this regard.

Alpman (2013) proposes a different approach to analyze the persistence of inefficient social norms in a society of rational individuals. The disutility of disobeying a

[63] Bicchieri and her co-authors contribute experimental evidence to the phenomenon that people will exploit such ”moral wiggle room”, where there is uncertainty about the norm. They will self-servingly interpret information in order to avoid a norm being applicable or to manipulate a norm. Cf. Bicchieri and Chavez (2010, 2013) and Bicchieri et al. (2021). Cf. also Dana et al. (2007).

social norm is to be found in a loss of social interactions rather than a loss of identity (Kranton and Akerlof, 2000) or a loss of reputation (Akerlof, 1980). In his view, reputation serves only as a means to easier interact with others and does not carry a value in and of itself. The interaction with others is therefore a source of utility in addition to consumption. However, these social interactions have to be produced by the individuals themselves. In this context, inefficient social norms are rules of behavior that proscribe economically efficient exchanges. Alpman (2013) makes the assumption that disobeying inefficient social norms makes the production of social interactions more difficult but at the same time increases the utility from consumption. The individual has to trade off social interaction and consumption and has to find the optimal level of norm adherence. Alpman (2013) finds that there exists an optimal level of disobedience, which depends on the type of the individual and their preferences for consumption and norm compliance as well as how strict society reacts to norm transgressions. Taken together, this offers an explanation for the observation of partial compliance with social norms.

In Alpman (2013), the player's utility depends on their production of social interactions Z_s and a consumption good Z_c, which are imperfect substitutes.[64] The player produces Z_c with time t_c and market goods C_c, according to a Cobb-Douglas production function, $Z_c = C_c^{\beta} t_c^{1-\beta}$, where subscript c stands for consumption. For the production of Z_c, social norms are irrelevant. The production of the social interaction commodity is also modeled with a Cobb-Douglas production function $Z_s = \mu(x, \delta) C_s^{\alpha} t_s^{1-\alpha}$, where C_s is the market good and t_s is the time used to produce social interactions, with subscript s standing for social interaction. For our purposes, the most interesting factor is μ, the total factor productivity of C_s and t_s and a function of x, the disobedience level, and δ, the intolerance of society towards disobedience. A higher x represents a higher disobedience level, where $x \geq 0$. A higher δ stands for a more severe punishment executed by society for disobedience, where $\delta > 0$.

The assumptions on x and δ drive Alpmans (2013) model. A higher x deteriorates reputation, which complicates the production of social interactions. Technically, a higher x reduces the total factor productivity μ: $\partial\mu/\partial x < 0$. μ attains its highest value for $x = 0$. Additionally, Alpman (2013) assumes that society does not punish disobedience in a linear fashion, i.e. higher levels of x are punished disproportionally more severe. This effect of a higher x on μ is described by: $\partial^2\mu/\partial x^2 < 0$. Furthermore, productivity is reduced by the intolerance of society for disobedience for all levels of x above zero, i.e. $\partial\mu/\partial\delta < 0$ and $\partial\mu/\partial x\partial\delta < 0$. For a sufficiently high x, $\mu = 0$, which means that the agent is banished from society and cannot produce social interactions.

[64] Alpman (2013) follows the approach of Michael and Becker (1973), who contributed to consumer behavior theory by modeling utility as derived from commodities produced by the consumers themselves, using purchased market goods and the consumer's own time as production factors (Michael and Becker, 1973, p. 381).

The efficiency, with which the social interaction good is produced, is reduced if the player transgresses the norm, the more, the more vindictive the society is.

Furthermore, in order to purchase consumption goods C_i, $i \in c, s$, the player has to work. In addition to a market wage rate W they earn, they can gain additional income $\gamma(x,\delta)$[65] by breaking the inefficient norm. The size of γ depends on the level of x. A higher x improves the additional income but also increases the punishments through society. The marginal utility of breaking the norm is diminishing.

The consumer faces an income constraint and a time constraint. They have to procure C_i at the market price and have an income, that depends on the fixed market wage W, their level of disobedience x, the strictness of society δ, and the time they dedicate to working. They have to divide their total disposable time between producing consumption goods and social interaction goods and working. Alpman (2013) continues with determining the optimal levels of t_i, C_i, Z_i, and x in order to maximize the player's utility.

The most important conclusion for our purpose is: if people have a higher preference for social interaction, they will tolerate more inefficient norms. Social interactions, and not, for example, concern for reputation, are, according to Alpman (2013), why we adhere to social norms. However, replacing his production of social interaction with the production of reputation would lead to the same results with reputation instead of social interaction. Nevertheless, his description of society might be more accurate, since we behave according to norms in society to be part of society and to continue being part of it, which implies continuously interacting with other people. Interactions are constitutive for human beings. Reputation might only be a secondary gain we obtain from them. The social interaction commodity is positively connected to following inefficient norms, since breaking these norms complicates producing social interactions. This can be sufficient to keep inefficient norms in place. That is, the social norms make people take choices of consumption they would not make otherwise. If society punishes deviation sufficiently harsh, inefficient norms can be sustained over time. Alpman (2013) does not model the inefficient social norm explicitly. Instead he assumes that an inefficient social norm exists. He explicitly models a society where social norms organize social coherence.

Michaeli and Spiro (2015) focus on social norms in a society that are in conflict with privately held beliefs of an agent. Agents are heterogeneous with respect to their private opinions. Every agent would like to announce their opinion according to their type. However, individuals in this model have to make a public announcement on where they stand on the continuum of possible opinions. They therefore face a tradeoff: complying to an established norm of opinion in society and suffering individual pain from standing for something one does not believe in or standing up for one's val-

[65] Alpman (2013) adds another variable that determines γ, which is not vital for our purpose here.

ues and being exposed to social punishment if the uttered opinion differs from the societal norm. The interesting aspect Michaeli and Spiro (2015) introduce into their model of social norms is the idea that different societies react differently to norm transgressions. They differentiate liberal and strict societies. In liberal societies, small to medium deviations from the norm are punished only mildly, if at all. Only extreme positions are punished but these are punished severely. Theses societies are said to apply convexly curved social pressure, with the size of the transgression on the x-axis and the severity of punishment on the y-axis. Strict societies, on the other hand, punish even small transgressions harshly. This leaves strict societies with little scope to increase the severity of punishments for larger transgressions. Strict societies therefore apply concavely curved social pressure. Michael and Spiro combine several interesting aspects in their model: social pressure can be applied in different ways and strengths, the severity of the transgression is important for the severity of punishment, and different societies can accustom divergent behavior variably well, which has repercussions for protest movements and societal stability.

In the model, an individual has a certain type t, who has to make a public announcement s of their opinion concerning a certain issue. If t and s differ, the individual suffers from this cognitive dissonance D. Formally,

$$D(s,t) = |s-t|^{\alpha}, \tag{117}$$

with $\alpha > 0$. The parameter α represents the sensitivity of an individual with regard to large and small deviations. For $\alpha < 1$, the individual's cognitive dissonance is concave, i.e. they strongly dislike even small differences between their values and their public announcement. For $\alpha > 1$ the cognitive dissonance is convex. Such an individual is not distressed by small levels of cognitive dissonance. In addition, by publicly declaring s to be their opinion on the issue, the individual also feels social pressure to conform to the prevailing social norm. The pressure they feel is modeled similarly to their inner discomfort as

$$P(s,\bar{s}) = K|s-\bar{s}|^{\beta}, \tag{118}$$

where $\beta > 0$ and $\bar{s}$ is the social norm. K measures the relative strength of the two forms of discomfort, i.e. external social pressure and internal cognitive dissonance. As above for α in the cognitive dissonance case, for $\beta > 1$. society does not mind small deviations too much, i.e. it is liberal; and for $\beta < 1$ it is strict and punishes even small deviations heavily. An individual will try to minimize their disutility from both forms of discomfort, cognitive dissonance and social pressure. The social norm $\bar{s}$ is exogenous for the individual, but in equilibrium it is determined by the average publicly announced opinions s, i.e. $\bar{s} = E[s^*(t)]$ with $s^*(t)$ the pronunciation that minimizes the loss of individual t.

The main findings of this model (for the case of $\alpha = 1$) are that in liberal societies types close to the norm state their types correctly (a corner solution), while types further away from the norm have inner solutions, i.e. they neither pronounce their true type nor do they conform completely to the norm. In strict societies, types close to the norm do not deviate and types far from the norm truthfully utter their type. In liberal societies, therefore, extreme types moderate their stance to avoid the harshest punishments, whereas close to the norm several differing opinions can be accommodated, i.e. nobody fully conforms and everybody chooses an s on either side of the norm. In strict societies, there will be high congruence between public stances and the social norm and only extreme types will speak their mind. Strict societies thus can induce conformity over a large range of types. A strict society does not allow for compromise and pits the extremists, who despise the society's norm, against the conformists, whereas a liberal society allows for many levels of norm adherence. Finally, liberal societies have norms that align with the average public opinion in equilibrium, whereas strict societies can maintain a biased norm, i.e. the norm favors some types and is closer to them. Michaeli and Spiro (2015) emphasize that the relative strength of α and β, i.e. the curvature of cognitive dissonance and social pressure, respectively, drive their results and continue with going through all possible combinations, which complicates the straightforward results of the base model presented here to some extent.

Michaeli and Spiro (2015) model a conflict between inner values and social pressure to conform to a norm. They highlight that the way transgressions are pursued can result in very different societal structures. The way external sanctions are administered is here the key to the societies' character and the conformity or relative non-conformity that can be modeled. Social norms prove to be context sensitive in this respect. As in Alpman (2013), the severity of external punishment is the deciding factor for how much norm transgression an individual will risk in order to improve either their economic prospects or their self-image. At least some norm transgression will take place in both models. Norms, modeled according to these authors, are seldom universal or encompassing.

V.3.6 Matching games

The models in this subsection analyze norms in infinitely repeated matching games. They show that norms contribute to the ability of societies to coordinate players' behavior and achieve better outcomes in conflict situations.

The model of Kandori (1992) is concerned with one pertinent result of the infinitely repeated games literature, the Folk Theorem. It states that with infinitely repeated play, virtually any payoff can be sustained in a subgame-perfect equilibrium if players are sufficiently patient (Fudenberg and Tirole, (1991), pp. 150-165.).This means that sufficiently patient players can maintain cooperation with informal personal enforcement, i.e. the players themselves punish deviations by their opponents. This striking result hinges, among other factors, on the high frequency with which the same players interact and the common knowledge of all relevant information for all players. Kandori (1992) argues that infrequent interactions are an important part of economic exchange and that important information is often private. The enforcement mechanism for cooperation in these circumstances cannot come directly from the player who has been cheated on. Instead, it has to come from the community of players, i.e. a defector is informally sanctioned by the community of players instead of by their direct opponent in the future rounds of play. A social norm in this specification is the description of the condoned behavior and the punishment rules for defectors in the community. Kandori (1992) shows that social norms can lead to desirable outcomes in such games with infrequent interactions.

Kandori (1992) models two equally large sets of players, with n players each. In each round of the infinitely repeated game, each player of one group is matched[66] with one player of the other group and they play a Prisoner's Dilemma stage game. The expected payoffs are the discounted sums of the stage payoffs of the players. The players are selfish. Therefore, the enforcement mechanism for cooperation must provide sufficiently large incentives (or high penalties) in order to maintain cooperation. Additionally, players only know the history of the games they were part of. Kandori's (1992) main research focus, since information about past play is private here, is what the minimal amount of information about past play is that has to be transferred between the players to maintain cooperation with community sanctions. The private information will get more and more complicated after each round, since different players may observe different behavior and, after a deviation, the players who experienced a deviation and those who did not, no longer share a common prior, since past experiences differ. In order to show that a social norm of cooperation can be sustained in equilibrium, Kandori (1992) examines different information transmission mechanisms.

The first information transmission mechanism under inspection is no transmission at all, such that players only know what happened in the games they were part of. Despite this humble amount of information, Kandori (1992) identifies a sequential equilibrium in this case, where cooperation can be maintained for sufficiently patient players and sufficiently large punishments. This equilibrium is called a contagious equilibrium. The strategy to maintain this equilibrium is: every player who observes defection in one of their games, henceforth defects as well, i.e. a classical tit for tat strategy. Non-cooperation is spreading like a disease in the community after defection. However, the defector is unlikely to be punished immediately for their transgression, which weakens the enforcement power compared to the case with perfect information. Moreover, and more importantly, it takes a long time in large communities for the deviations to spread, thus making sustained cooperation harder or impossible in large

[66] The matching mechanism used is not critical.

communities. Furthermore, this equilibrium punishes not only cheaters but many innocent players as well. Finally, the equilibrium is susceptible to collapse for small deviations, i.e. it is not trembling hand perfect. The instability of the previous equilibrium leads Kandori (1992) next to local information transmission as a second mechanism in order to identify more robust equilibria. With local information transmission, at least some information is truthfully processed (exogenously) and made available to players. In this setup, each player carries a visible label. Before playing the stage game, the two players observe each other's label. After an interaction, players' labels are updated according to a rule, given the original labels and the actions taken in the game just played. This describes what is meant by local information, since the possible change of label is only based on information available when observing the specific stage game. All the necessary information for the interaction is contained in the label. Kandori (1992) highlights that for important classes[67] of stage games, the social norm of cooperation can be sustained with local information transmission for any size of the community and for any matching rule. Kandori (1992) argues that a realistic equivalent for the information contained in the label of the other players could be the status of a person or them belonging to a certain club or cultural subgroup with a convincing identification device. However, the result hinges on the costless, exogenously produced label. The information processing institution would have to produce the label costlessly and either be assumed to be neutral or endowed with the right incentives. Kandori (1992) contributes to the social norm literature by showing that for an important class of repeated matching games, equilibria with a social norm of cooperation can be stable with community punishment. The information requirement for this equilibrium is, admittedly, high.

Okuno-Fujiwara and Postlewaite (1995) also analyze an infinitely repeated matching game but with uniform random matching. Their impetus is also the unrealistic assumption of common knowledge of all relevant information in a game, constitutive for the folk theorem. They aim at showing how norms can coordinate behavior of players in infinitely repeated games, where the stage game represents a conflict between the players. Players are drawn from two continuous sets, [0,1],[68] and in every round the players of one set are matched with the players of the other set. Additionally, in each round every player is assigned a status element, in the first round drawn from a finite set of status and later updated according to an updating rule. Status provides indirect information about past play. While not explicitly stated, a higher status is assumed to be better in the model. As in Kandori (1992), players have only local information. They know their own status type and they can observe the status of their opponent. After the round is played, a player's status is updated according to a rule which determines

[67] The setup with a finite set of players requires some restrictions to avoid dealing with incentive problems off the equilibrium path.

[68] In one extension of their model, the authors illustrate their model with two finite groups of players.

the status in the next round. The updating rule considers the current status level of the player, the current status of the opponent, and the action the player took in this round. In this society, there is a prescribed adequate behavior for every type of status, of what to play, given the status of the other player. The updating rule together with this prescription of adequate behavior are called a social norm in this model. The social norm will guide behavior by establishing what one is expected to do and what to expect the other player will do.

Agents are modeled as selfish and act according to their best self-interest, i.e. if it appears advantageous, they will break the norm. A social norm can only be sustained, i.e. be a stable equilibrium of the game, if it is optimal for each player to follow the norm prescribed behavior. That means that the short term gain from deviating in the stage game must be less than the discounted loss in the future, due to a changed status. Okuno-Fujiwara and Postlewaite (1995) continue this line of reasoning with determining conditions for when a norm of cooperation can be sustained as an equilibrium in a game, by way of examples.

Kandori (1992) and Okuno-Fujiwara and Postlewaite (1995) contributed to the literature on social norms by showing that such norms can be effective in increasing cooperation in conflict situations when the information structure is not perfect. In addition, this line of research has highlighted that the norm equilibria do not require common knowledge on all aspects of the game by all players.[69] Nevertheless, as both contributions state, the information requirements are still relatively high. A player needs exogenously given information about the other player's characteristics in order for there to be an equilibrium. The information structure employed is thus still demanding and costs for information transmission and for establishing and maintaining a neutral labeling or information processing organization with the right incentives have been blended out in both approaches.[70] The Matching games and the norm equilibria they found are in some respects only the prelude for the evolutionary game theory models in the next subsection. There, people will be matched, although there will usually only be one population set. Information transmission mechanisms will be replaced with learning mechanisms. Norms will be equilibria of the game in time, but their stability concepts and dynamics of play will come into focus.[71]

[69] Both contributions present a Folk Theorem with only local information processing.

[70] An empirical application of these concepts can be found e.g. in Munshi and Myaux (2006), who develop a model of norm change building on Kandori (1992). In the context of the acceptance of family planning methods in rural Bangladesh, traditionalists and modernizers are matched (although here, people from the same group can also be matched). If enough people gain sufficiently from modern contraception, local transmission of information can lead to a different societal outcome, where modern contraception is accepted in conservative societies.

[71] There is also a model by Cole et al. (1992), who combine an infinitely repeated matching game with a capital accumulation process. Here, social norms impart status, i.e. determine a ranking system in society. In equilibrium, people match according to their rank. The highest man with the highest woman and so on. They employ two social norms to illustrate the effects of their model: plutocracy and

V.3.7 Evolutionary game theory

Traditional evolutionary game theory proposed that players have a fixed strategy which they follow without ever changing it, e.g. a gene in biology. The more successful strategies would allow its carriers to survive because they were able to reproduce more quickly in the long run. Once economists applied these biological concepts to social interaction, some adaptations were necessary. Humans learn behavior and internalize social norms, i.e. expected group behavior. Since the strategies refer to behavior, individuals can, in principle, change and adapt their strategy comparatively quickly. However, evolutionary game theoretical models, once accounted for the social realities they are supposed to represent, are capable of illuminating the existence, development, and some empirical characteristics of social norms. The norms are interpreted as the equilibria of the evolutionary game. Since these kind of games have usually many possible equilibria, which norm asserts itself, is due to historic accidents, starting positions, or exogenous shocks (Ostrom, 2000). This constitutes a weakness of evolutionary models of social norms. They cannot predict which norm will obtain for what reasons at a certain point in time in some cultural setting.[72] Their strength consists in showing that social norms (of cooperation, for example) can be sustained in a dynamic interaction that involves different types of players. These norms are not necessarily eradicated by selfish behavior. Another advantage of evolutionary models is that players in these models do not have to be perfectly rational. They are even allowed to make small mistakes in some specification. The modeling focuses on the eventually established rules in a society (Matsui, 1996). An important building block of evolutionary game theory with regard to social norms is a learning mechanism of how players learn about successful strategies (Binmore and Samuelson, 1994). In the following, three of these learning mechanisms are presented: adaptive learning, simple replicator dynamics, and cultural transmission. The literature on evolutionary models of social norms is extensive.[73] This subsection is designed to highlight the basic ideas of how evolutionary processes can lead to stable norms.

Young (1993) develops an evolutionary model of conventions or norms (Young, 1998a), where a convention or a norm is an equilibrium of a game that everyone ex-

aristocracy. With plutocracy, money is everything and the status that ranks the people for the matching process is wealth. With aristocracy, the status depends only on the status of the ancestors. These two norms lead to very different economic decisions, matching decisions, and societies. Cole et al. (1992) are mainly focused on combining their matching game with their growth model. Social norms appear more like an addendum. They do not discuss other ranking norms but the two extreme forms. Their modeling choices make it highly complicated to model a ranking norm that relies on money and inherited status for current status. It provides an interesting idea but unfortunately contributes little to social norm modeling.

[72] One attempt to remedy this shortcoming can be seen e.g. in Roos et al. (2015). They illustrate how external threats can push a society to be more prone to norm adherence and administering higher punishments for norm transgression.

[73] For a first orientation consult e.g. Sethi and Somanathan (2003) or Young (2015).

pects to be played. The main focus of Young's (1993) approach is the question of how norms come into being. In his view, a norm helps organize social interaction if more than one equilibrium exists. In order to fulfill this role, playing the norm prescribed action in a game must have positive feedback effects for the players. In this setup, norms develop and are changed by many small actions of individuals in a society. Eventually, the evolutionary dynamics of play may lead to the selection of the conventional equilibrium and thus the norm, for a given stability concept. Actions in the past guide action in the present, as agents mold their best action according to their knowledge of past play. The past therefore influences the present, and present actions become the guideline for future action.

Technically, a stage game is played by a random selection of n players from a large finite population once in every period. Each player forms beliefs about which strategy is most promising, given that they have limited knowledge of recent play in the past. An agent who is selected to play can take a sample k of the histories of past play up to m periods in the past, with $1 \leq k \leq m$. Histories of play that are further than m periods back in the past are deleted and hence do not influence play in the presence or future any more.[74] The player may now choose a strategy of play from this sample, or they may experiment with a new strategy. The player is also allowed to make mistakes in some extensions of the model. Learning in this model is on the societal level. A single agent only plays the game once. Therefore, it is the information available to all agents about past play that allows identifying successful strategies. If the players do not make mistakes, i.e. always play a best response to their sample k, Young (1993) shows that a weakly acyclic game (e.g. a coordination game) converges with very high probability to a pure strategy Nash equilibrium. This equilibrium will be the norm in this society. However, which equilibrium is selected cannot be determined a priori but depends on the initial state and the unpredictability of the process. Given some initial m and a state space, the system can be described by a Markov chain, with the transition probability from the current state to the next state depending on the product of the actions chosen in the current state by each agent as a best reply to their sample k. If players are allowed to make mistakes, i.e. their selected action is not necessarily a best response to their sample k, this Markov chain development is disturbed. Mistakes pull the process away from equilibrium. With mistakes, different equilibria can be reached with a certain probability. Interestingly, this process can flow from one equilibrium to another for relatively frequent mistakes. If the probability of mistakes approaches zero, some subset of equilibria will be observed with higher frequency than others, in the long run. For a very small probability of mistakes, most of the probability weight will be on one equilibrium, called a statistically stable equilibrium. Such an equilibrium is restored after a small shock to the system and thus relatively stable, i.e. frequently (almost ex-

[74] This concept of learning is called adaptive play. For a detailed introduction cf. Young (1998b), chapter 2.

clusively) played.[75] A social norm is hence the outcome of many small interactions and, to some extent, trial and error processes.

The next three models by Alger and Weibull (2013), Nyborg and Rege (2003), and Azar (2004) employ replicator dynamics as learning process. Replicator dynamics formalize how a replicator (a gene, a strategy etc.) changes under evolutionary forces. For example, for a strategy in a large population, the replicator dynamic is $\dot{p}_i^t = p_i^t(\pi_i^t - \bar{\pi}^t)$, with p_i the frequency of strategy i in the population, π_i^t the payoff of strategy i, and $\bar{\pi}^t$ the average payoff in the population. Superscript t indicates the point in time. That means, the frequency of a strategy increases if its payoff is above the average payoff in the population. For social applications, the difference in payoffs makes people modify their behavior if they observe a higher payoff for another player. This is a form of learning, as the players learn about successful strategies and if the payoff differential is large enough, they imitate successful behavior (cf. Gintis, 2009, chapter 12.).

The model by Alger and Weibull (2013) posits homo moralis, a mixture between pure self-interest and doing what is right. The norm or doing what is right in this model consists in choosing an action that, given that all other players play this strategy, would maximize the collective payoff. Acting according to this strategy is the morally right thing to do. Alger and Weibull (2013) want to show that this homo moralis can be an evolutionary stable type.[76] They extend the concept of evolutionary stable strategies beyond only being applicable to hardwired (genetic) strategies and show that it can also be used when rational players optimize and have correct beliefs about the statistical characteristics of the population. The evolutionary game theory apparatus in place is the following. There are two types of players in a population θ and τ, with $\epsilon \in (0,1)$ the portion of type τ in the population, and $\theta, \tau \in \Theta$, the type space. The two types play a (fitness) game. The two types and their population share define a state of the world $s = (\theta, \tau, \epsilon)$. The set of states is therefore $S = \Theta^2 \times (0,1)$. When type θ plays strategy x, and type τ plays strategy y, type θ gets a payoff of $\pi(x, y)$. Given a matching mechanism (e.g. random matching or assortative matching), types meet their own kind with some probability and the other type with the complementary probability. The strategy pair (x^*, y^*) constitutes a Bayesian Nash Equilibrium of the fitness game if x^* and y^* are maximizing the respective expected utilities of the two types. A type θ is said to be evolutionary stable if a share of a type τ, $\bar{\epsilon} > 0$, exists, such that in all Nash equilibria in all states s, the payoff of type θ is larger than the payoff of type τ if $\epsilon \in (0, \bar{\epsilon})$. This has to hold against all types $\tau \neq \theta \in \Theta$.

[75] Asymptotic stability: starting the process in the neighborhood of this stable equilibrium (i.e. its basin of attraction) the process will always converge back to the equilibrium in the long run, cf. Ashby [1956] (2015), Young 1998b, chapter 3.

[76] A population of a certain type is said to be evolutionary stable if a small invasion of a different type (a mutation) cannot replace the first type, cf. Maynard Smith and Price (1973).

A type is now called homo moralis if they have the utility function

$$u_\kappa(x,y) = (1-\kappa)\pi(x,y) + \kappa\pi(x,x), \tag{119}$$

where $\kappa \in [0,1]$ measures how much the player values acting morally. This is a convex combination of selfish behavior and following the norm. For $\kappa = 0$, the player is homo oeconomicus and, given any y, will use a strategy to maximize their own payoff. For $\kappa = 1$, the player will always choose a strategy from the set of actions the norm prescribes, irrespective of what the other player may play. In the following, Alger and Weibull (2013) establish that homo moralis can be evolutionary stable with strategies evolving according to a replicator dynamic and for matching processes with correlations (e.g. family) as well as with uniform matching processes. Alger and Weibull (2013) show that, if people expect others to act according to a mutually beneficial strategy, it can be the best response to act according to this strategy as well. Especially in assortative matching situations (which are likely more important for human interactions, i.e. friends, family, or colleagues), homo moralis is evolutionary stable, while homo oeconomicus does not survive. This may illuminate why cooperative behavior exists in human interaction and is not eroded by small invasions of selfishness. Alger and Weibull (2013) thus contribute to the literature on social norms by modeling a player that is a combination of economic and sociological understandings of the human disposition.

Nyborg and Rege (2003) model the lasting effects of interventions on the evolutionary equilibrium. The background is the evolution of smoking behavior under the influence of legislation banning smoking from certain public areas. The learning process, modeled as a replicator dynamic, driving their model can be illustrated with the following story. Smokers are forced by legislation to reduce smoking, for example in restaurants or in the work place. This reduces the level of tolerance non-smokers have for being exposed to passive smoking. This in turn increases the price to be an inconsiderate smoker, i.e. smoking in the presence of non-smokers anywhere, which can lead to a new social equilibrium.. Technically, a smoker has to divide their time between places they can still smoke, $R \in [0,1]$, and and places they cannot smoke, $(1-R)$. A smoker i's payoff function in this model is

$$U_i = U(\gamma_i, \bar{\gamma}) = (k\alpha(R + (1-R)\bar{\gamma}) - c)\gamma_i \tag{120}$$

where $\gamma_i \in \{0,1\}$ is i's consideration level, with $\gamma_i = 1$ meaning the smoker is considerate, i.e. leaves the room to smoke. Furthermore, $\bar{\gamma}$ is the average consideration level, k is the public belief about negative health effects of passive smoking, α is the frequency with which a smoker meets a non-smoker, and c is an inconvenience cost of considerate behavior. The stage game has three Nash equilibria, one in which all smokers are inconsiderate, one where all smokers are considerate and a mixed equilibrium where a share of smokers is considerate. The replicator dynamics illustrate the evolution of the behavior. They are characterized by

$$\dot{x} = (U(1;x) - \bar{U}(x)) = x(1-x)(k\alpha(R + (1-R)x) - c), \tag{121}$$

where $\bar{U}(x)$ is the average utility in society and x is the share of considerate smokers. Only the fully considerate smoking equilibrium, $x = 1$, and the not considerate smoking equilibrium, $x = 0$, are asymptotically stable. Nyborg and Rege (2003) further show that introducing smoking regulation can move a society from a state $x = 0$ to a state $x = 1$, given R is sufficiently large, i.e. that smoking is prohibited in enough areas. This stable equilibrium will be maintained even if the regulation is abolished later.[77] In this view, interventions to change social norms are not easily reversible. The evolutionary forces maintain the equilibrium and it takes a strong shock to change them. Even then a direct return to an old equilibrium is not guaranteed.

Azar (2004) illustrates that punishments in form of social disapproval alone are not sufficient to maintain a norm. Otherwise, a costly norm (like tipping in a restaurant) would erode over time. People need to gain other benefits from norm compliance, e.g. a better self-image, for the norm to survive. Azar (2004) makes the point that for some conditions norms can erode completely if social disapproval for disobedience is the only punishment or reward effect. Azar (2004) proposes an additively separable utility function with a social norm component and a concern for monetary payoff,

$$u(g; n_t, \theta) = d(g - n_t) + \theta p(g) - bg. \tag{122}$$

Here, n_t is the norm that specifies the size of a tip as a percentage of a given bill, g is the actual tip, d is a function representing the disutility from disobeying the norm, i.e. social disapproval, with $d'(x) \geq 0$, $\forall x < 0$ and $d'(x) \leq 0$, $\forall x > 0$, implying that the individual wants their tip to be as close to the norm as possible, since $d'(0) = 0$. Furthermore, $b > 0$ is the size of the bill, and p is the utility of the good feeling derived from tipping with $p' \geq 0$, and type $\theta \geq 0$, a idiosyncratic weight measuring the strength of the positive feeling the agent gains from tipping. The norm evolves according to a replicator dynamic, where the norm prescribed tip in the current period is the average tip from the proceeding period. Azar (2004) continues by showing that if a positive norm for tipping were to exist, but all people were of type $\theta = 0$, i.e. whose only concern for conformity followed from social disapproval, the norm of tipping would erode. This is due to the fact that people would always choose to tip strictly less than the current norm,

[77] Similarly, Janssen and Mendys-Kamphorst (2004) show that the effects of interventions can be persistent. In a simple evolutionary model, introducing financial rewards to contribute to a public good can reduce contributions if norm followers and selfish agents both care about reputation. Abolishing the monetary incentive afterwards reduces contributions even further instead of reestablishing higher contributions. Reestablishing the norm is then costly, and success uncertain.

thus reducing the average tip and consequently the norm in the next period, which would eventually erode the norm over time. However, empirically, tipping has existed for a considerable amount of time and the average tip size has even increased. This indicates that external punishment alone is not sufficient for the survival of a norm. Azar (2004) supports this claim by showing that a two type version with simple functional forms (e.g. $d(x) = -x^2$, $p(g) = g$) of his more general model is adequate to describe the empirical evolution of the tipping norm in the United States.[78]

Several other contributions employ simple replicator dynamics to analyze the evolutionary development of social norms. They all find that norm adherence of at least a part of the society can be justified as a strategy in evolutionary stable equilibria.[79] Sethi (1996) adds that initial conditions and shocks to the system can lead to different norm arrangements in the long-run, thus delivering an explanation for different norms in different cultural settings, even if the starting conditions were similar. Sethi and Somanathan (1996) show that rules of restraint and punishment can be evolutionary stable and can resist the invasion of selfish exploitation attempts of commonly owned property. Fershtman and Weiss (1998) determine that in one evolutionary stable equilibrium in their setup, where some people care about their reputation, a norm to contribute to a public good can exist. In this equilibrium the negative effects of externalities can be attenuated. Boyd and Richerson (2002) show that groups who have little interaction with each other and employ different strategies in their respective evolutionary stable equilibrium can adapt quickly to new group-beneficial strategies if individuals of the two groups interact and one group has a more successful strategy. This provides a possible explanation why norms can be stable for a long time but can sometimes change relatively rapidly. Carpenter and Matthews (2010) want to model how mutations could develop in a society. They employ enhanced replicator dynamics. Most people in society are modeled as in the standard replicator dynamics. Some people in society, however, have a specific payoff level in mind that they want to achieve in their interaction, no matter what. If they fall short of that, independent of the payoff of their interaction partner, they choose a different strategy at random. This later group can be interpreted as causing additional variation, i.e. mistakes, in the evolutionary game.

In the last model of this subsection, Mengel (2008) shows the effects of several channels of pressure affecting behavior at the same time. Two types of agents exist in this evolutionary model. One type has internalized a social norm to cooperate. The

[78] Conlin et al. (2003) use survey data to investigate tipping behavior. They find that tipping behavior depends on many factors. Among them, customers suffer a negative utility if they do not adhere to the correct norm of tipping and customers trade off higher material tips with smaller emotional costs of being a norm violator.

[79] The modeling differences, compared to the above models with regards to social norms, do not justify a more detailed discussion of these interesting papers here.

stronger the norm is, the more transgression hurts this type. The other player has not internalized the norm. These two types are matched to play a stage game Prisoner's Dilemma, where they know only the distribution of types in society. The game is played under different degrees of assortative matching, ranging from full separation to random matching of the two types. The learning mechanism in Mengel's (2008) model is cultural transmission. There are three factors important for cultural transmission of type: peer experiences that follow a replicator dynamic, institutional pressure that budge the non-cooperating type to follow the social norm, and vertical transmission, which here means a player is replaced with a new player at death, who has exactly the same type. These three factors now influence the dynamic of the frequency of strategies in the population.[80] Mengel's (2008) analysis shows that whether or not cooperation can be sustained as an equilibrium depends on the strength of the norm and the grade of separation between the two types. A strict social norm can be sustained for high levels of segregation or high institutional pressure. For intermediate social norms, cooperation can be an equilibrium for high separation and high integration, with high institutional pressure required for some parameter specifications in the latter. If norm strength is endogenized, i.e. the strength of punishment is positively correlated to the number of norm following types, two scenarios can be justified as equilibria. One with high segregation, high institutional punishment and strong norms. The other has norms of intermediate strength and little to no separation of types. In this case, conditional cooperation is commonplace, which is compatible with empirical findings.[81]

For social norm analysis in general, evolutionary game theory has contributed dynamic modeling possibilities to reflect some empirical findings on norms. According to Young (2015), the evolutionary models can account for the following empirical observations on social norms: First, norms are long-lived, even if they are not optimal. Second, if norms change they often do so swiftly. Third, combining the two above facets of social norms, one can illustrate historical norm change, where long standing norms were replaced with new norms rapidly, sometimes within one generation. Fourth, social norms, once established, prescribe behavior and limit the options individuals have and reduce the variety of observable behavior in a group. Finally, dif-

[80] Cultural transmission is a more richly layered learning mechanism than simple replicator dynamics. In fact, replicator dynamics can be integrated for one or several of the transmission channels employed in cultural transmission. The literature on cultural transmission usually differentiates one direct and two indirect transmission channels: direct vertical socialization of cultural traits through parents, indirect socialization through either horizontal transmission, i.e. for example through peers or colleagues, or oblique transmission, through not always obvious or straightforward societal processes (Bisin and Verdier, 2008). With regards to norms, this approach can highlight that norm dissemination is highly complex. Cf. Bowles (1998) for the influence of the design of economic institutions on values and norms, Bisin and Verdier (2001) for a seminal theoretical contribution, and Bidner and Francois (2011) for cultural transmission of trust norms. Cf. Bisin and Verdier (2008) for a review of the cultural transmission research.

[81] Cf. e.g. Frey and Meier (2004), Fehr and Fischbacher (2004).

ferent groups may have different norms, even if their socioeconomic conditions are comparable. With the evolutionary game theory apparatus, e.g. (Young, 1993), these empirical features of social norms can be illustrated. From an evolutionary perspective a long-lived norm is a stable equilibrium. For a sufficiently great shock or number of mistakes, however, play can escape an equilibrium's basin of attraction and a new equilibrium can be established after relatively few rounds of play. In any given equilibrium only some strategies are optimal or permissible, which reduces the variety of play. Finally, due to the idiosyncrasies of the evolutionary process, small deviations can land the dynamics in a different equilibrium, even if they started from a similar position. Furthermore, evolutionary dynamics depend on many small interactions, something some sociologists would agree to, thus relating two social sciences interested in human behavior.[82]

Despite these contributions to understanding the evolution and development of social norms, there are two difficulties with evolutionary game theoretical contributions. First, without sufficient reference to other social science accounts of the characteristics of specific, observable social norms, the approach risks being self-serving. Specify any regular form of behavior and show for which parameter constellation it is an equilibrium, and postulate this to constitute a social norm. Second, it has proven difficult to relate the models to case studies and other empirical work.[83]

V.3.8 Norm entrepreneurs

Contrary to an evolutionary game theory or a sociological approach to social norms that state that social norms are most likely the outcome of many diffuse social interactions, the literature on norm entrepreneurs posits that a ”norm is [...] the product [...] of the purposive actions of discrete individuals, especially those who are particularly suited to providing the new rule and those who are particularly eager to have it adopted”(Ellickson (2001), p. 2). Thus, particularly endowed or gifted people can enact norm change, which usually benefits themselves. We present three formal models of norm entrepreneurship, i.e. how (influential) actors can generate or shape norms: Corneo and Jeanne (1997), Kübler (2001), and Acemoglu and Jackson (2015).

Corneo and Jeanne (1997) concern themselves with the origin of social norms. They use Akerlof (1980) to model how a firm, as norm entrepreneur, might create a consumption norm for their good. The model posits a continuous time economy with a continuum of agents who can consume a good at each point in time. The good has no intrinsic consumption value. However, it has a reputational value, insofar as not consuming it may reduce an agent's reputation. Buying the otherwise worthless good,

[82] Cf. e.g. Berger and Luckmann (1966).

[83] Cf. Young (2015) for more details and successful examples.

however, one has to pay its price. The authors adapt the model in Akerlof (1980) in order to make reputation dependent on the purchase of the good. The critical update to generalize Akerlof (1980) consists in making the long-term change of the norm ($\dot{\mu}$ in subsection V.3.1) dependent on a function of the number of consumers who currently buy the good, instead of just the share of consumers directly. Allowing for the function to take different forms enables the results below.

Initially, the consumption good has no reputational value at all. Therefore, an infinitely lived producer needs to invest into the reputational value of the good in order to create a consumption norm. Corneo and Jeanne (1997) compare this problem of the firm with an optimal investment program, with the reputation of the good taking the place of capital. The firm has to balance short-term profits with the long-term reputational value of the good. Since such investment is costly, only firms with sufficient market power will try to establish a consumption norm.[84] Corneo and Jeanne (1997) show that an adequate intertemporal pricing strategy may be optimal for a firm with market power. They describe two possible consumption norms: bandwagoning norms, where the good is more attractive the more people already have it and its reputational value increases steadily in time as more people buy it, and snob norms, where the reputational effect of the good is strongest when only few people consume it in order to distinguish themselves from the mass. The reputational value of a snob good increases fast but falls off in time as more people buy it. Thus, according to Corneo and Jeanne (1997), a firm with sufficient market power can establish a norm. It is, however, not entirely clear, whether these consumption norms are really norms that guide behavior, or whether they would be more adequately described as fads and trend-setting.[85] Nevertheless, Corneo and Jeanne (1997) add to Akerlof's (1980) reputation model by allowing for more sophisticated long-term effects. The effects, at least with regard to snob norms, would be more adequately described as fads or trends. They thus provide a tool to investigate these phenomena. The requirements to act as a norm entrepreneur are quite high, as sufficient market power and financial strength for a risky long term investments strategy are required.

Kübler (2001) studies how influential parties, e.g. government agencies, NGOs, or interest groups, can change inefficient norms, where a change might improve welfare. A norm change can involve the abolition of an existing form of behavior or the establishment of desired behavior. Kübler (2001) illustrates two ways in which these changes can be enacted. The norm entrepreneur either changes the payoffs of the agents or they change the reputation gained by following a norm. Subsidies, punishments, and prohibitions are the proposed policy tools to erode an inefficient norm via the change

[84] With perfect competition and homogenous goods, an investing firm would only provide a positive externality to its competitors.

[85] The authors', in hindsight, telling example for such a consumption good are Pogs, a popular toy for children in the 1990s.

of the payoff route. This changes the relative prices by making the social stigma less stingy. Reducing the reputational payoff of a norm, on the other hand, involves making the inefficient norm appear unfit or inadequate for its purpose or discrediting the action the norm prescribes. Kübler (2001) builds both avenues of norm change into Akerlof's (1980) model. We start with the option to change payoffs by offering incentives. In the short run, the utility of agent i is given by

$$U(a_i) := (y - a_i\mu + m)A, \tag{123}$$

where $y > 0$ is the additional utility obtained by breaking the inefficient norm and pursuing an individually valuable enterprise, a_i is the agent's type of how much they care about reputation, distributed uniformly on the unit interval, μ is the proportion of people who believe in the norm, m is a material incentive the norm entrepreneur can provide, and A is a dummy variable, taking the value of one if the agent breaks the norm. Without reputation being important to the agent, they would always break the norm. As in Akerlof's (1980) model above, the higher a_i and μ the more disadvantageous breaking the norm becomes. If m is high enough, some agents prefer to break the code. In the long-run this reduces μ, thus leading to the erosion of the norm. Agents who break the norm are punished with loss of reputation but are compensated materially. This avenue to erode inefficient norms can therefore be costly. The second way of eroding an inefficient norm recommends reducing the reputational gain of adhering to the norm. The proposed intervention operates via the influence over a newly introduced parameter, the long term reputational value of a norm $v(x_t, r)$. The parameter r is the norm entrepreneurs' action influencing the meaning of the code. Kübler (2001) refers to advertising campaigns informing about the inefficiency of the norm or public interventions by influential people as examples for r. Equation (102) from subsection V.3.1 takes here the form

$$\frac{\partial\mu}{\partial t} = \delta(v(x_t, r) - \mu_t). \tag{124}$$

The parameter $\delta \in (0,1)$ measures the rate at which a change in norm strength will take place, i.e. an added measure of norm persistence. In order for the norm entrepreneur to be able to discredit the norm, Kübler (2001) assumes $\partial v/\partial r \geq 0$. It is acknowledged that this need not be the case and the entrepreneur may fail and, against their intentions, may strengthen the inefficient norm, i.e. $\partial v/\partial r < 0$. If the reputational value v is reduced below μ by reducing r, the norm erodes. There are two effects, the direct effect on μ and the indirect effect of a lower μ lowering the number x of people who actually follow the norm, thus reducing v. Regulating the meaning attached to a norm, reduces the reputational value of the norm, and alleviates the pressure to follow the norm. In the long-run the norm erodes. An advantage of the this approach to change a norm is that no agent suffers stigma due to reputation loss, unlike in the scenario with material incentives to violate the norm, where the agent is only materially compensated for the loss in reputation. Kübler (2001) analyzes how the two approaches to norm entrepreneurship work in different circumstances. The optimal policy mix depends on the specific norm that the norm entrepreneur wants to change. According to Kübler (2001), bandwagoning norms are eroded efficiently with changing the material incentives, snob norms by changing their meaning.

The model of Kübler (2001) is intentionally kept abstract to indicate the main possibilities for norm change in a model inspired by Akerlof (1980). The effects of the proposed policy instruments are not surprising. In case of reducing the reputation of the norm, the policy advice would have to be detailed further and designed adequately for the specific situation. The material incentives in the model have the desired effect of compensating agents for suffering a loss in reputation, but the channel appears overly simplified. Above, Bénabou and Tirole (2006) highlight that the interactions of different motivational factors might not be this straightforward. Thus, it might be the case that the desired effect is not achieved fully or that unwanted side effects may occur, making the task of norm change as a policy tool a sensitive one. There is hence no guarantee of successful norm change with the proposed strategies, and if norm change were actually observed empirically, it is unlikely that one could isolate the one measure that changed it. It is probable that a multitude of effects acts at the same time to change a norm. However, the model provides a bedrock of future research into norm change. Interestingly, recent empirical research on information campaigns to change norms lends support to the ideas in Kübler (2001) with regard to influencing the meaning of a norm.[86]

Acemoglu and Jackson (2015) contribute to the literature on social norms by highlighting the role of history and historical accidents in the development and change of norms and the role prominent figures can play as norm entrepreneurs. They model an overlapping-generations model, where each generation consists of one agent. They build their model around a simple coordination game with two strategies, high and low, and two pure-strategy Nash equilibria, one equilibrium where both players play 'high', which leads to a higher payoff for both players, and another equilibrium where both play 'low'. An example can be seen in Figure 11, with $\alpha, \beta > 0$.

A social norm in this setting exists, for example, when most players play the strategy 'high', which would constitute a norm of trust or cooperation. Acemoglu and Jackson (2015) are particularly interested under which circumstances this cooperative behavior obtains. In this society, agents live for two periods. In the first period they are young

[86] Aloud et al. (2020), Grewenig et al. (2020), Bursztyn et al. (2020b), as well as Jayachandran (2020) investigate how gender norms influence labor market outcome differences between men and women. All four contributions are reservedly optimistic that policy intervention, i.e. information about misperceived social attitudes, can be effective in attenuating labor market gaps between men and women in their respective contexts.

		Player 2	
		High	*Low*
Player 1	*High*	β, β	$-\alpha, 0$
	Low	$0, -\alpha$	$0, 0$

Figure 11: A simple coordination game, cf. Acemoglu and Jackson (2015), p. 431.

and in the second period they are old. Their payoff depends on the agent's own action and on the action of the previous and the next generation. Every agent plays the same strategy in both periods and decides at the beginning of the first period what their strategy will be. An agent born at time t will have total utility of

$$(1-\lambda)u(A_t, A_{t-1}) + \lambda u(A_t, A_{t+1}), \tag{125}$$

where A_t is the action of the agent born in t, A_{t-1}, and A_{t+1} are the actions of the agents born in $t-1$ and $t+1$, respectively, and $\lambda \in [0,1]$ is a weighing factor of how the agent born in t values interaction with the other two generations, i.e. how forward looking the agent is. Agents are heterogeneous in two independent aspects. First, with regard to how they choose which action to play, and second, with regard to their prominence. There are two types of agents when it comes to action selection. With probability $1-2\pi$ agents choose their action by maximizing (125), while with probability 2π it is a dominant strategy for agents to always play either 'high' or 'low', with $\pi \in (0, \frac{1}{2})$. With respect to prominence, agents can either be prominent, i.e. an important, well-known figure, or not. Agents are prominent with probability q and not prominent with probability $(1-q)$, with both types of agents assumed to be present, i.e. $0 < q < 1$. All four combinations are possible. For example, a prominent agent who always plays 'high' has a probability of $2\pi q$. The information structure is as follows. The agent at t knows some of the history of the past play up to $t-1$. Specifically, they have perfect information about the actions of all previous prominent agents. They do not know the actions of non-prominent past agents, only how many there were. Furthermore, agent t receives an imperfectly informative signal, s, about the action A_{t-1} of the agent in the previous period, unless, of course, the previous agent was prominent, in which case the signal is perfectly informative. Agent t updates their belief about the likely play of the non-prominent agent $t-1$ with the signal s they receive.

A Bayesian equilibrium[87], in this setup, is a profile of the strategies of the players without dominant strategies and a stipulation of beliefs, dependent on the possible histories and signals. The strategy of a player without a dominant strategy is a best response to this profile, given their beliefs and prominence. Formally, the best response

[87] The set of Bayesian and perfect Bayesian equilibria coincides in this setup.

of such an agent will be to play ’high’ if and only if

$$(1-\lambda)\phi_{t-1}^{t}+\lambda\phi_{t+1}^{t}\geq\gamma, \tag{126}$$

where ϕ_{t-1}^{t} is the equilibrium probability that t gives to agent $t-1$ having played ’high’ in the previous period. Similarly, ϕ_{t+1}^{t}, for $t+1$, which, whats more, also depends on agent t’s play at time t. γ measures how risky it is for agent t to play ’high’. To be precise, γ gives the probability that the other player should also be playing ’high’ that makes agent t indifferent between playing either ’high’ or ’low’. In the following, we will concentrate on γ as the central unit of analysis.[88] As usual in these kind of models with a large number of possible equilibria, Acemoglu and Jackson (2015) focus on a tractable subset of possible equilibria, here, greatest equilibria. They find that all equilibria are cut off equilibria, where the choice of action depends on the size of γ, λ, π, and q. Play will, in general, depend on the choice of the last prominent player. For example, fixing the other parameters, there is then a threshold of γ for below which it is optimal for players without a dominant strategy to play ’high’, irrespective of their signal. This equilibrium is called a social norm for these players to play ’high’. Above the threshold level of γ, playing ’high’ becomes riskier, and always playing ’high’ is not an equilibrium strategy anymore. Increased risk of meeting a player who plays ’low’ means that players without a dominant strategy shortly following a prominent player who played ’high’, will still play ’high’, but the further back the prominent player is in time, the more likely it gets that the player in the previous generation plays ’low’. For a sufficiently high γ, players always play ’low’, since there is always a risk of meeting a player with the dominant strategy to always play ’low’, and the possible benefit of playing ’high’ is too unlikely. Summarizing, for different thresholds of γ, different play in equilibrium occurs. There are two social norms, where players choose the same action as the last prominent player, irrespective of their signal. In addition to the ’high’ social norm above, a similar argument can be made for a social norm to always play ’low’.

The work of Acemoglu and Jackson (2015) highlights the role of history for norm maintenance. A social norm is determined by the act of a prominent player and lasts for some (possibly long) time, depending on parameter specifications. The norm can be changed with the arrival of a new prominent player, who may select the opposite action. In the intermediate parameter space, where no social norm exists, some signals may make it highly likely that a previous player played ’low’. This is due to a monotonicity result, which states that cutoff values are either always non-increasing or non-decreasing. In the frame of the model, the likelihood that the player in the previous period played, for example, ’high’ decreases monotonously with distance from the prominent player who played ’high’. There are some signals for which it is likely

[88] Since the notation is already involved at this point, the following will try to specify the main ideas without introducing more variables.

that the previous agent had a dominant strategy of playing ’low’, or that one of the previous players received a highly indicative signal to play ’low’. This fact will lead to requiring a stronger signal of ’high’ play in the previous period to convince an agent to follow the signal. This effect does not only influence choice with respect to the past, but agent t also has to consider what signal will obtain in the next generation, and how this next generation will interpret it. There is thus a dynamic where, for intermediate parameter values, play will switch along the equilibrium path from the last prominent agent’s action to the opposite and back.[89] Norm change, for example, from a social norm of ’low’, is possible for prominent agents without a dominant strategy if their signal of play in the last generation indicates that they might be playing an agent with a dominant strategy to play ’high’. For a sufficiently strong signal, such a prominent agent will find it in their interest to play ’high’, which will break the norm to play ’low’ and lead to a future social norm to play ’high’.

Acemoglu and Jackson (2015) provide an accessible model, where simple building blocks fit together to generate an initially complex but eventually very tractable model. They highlight the importance of history for social norms, as agents take the noisy information they receive and orient themselves on historic events that give them additional guidance. Norm entrepreneurs are still specially gifted. They face lower risks than other players because they know that at least the following generation will know for sure what they played. A social norm is here modeled as a backdrop, as a frame that orders history and shapes agents’ interpretation of their signals. This orients or even determines choice in the present, despite the fact that the decisions of the distant past are irrelevant for payoffs today. The second contribution is the ability of the model to represent norm durability as well as behavioral dynamics without a norm. Norm change can occur through new prominent figures, who find it in their interest to act.[90]

What all reviewed models of norm entrepreneurship have in common is that the challenges inherent in social norm change are largely faded out. For example, Corneo and Jeanne (1997) assume that firms invest successfully into their goods’ reputation, without clarifying what that entails. Even a well-planned investment may still fail if circumstances prevent that people actually value the reputation the good conveys. Kübler

[89] This dynamic of ebb and flow, from playing ’high’ to playing ’low’, is different from the dynamic in evolutionary models. Here, this dynamic will occur in the intermediate parameter space with optimal play. In evolutionary models, first, change from one equilibrium to the next requires mistakes or shocks. Second, further mistakes or shocks in the new equilibrium need not bring the system back to the original equilibrium.

[90] On the importance of history: Hoff et al. (2011) find that lower caste members in India are less willing to punish norm violators of a cooperation norm that hurt members of their own caste. This reduced capability (Sen, 1993) to punish norm violators inhibits their ability to reap the benefits of well working social norms of cooperation, since they find it harder to maintain them. They are, for example, less capable to enforce contracts, trust is reduced, and opportunism cannot be reigned in as well, which allows higher castes to employ a strategy of divide and conquer in order to exert control and stay in power. This highlights the importance of history and context when analyzing norms and their possible effects in society.

(2001) is optimistic that information campaigns can engender norm change. However, it is not a given that the effects of the intervention will be as strong as planned or in the right direction, or will not engender unwanted side effects (Bird, 1999). Acemoglu and Jackson (2015) circumnavigate this issue by selecting the simple coordination game, which allows for excluding these problems, at the cost of realism. In all three models, norm entrepreneurs have special knowledge, skills, power, or stand to gain particularly much from a norm change. The level of abstraction is high and the actual strategies stylized. However, this shortcoming should not be overstressed. These models are highly illuminating because people actually intent to form and change social norms and sometimes are successful in doing so. Furthermore, these models are capable of drawing up important features, relevant for understanding societal processes. They provide ideas for shaping political agendas but at the same time warn the would-be norm entrepreneur about the difficulties they face.[91]

V.3.9 Public goods and positive externalities

In several models discussed above, the main concern was the punishment an individual would incur if they violated the norm. In this subsection we will detail two models, where a positive contribution to a public good is the norm, and how positive externalities may play a role for norm conformity.

Rege (2004) analyzes the role norms based on social approval can have for the voluntary provision of a public good. In this model, a continuum $[0,1]$ of agents can contribute to a public good. Agent i's contribution to the public good is denominated by g_i. The agents can choose to have one of two types. Either be a contributor to a public good, $g_i = 1$, or a non-contributor, $g_i = 0$. After having made their decision and having contributed accordingly, the agent meets other people in the society, who know what action the agent took. If the agent did not contribute and they believe to meet a contributor, they lose utility. If the agent contributed they gain reputation, if they believe to meet another contributor. Believing to meet a non-contributor does not change the reputation of either type of agent. In addition to the reputation effect, people gain utility from private consumption and consumption of the public good. The utility function for i is

$$U_i = c_i + q_i + w(\bar{g}), \tag{127}$$

where c_i is private consumption, q_i is the reputation effect, and $\bar{g}$ is the average contribution to the public good, with $w' > 0$ and $w'' < 0$. When i meets m people, the effect of reputation on their utility is the average of the reputation gain from each meeting,

[91] In contrast, Offiaeli and Yaman (2021) report on the effectiveness of the introduction of waiting zones on train platforms to reduce waiting times and increase capacity. They judge this simple nudging technique to be equivalent to establishing a social norm without much difficulty.

i.e.

$$q_i = \frac{1}{m} \sum_{j \in m_i} \lambda(g_i - \bar{g}) E_i(g_j). \tag{128}$$

$\lambda > 0$ measures how much contributing can potentially benefit each person, while $E_i(g_j)$ is i's expectation with regard to the contribution of j. The agents' contributions constitute an externality on each other, i.e. $(g_i - \bar{g})$. If the average contribution, $\bar{g}$, in society rises, the reputation effect for each person is reduced. Rege (2004) continues with assuming viscosity[92] in the meeting of others, i.e. contributors are more likely to meet a fellow contributor. Having provided only the details relevant for the modeling of social norms here, Rege (2004) shows that this game is a coordination game with three Nash equilibria for some viscosity parameters, one in which everybody contributes, one in which nobody contributes, and one in which a portion of society contributes to the public good. Using evolutionary game theory, Rege (2004) shows that only the full contribution and the no contribution equilibria are asymptotically stable. The one where everybody contributes is Pareto superior to the one where nobody contributes. The third equilibrium is not stable but serves as a tipping point. Above its value of contributors, the Pareto superior equilibrium obtains in the evolutionary process and vice versa. The Pareto superiority is not due to the reputation effect but to the positive effect of the provided public good, since if everybody contributes the same amount, nobody gains social approval. The logic why the two equilibria are stable is the following. If non-contributors meet relatively many contributors, the reputation gain from changing to contributing is large. This will lead to everybody being a contributor. For a similar argument, the non-contribution equilibrium obtains. Rege (2004) further shows that subsidies for the public good may move a society to the Pareto superior situation because gaining reputation becomes cheaper and more people find it beneficial to contribute. If the subsidy is in place long enough, such that the portion of contributors is high enough (crowding in), the evolutionary dynamics will drive society to the desirable state, even if the subsidy is removed. Contrarily, if the government provides the public good, they crowd out voluntary participation and this may drive a society to the less desirable equilibrium state.

Rege (2004) states that the effect of conforming to a norm can also be positive if there are positive externalities in doing so. The motivation to contribute to a public good stems from a desire for social approval. Rege (2004) explains that if a social norm for contributing to the public good is enforced (she does not explain what that entails), then everybody contributes. Unfortunately, Rege (2004) does not model this norm explicitly. She does not need it for her formal analysis and she instead interprets

[92] Introduced by Meyerson et al. (1991) for biological applications: close relatives are more likely to meet than distant relatives. Adapted to social contexts, peers are more likely to meet each other than non-peers. Technically, with some probability, the type of a players' opponent is drawn from the same peer group, and otherwise is drawn from the overall population, cf. Rege (2004).

the outcome of either full contribution or of no contribution as equivalent to a social norm. These equilibria as social norms, however, are obtained without ”enforcement” and do not require enforcement to be maintained. Norms are hence the equilibria in a coordination game in this setup.[93]

In Nyborg et al. (2006), the focus rests on internalized moral motivation to behave socially beneficial. They thus focus on other motivational motives, different from reputation, important for social actions. They highlight that people may have an internal norm of ideal behavior they would follow if everybody else did so, but in a world of free-riders may find that it is not their responsibility to act in a socially preferable way. They behave only conditionally prosocial. The rewards of moral behavior are internal in the form of an improved image of oneself and depend on two factors. First, the agent believes that their action constitutes a positive externality for others and, second, that the socially beneficial action is prescribed by a norm that is, in general, observed in the society. In other words, the agent has to understand that their actions can have positive consequences for others and that this makes the norm of societally beneficial behavior applicable. Whether or not the agent will obey the norm depends on whether they feel personally responsible to take the action and bear the costs the norm imposes. In Nyborg et al. (2006), the feeling of responsibility has its roots in two concepts. First, descriptive norms, i.e. the perception of how other people actually behave in some situation, irrespective of whether this action is actually socially accepted. Second, responsibility also stems from fairness and reciprocity concerns. The responsibility to contribute in the socially desirable norm prescribed action is then likely to increase, the larger the portion of people in the society that actually follow the norm.

Nyborg et al. (2006) model a society of a large finite number of agents n, who have self-image concerns S. Each agent has the option to either buy a normal consumer good or one that is environmentally friendly i.e. green. The latter is more expensive, with additional costs C, but it provides a small benefit b, both for the agent and all other members of society. It is assumed that $b < C$, i.e. it is not individually optimal to buy green without concerns for self- image, and $nb > C$, i.e. the positive effect on society from the agent buying green outweigh their costs. Choosing the green good is assumed to be considered morally superior by all members of society. Buying green thus leads to an individual benefit p_i of

$$p_i = (S + b - C)x_i = (S - c)x_i, \tag{129}$$

where x_i is a dummy variable, with $x_i = 1$ if the agent buys green. The net costs of buying green are $c = C - b > 0$. The perception of agent i of themselves, S, depends positively on the size of the positive externality for the other members of society from

[93] Bardsley and Sausgruber (2005) add that if others contribute more, the desire for conformity contributes about one third to the above average contributions in experimental public good games.

the agent buying green, $B = b(n-1)$, and how much they feel that buying green is their personal responsibility, which depends positively on the portion of society a that also buys green, $a = \sum x_i / n$, i.e. the descriptive norm of behavior in society. Therefore,

$$S = s(B, a), \tag{130}$$

with $s(\cdot)$ a concave, increasing, and differentiable function. An agent chooses to buy green if they believe that their self-image improvement is larger than their net costs,

$$s(B, a) > c. \tag{131}$$

In contrast to reputation models, where the fear of external punishment maintains conformity, here only internal perceived norms of behavior guide the contribution to the socially beneficial cause.

Nyborg et al. (2006) simplify their analysis in the following. They assume that if no descriptive norm of buying green exists, i.e. no rule is believed to exist because one cannot observe anybody else buying green, then the agent will not buy green. Technically, this assumption can be expressed as $s(B, 0) < c$, i.e. if no one else buys green, the social image improvement of buying green does not cover its cost. No one buying thus constitutes a Nash equilibrium. Contrarily, if the share of people buying green increases, self-image concerns become more important. Technically, $s(B, 1) > c$. Everybody buying is therefore identified as a second Nash equilibrium. The form of the game can hence be seen as a coordination game, where either no one buys, or everybody buys green, the latter being Pareto superior. Additionally, there is a Nash equilibrium in mixed strategies where only a share of society buys green. As in Rege (2004), a dynamic analysis with the help of evolutionary game theory reveals that only the two pure strategy Nash equilibria are asymptotically stable and the mixed strategy equilibrium serves as a tipping point. For values of a above the one specified in the mixed equilibrium, the Pareto superior equilibrium obtains and vice versa. Therefore, adequate taxes or subsidies may change behavior in society.

Nyborg et al. (2006) propose that information campaigns, highlighting the beneficial societal effects of responsible consumption, B, or informing on the actual share of people engaging in green behavior, a, may be a cheap, effective policy intervention. Nyborg et al. (2006) also point out that real world attempts at such information campaigns, aimed at influencing people to embrace more environmentally friendly behavior, have met with mixed results, sometimes engendering a behavioral change and sometimes not. They argue that if the externalities are perceived to be low, then a high perceived level of behavior is required to land in the Pareto superior equilibrium. Additionally, since in this model only perceived behavior of others enters the consideration of the agent, Nyborg et al. (2006) state that it is not certain that an actual real life change of behavior in society would have any effect on an agent's action. That is because, even if others buy more green, and the tipping point for parameter a in society is reached, this does not entail that the perceived norm of the agent is changed, i.e. they believe parameter a is still lower and do not change their behavior.

The two models with positive external effects required no external punishment in order to show that a norm of prosocial behavior can be an equilibrium. Rege (2004) models a reputation effect that depends on a simple threshold specification, but the strength of the norm depends on the beliefs of the agent concerning the likelihood of meeting a contributor. The higher the likelihood, the more likely is contributing, strengthening a social norm of contributing. Nyborg et al. (2006) show that an internalized norm to contribute to others' welfare can, unsurprisingly, be effective if the size of the externality is large enough. In addition to this internalized injunctive norm, they add the perceived descriptive norm a. Interestingly, the two norms can work against each other, highlighting that social norm interaction is difficult to predict. The fact that the two models can maintain a social norm with positive contributions as one possible, stable equilibrium, may come down to both setups effectively reducing to coordination games. In a 2×2 coordination game, both pure strategy Nash equilibria can be justified as a social norm, since not coordinating is already costly enough for the players. Which one obtains, is coincidence, even if one is Pareto superior (Cf. Ullmann-Margalit ([1977] (2015), p. 89.). The contribution here is that this static insight is applied in a dynamic environment.

V.3.10 Norms in the welfare state

All the models in this subsection postulate that a norm to work hard and to live off of one's labor exists in society. This norm puts social pressure on people who rely on transfers to make a living.[94] In front of this societal backdrop, welfare benefits, especially unemployment benefits, are investigated in the following.

As such, Lindbeck (1995, 1997, et al. 1999) postulates that there exists a norm to work hard and to be able to support oneself instead of relying on transfers to make a living, thus making it socially less attractive to be unemployed. The individual has a choice to make: either work, receive a wage, and pay taxes that finance the transfer scheme of the welfare state, or do not work, receive a transfer, and face the ignominy of not working. The norm to work is exogenously given, the strength of the norm, i.e.

[94] Empirical support for this rather stylized norm comes from a panel study by Clark (2003), which shows that the amount of social pressure associated with becoming unemployed has an impact on a persons' well-being and how likely it is that the person finds a job quickly. If many people in the close environment of the unemployed are also currently out of a job, the negative effect of unemployment on the person's well-being is smaller and the pressure to find a new job is reduced. Stutzer and Lalive (2004) add, with a field study in Switzerland, that a stronger norm to work reduced the life satisfaction of the unemployed and increased the speed with which they found a new job.

how much non-compliance affects the individual negatively, depends on the number of unemployment benefit recipients in the society. If many people are on welfare, the sting of not working is weaker. Formally, Lindbeck's (1997) model introduces the following tradeoff for individual i:

$$u([1-t]w_i) > u(T) + \mu - v(x), \tag{132}$$

with $u(\cdot)$ the utility of income, either from working at the wage rate w_i and paying taxes at rate t or from receiving a lump sum transfer T. μ is a factor that captures the difference in utility gained from leisure and the intrinsic utility of working, and x is the share of people on welfare in the society. The function $u(\cdot)$ is increasing in its arguments, and the function $v(\cdot)$ is decreasing in x, capturing the idea that if more people are on welfare, the less stingy the social punishment for disobedience of the norm becomes. Transfers are financed only through taxed income. When making a decision on whether to work, an individual takes the tax rate, the size of the lump sum transfer, and the share x as given. Assuming that individuals differ only with respect to their wage rate, distributed according to some cumulative probability distribution function, and that the functions comply with some standard mathematical properties (e.g. continuity), standard arguments reveal that there exists a unique cut off wage rate that solves equation (132).[95] Those with higher wages work, the others rely on the transfer. In equilibrium, people expect there to be exactly a share x of unemployed people and make their decisions accordingly, which leads to the actual realization of x. In the more detailed analysis of Lindbeck et al. (1999), it is shown that for every level of x there exists a unique pair of tax rate and transfer that equates income from taxes and expenditure for transfers, leaving a balanced budget. If people in society can vote on the tax rates, Lindbeck (1997) demonstrates, two political equilibria (where no other equilibrium with a balanced budget is preferred) with selfishly motivated agents exist. One where everybody works and taxes and transfers are zero, and one where a majority lives off benefits. The full working, no taxes equilibrium can be converted into a low-taxes, some transfers equilibrium if one allows for either some altruism towards individuals who are worse off than oneself, or by allowing for the possibility for individuals to lose the ability to work.[96]

The two generation model in Lindbeck and Nyberg (2006) is interesting because it explicitly models the attempt by parents to instill the norm to work hard in their children, and how welfare benefits may change their incentives to do so. Parents, in a first

[95] Equation (132) must of course be written as an equality.

[96] In Lindbeck et al. (2003), individuals with equal productivity and thus equal wage differ with regard to how much they value leisure. There exist thus two subgroups in the population share x. Those who do not want to work and those who currently cannot find a job. There now exists a cutoff value for the parameter measuring the prevalence for leisure. All individuals with a higher value will not look for a job, all individuals with a lower value will look for a job.

step, try to instill norms to work hard in their child. In a second step, the child then has to exert costly effort on the labor market to either get a well paying job or a low paying job. The job the child gets depends on their effort and on random events. The two wage levels of the two jobs are fixed. Parents care about their child's well-being, while children are selfish. In a final step, after having observed their child's success on the labor market, the parents decide whether they want to give a transfer to the child. The child's welfare is additionally insured by a more or less liberal social insurance scheme, which provides a fixed transfer for those in the low paying job. As before, the social norm to work hard leads to disutility if an individual depends on transfers from the social insurance. The utility of a parent is composed of the parent's own consumption and the weighted utility of the child, i.e. the parent's altruism. Using backwards induction to solve the game, parents, in the third step, maximize their utility with respect to the transfer they provide to their child. The child then decides how much effort to exert in order to get a high paying job, considering the likely support of the parents and how much the social stigma of a low paying job and a transfer affects them. In the first step, parents can influence how sensitive their child is with regard to adhering to the norm, but this is costly. They maximize utility, taking the child's presumed effort choice in the next round into account. Lindbeck and Nyberg (2006) identify three possible cases, depending on the parents' income, the result of the child on the labor market and the strength of the parents' altruism. Parents do either not support their child at all, they only assist a child if it failed on the labor market, or they always help their child. Higher social insurance transfers lower the incentives for children to extend efforts and lower the incentives for parents to instill norms to work hard in their children if they provide a positive support transfer to the child. Social insurance leads here to free-riding by the parents, who want other parents to instill the norm to work hard in their children, but lower their own efforts to inculcate the norm to work hard in their own child.

In Dufwenberg and Lundholm (2001), the question is again whether or not to take up a job. In order to get a job, the individual has to exert effort x. The extension of the model here is that the individuals differ with respect to their talent t, and that the social norm applies to each individual according to their talent.[97] An agent's probability to get a job is the product of their talent and their effort, tx. There is only one kind of job with wage normalized to one, and there is a fixed social unemployment transfer, irrespective of effort or talent. An agent's material payoff when there is no concern for social rewards for norm compliance is

$$U(x,t) = tx + (1 - tx)u(\beta) - \frac{K}{2}x^2, \tag{133}$$

where u is assumed to be a strictly increasing and sufficiently concave function, β is the

[97] Talent is distributed according to a strictly increasing, continuously differentiable distribution function with respective density function. The functions are common knowledge.

unemployment transfer, and K is a positive parameter to insure that effort is always smaller than one. The agent's talent is private information. The interesting feature of Dufwenberg und Lundholm (2001) is that the social norm in this model does not prescribe the same behavior for all people because society expects the more talented to apply themselves more. In particular, the social norm applying to an individual of talent t requires them to exert at least as much effort as they would have exerted if there were no social transfer. This level of effort is found by maximizing the material payoff of a player with talent t with respect to effort, which is

$$x = \frac{1}{K}[1 - u(\beta)]\,t. \tag{134}$$

The authors assume that the effort choice can be observed by some close neighbors. The level of talent is then endogenously inferred from the observed effort choice and labeled τ. Therefore, the social benefit for a player of inferred talent τ is given by

$$\sigma\left[x - \frac{1}{K}[1 - u(\beta)]\tau\right], \tag{135}$$

with $\sigma > 0$ a sensitivity parameter toward social rewards. Social benefits accrue to people who exert more effort than their inferred talent demands. Note that since the norm asks for people to exert effort according to their talent, each different talent should exert a different effort according to the norm. In sum, the player of talent t has the utility function

$$U(x, t, \tau) = tx + (1 - tx)u(\beta) - \frac{K}{2}x^2 + \sigma\left[x - \frac{1}{K}[1 - u(\beta)]\tau\right]. \tag{136}$$

With some restrictions on the talent inference function, Dufwenberg and Lundholm (2001) find that there exists a unique separating equilibrium for this signaling game, where all but the highest type of talent shade their talent to be held to a weaker social norm, i.e. the inferred τ is lower than t. With respect to the welfare state, they find that effort in a model with social rewards is higher than in a benchmark model without social rewards, which lowers the impact of moral hazard. Interestingly, increasing the unemployment transfer has contrary effects for high and low talents. There are two effects of a higher transfer. First, a higher transfer lowers effort for all players. The most talented players profit the most from this because it is easier (cheaper) for them to turn effort into employment probability. The second effect of a higher β is that it is now less profitable to imitate a less talented player because the impact of the perceived talent τ on the social norm, to which one is held, is reduced, cf. (135). Therefore, all but the highest talent can reduce shading their effort without being held to a tougher social norm. This increases their employment probability and hence their utility. For the players with the lowest perceived talent, this second effect can dominate the first effect, i.e. an increase in the transfer increases their effort provision. Despite these interesting results, Dufwenberg and Lundholm (2001) are cautious with regard to the welfare benefits of social rewards. First, because low level talents exert themselves too much and do not increase their chances to be employed sufficiently to justify the exertion. A society with many low talent individuals may therefore not benefit. Second, they find that high talents gain systematically higher social rewards than low talents.

Dufwenberg and Lundholm (2001) chose a different heterogeneity feature from Lindbeck and his co-authors above. In their setup, the social norm is the same for everyone, but since the norm refers to an individual characteristic to determine social rewards, its consequences are unequally severe for the same action for two individuals of different talent. Modeling type dependent norms is a unique contribution to the literature on social norms. They furthermore show that people might try to avoid to conform perfectly to the norm specified behavior if information about their type is imperfect. This is akin to the model of d'Adda et al. (2020) in subsection V.3.4, where people exploited the fact that the information about the precise level of the norm specified behavior was imperfect.

With regard to this subsection, Lindbeck and his co-authors have analyzed a social norm according to the idea of Akerlof (1980).[98] Any norm gains strength the smaller the group of people who do not adhere to it is. The contribution consists in linking this basic insight of economic analysis of social norms to questions of how to finance social transfers. The norm prescribes to extend effort to find a job. Here, the strength of the norm to work depends on the share of unemployed people in society. Lindbeck and his co-authors have considered different productivity of the agents, or different characteristics, and incorporated the idea into a dynamic game setup. The modeling of the social norm has varied little. Dufwenberg and Lundholm's (2001) contribution is a unique extension of the basic Lindbeck (1997) model. It hints at norms being applicable discriminately to different people.

[98] For additional contributions following Lindbeck see e.g. Weibull and Villa (2005) and Traxler (2010). Weibull and Villa (2005) employ the Lindbeck et al. (1999, 2003) model with regard to crime and punishment. They replace the norm to work hard with a norm against committing crime. The more people work legally, the stronger is the negative utility effect of the norm. Additionally, crime is a negative externality for all agents. They show, for example, how a higher income tax or a change in the strength of punishment can affect the crime rate. Traxler (2010) uses the Lindbeck et al. (1999) model to investigate tax evasion. The social norm is to comply with the tax code. The strength of the punishment for tax evasion depends on the perceived share of people cheating on their taxes, the sensitivity towards norm compliance, and the amount of income concealed. Traxler (2010) specifies the optimal level of tax evasion in the social equilibrium, given the size of punishments, the likelihood of detection and the effects of a tax increase. Among the interesting findings the model proposes are, first, that a high compliance rate despite low prosecution levels is compatible with a social norm prescribing tax compliance, and, second, that the perceived behavior of others is very influential for tax compliance, potentially allowing for policy tools to influence these perceptions.

V.3.11 Social Distance

Models of social distance maintain that it is easier to uphold cooperation with people that are socially closer, i.e. family or friends, than with a total stranger. This leads to social norms that are applicable with socially close people but not in interactions with strangers. It can also lead to the perpetuation of certain behavior from one generation to the next, which is detrimental in case of inefficient norms.

Tabellini (2008) focuses on internalized values parents instill in their children. It is assumed that there exists a norm of cooperation that parents teach their children. In this society, people are uniformly distributed on a circle[99] and randomly matched with another player. The matched players observe their distance from each other and play the Prisoner's Dilemma game once. The idea is that it is not reputation that can sustain cooperation, since they play the game only once, but that the values they learned from their parents help maintain cooperation, because players gain a non-monetary benefit if they cooperate, irrespective of how the other player plays. This benefit, however, decays with distance, such that it is much easier to maintain cooperation with close players than with players further away. The model thus highlights that social norms of good behavior, albeit Tabellini (2008) only refers to them as internalized values, may only be applicable in a certain subgroup, or among friends and family, and that they do not apply to more distant individuals. The contribution thus is that subdivisions of a society may have their own social norms.

In a second model on social distance and social norms, Akerlof (1997) distinguishes between conventional economic decisions and social decisions. Whereas the former do not affect other people, the latter do. Social behavior, in this view, is intricately linked to externalities and therefore often leads to suboptimal allocations. Players in the model have to choose a decision variable x and each choice of x carries an intrinsic value. Additionally, players occupy different places in social space. The closer two players are to each other, the more beneficial trade will be for both. The social location at a point in time is conditional on the choice of x and conditional on the inherited social position from the past. Finally, people take the positions of other players in social space as given when they make their decision. Employing a modified gravity model to analyze the stylized facts above, Akerlof (1997) can explain the existence of social subgroups with their particular norms. People's intrinsic, individually efficient choices of x may be undermined by incentives to choose an x to conform to other players close to their inherited social position. Moving the social position reduces the benefits of trade with formerly close players. If this loss is sufficiently large to not pursue more beneficial but distant enterprises, otherwise efficient trades are not realized. This can be interpreted as a social norm which makes it difficult to distance oneself from one's

[99] Cf. Salop (1979).

upbringing. This may entail repeating the same choices prevalent in one's social subgroup and forgoing opportunities that are more distant in social space.[100] Therefore, the model hints at the fact that small social groups may have their own social norms and that it can be difficult to escape even a socially harmful social environment and its norms. In this respect, inefficient norms can be sustained even if they are disadvantageous for individuals and can have adverse effects for society.[101] The two models on social distance postulate that norms can be subgroup specific and that they are tied to the social environments they were formed in.

V.4 Implications of empirical work

Although this review is focused primarily on theoretical contributions to the modeling of social norms in economics, we have already integrated models based on mainly empirically oriented papers (e.g. in subsection V.3.3). These have used a relatively simple specification, which is not meant as a critique, since the models in these papers serve a different purpose of justifying, orienting and buttressing the respective empirical investigation. In the first part of this section, we are looking at further specifications of models used in empirical work. Despite their simplicity, it is important to be aware of their structure, since they guide a lot of interesting experiments and field studies regarding social norms, which have always enriched and inspired further theoretical modeling. In a second subsection, we consequently identify topics in empirical discussions which have so far, to our knowledge, been insufficiently treated by theoretical modeling and therefore present avenues for future research.

V.4.1 Further theoretical models in empirical work

In their influential paper[102], Krupka and Weber (2013) aim to model injunctive norms. The power of the social norm to induce other regarding behavior is due to the willing-

[100] Case and Katz (1991) show that disadvantaged youths have a high probability of becoming disadvantaged adults. These neighborhood and family effects can be interpreted as learned social norms, that guide behavior, attitudes, and outlooks on life.

[101] For experimental contributions to social distance research see e.g. Hoffmann et al. (1996) and Bicchieri et al. (2022). Hoffmann et al. (1996) investigate giving in dictator game experiments. By game theoretical standards, dictators give too much. By increasing the social distance in their treatments, the authors can lower contributions by dictators, but two thirds of subjects still provide shares that are difficult to reconcile with game-theoretical predictions. This indicates that social distance may reduce norm compliance for some players. Bicchieri et al. (2022) are concerned with norm erosion if a norm follower observes non-compliance. They report experimental results in which observing an anonymous player taking money from a charity increases the likelihood that the observing player breaks the norm as well and takes money from the charity. Observing anonymous donating behavior in line with the norm does not increase observers' contributions to the charity. However, social proximity by creating a group identity alleviates the effect, insofar as observance of norm compliance now positively affects observers' donations, thus implying that social proximity is important for norm compliance.

[102] Cf. Kassas (2018), chapter 3.

ness of others to punish norm transgression on the one side and due to the positive (or negative) emotions one feels by conforming to (or disobeying) a norm on the other side. The model posits a set of K actions represented by $A = \{a_1, ..., a_K\}$. A social norm is a collective judgment by the community of what action constitutes good behavior and what does not. Technically, this is modeled as $N(a_k) \in [-1, 1]$. Adhering to the norm implies $N(a_k) > 0$, breaking the norm implies $N(a_k) < 0$.[103] Making use of an interval for the strength of norm compliance constitutes a new and interesting modeling choice. This allows for leveled reactions, i.e. punishments or rewards, dependent on how severe the transgression is or how well the norm has been fulfilled. This is in line with the critique in Romer (1984). It acknowledges the fact that the severity of punishment may correspond to the gravity of the transgression. The utility function in this setup is hence

$$u(a_k) = V(\pi(a_k)) + \gamma N(a_k), \tag{137}$$

with the increasing function $V(\cdot)$ representing how much the agent values material rewards π associated with action a_k. Furthermore, $\gamma \geq 0$ determines how important norm compliance is for the agent. The work of Krupka and Weber (2013), despite its simplicity, has become influential in experimental economics, due mainly for its norm elicitation procedure.

The work of Krupka and her co-authors (Krupka and Weber, 2003; Krupka et al. 2017), together with Kessler and Leider (2012) and Abbot et al. (2013), from subsection V.3.3, as well as d'Adda et al. (2020), from subsection V.3.4, can be seen as variants of a widely used, influential, standard theoretical approach in empirical research on social norms in economics.[104] The utility of an agent depends on their concerns for material payoffs (with an adequate function, e.g. $V(\cdot)$ in 137), and a measure of how well the agent conforms to the social norm, either like in Krupka and Weber (2013) or, even more prominently, with a norm as a threshold, i.e. $g(x - \bar{x})$, where x is the action taken by the agent, and $\bar{x}$ is the level of the action required by the social norm. The function g translates the norm (usually) transgression into disutility. Additionally, a parameter is introduced that measures the agent's sensitivity with regard to the effect of norm compliance or transgression, γ in (137). Different types of agents can be distinguished by different sensitivity parameters. Finally, other factors on utility that are of interest

[103] The case $N(a_k) = 0$ is unfortunately not discussed. One main change to the theoretical model in Krupka et al. (2017) is that instead of an interval for $N(a_k)$, just the two cases of compliance, $N(a_k) > 0$, and non-compliance, $N(a_k) < 0$, are given. A "neutral" action is not possible. A neutral effect on utility, however, can still be modeled by using γ, cf. equation (137) below. Additionally, Krupka et al. (2017) make the material payoff also dependent on the other players' actions.

[104] There are also contributions in experimental economics concerned with modeling norms in finitely repeated games in order to apply the concept to dynamic laboratory environments. Simplifying, there, a norm is a correspondence that prescribes behavior in any information set of the game. If an action is chosen that is not prescribed by the norm, there is a deviation from the norm, cf. Lopez-Perez (2008), Sontuoso (2013).

in the specific setup, e.g. altruism, social image concerns, inequity or guilt aversion, are modeled separately as another additive term in the utility function, e.g. z, with a function f with adequate mathematical properties, and usually with their own weighing parameter, e.g. ϕ. The standard utility function informing empirical research on social norms thus takes a linearly, additive form

$$u(x, \bar{x}, \cdots) = V(\pi(x, \cdots)) \pm \gamma g(x - \bar{x}) \pm \phi f(z(x, \cdots), \tag{138}$$

where dots $(\cdots)$ indicate that additional modeling choices are possible, e.g. the action of the other player(s) may play a role. The social norm is exogenous and cannot be influenced by the agent, which serves the purpose of most empirical research which analyzes behavior at a specific point in time where norm change is unlikely.[105] This standard model serves empirical researchers well for anchoring their ideas.[106] Despite this theoretical simplicity, empirical and theoretical work have fruitfully fertilized each other in the past and will continue to do so. For this reason, it is advantageous to look at current problems in empirical research to detect possible inspiration for future theoretical work, which is done in the next subsection.

V.4.2 Fertilization possibilities

In this subsection we describe four aspects that are currently debated in empirical research on social norms, three of which might be interesting for future theoretical work. It is not possible, nor our goal, here to provide a full overview over the empirical research on norms[107] and all possible theoretical links. We will highlight some ideas that have repercussions with the theoretical literature and may lead to the mutual fertilization of theoretical and empirical research.

An important issue in the empirical literature on social norms in the last decade has been the elicitation of norms.[108] Researchers find they need to elaborate first whether

[105] Investigating norm change is not very common. The time frames are usually too short to observe change. As evolutionary models have shown, norms can be relatively stable for long periods of times, but change might happen rather quickly (in evolutionary terms). Observing and exploiting a serendipitous external event that changes norms or long term norm change favors field studies compared to laboratory experiments in this regard. See e.g. Hallsworth et al. (2017) for a field experiment with reminder letters to pay taxes alluding to others' behavior, i.e. a descriptive norm, or Besley et al. (2019), who exploit a change in the UK tax code that lead to higher levels of tax evasion for their analysis of tax compliance norms.

[106] For additional examples of empirical research employing a variant of this modeling choice cf. e.g. Conlin et al. (2003), Alpizar et al. (2008), Andreoni and Bernheim (2009), Kimbrough and Vostroknutov (2016), Gächter et al. (2017), Danilov et al. (2018), and Bicchieri et al. (2021).

[107] For a first introduction to the vast literature see e.g. Fehr and Schurtenberger (2018) or Vostroknutov (2020).

[108] Important references for elicitation mechanisms are Burks and Krupka (2012), Krupka and Weber (2013), and Kimbrough and Vostroknutow (2016, 2018).

the social norms they are interested in are understood by the subjects and applicable to the situation of the experiment. There is now often concern for a norm elicitation task before the actual experiment. These elicitation tasks are often simple games or questionnaires. The gathered data, e.g. on norm sensitivity, is then used to control for how the test subjects of different types interact among themselves or with each other. Modeling norm elicitation is not the task of theoretical contributions in economics in general. Rather, this is a research question for the theory of empirical social analysis and the development of adequate research tools. Theoretical research could still contribute to this eminent topic in empirical contributions by focusing on modeling context sensitivity, the next important theme discussed in the empirical literature.

For experimental researchers it is important to know, how the context affects the applicability of norms. For example, why are worse proposals accepted in a ultimatum game setting, if the possible share has been earned by the proposer instead of it being a windfall?[109] Questions of context sensitivity are related to a host of other topics. For example, Gerber and Rogers (2009) as well as Farrow et al. (2020), highlight that framing an issue in a positive or a negative light influences the effectiveness of a social norm intervention. Related relevant issues of embedded sociality are, furthermore, historical perspectives (Hoff et al. 2011) and the focus on the right norm at the right time, so that people can act accordingly (Cialdini et al. 1990).[110] The question how these aspects influence the relationship between norms and influenced behavior is important for understanding social norms. This is where modern salience theory can connect to research on social norms, for example by highlighting reference points for behavior or introducing different probability weights on different norms.[111] Additionally, modeling players as only boundedly rational or naive, could lead to them not knowing of a salient norm or its exact prescription. This could be modeled to explain why there is sometimes insecurity about the applicability of a norm.

The next point concerns the possibly constitutive role of external punishment for social norms. The experimental evidence on altruistic punishment is large.[112] However, some theoretical models have highlighted that social norms could be maintained without external punishment, mainly relying on internal forms of punishment instead. The empirical question becomes, whether external punishment (or the threat thereof)

[109] On the existence of this ”entitlement effect”, cf. Demiral and Mollerstrom (2018).

[110] For additional empirical work on context sensitivity cf. Schultz et al. (2007), Helliwell et al. (2014), Roos et al. (2015), Whitson et al. (2015), Gneezy et al. (2016), Gächter et al. (2017), and Boonmanunt et al. (2020). For norms and focus cf. Nolan et al. (2008), Krupka and Weber (2009), and Reuben and Riedl (2013); For salience of norms cf. Alpizar et al. (2008), Falk et al. (2021), Offiaeli and Yaman (2021). For framing effects cf. Wenzel (2005), Cialdini et al. (2006), Goldstein et al. (2008), Allcott (2011), Yeomans and Herberich (2014), and Hallsworth et al. (2017).

[111] Introduced by Bordalo et al. (2013). For the latest overview, cf. Bordalo et al. (2022).

[112] Cf. Fehr and Gächter (2000a, 2000b, 2002), Fehr et al. (2002), Fehr and Fischbacher (2004), Fisman and Miguel (2007), Lopez-Perez (2008), Balafoutas and Nikiforakis (2012), Carpenter and Matthews (2012), Xiao (2013), Balafoutas et al. (2016), Fehr and Williams (2018), and Dimant and Gesche (2021).

is always needed to maintain norms. This is a challenge for empirical research, given that external punishment is easy to observe, whereas the other factors that contribute to norm stability are difficult to identify from an empirical perspective. Complicating the matter further is the fact that these other factors are not necessarily independent of external punishment as an important stabilizer of social norms. The other factors can thus be accompanied by external punishment, which would require disentangling strategies to measure the influence of these hard to observe causes for behavior. Balafoutas et al. (2016) hint at the possibility that altruistic punishment might not be enough to sustain social norms in case of severe transgressions. They interpret this as a possible reason for the development of formal institutions. Empirical evidence in field studies (Balafoutas and Nikiforakis, 2012) further reveals that the actual number of people willing to punish is relatively small (and smaller than laboratory results would suggest). There are several reasons for this, but a pertinent one is the fear of counter punishment. This renders punishment less forceful in maintaining social norms. A model where transgression is only followed by a punishment with a small probability could be a first step. This could also be modeled as an experiment. Additionally, models where punishment can result in counter punishment and adverse conflict with small probability could enrich the research landscape.

As a last avenue for future research we highlight the still not fully understood interaction of several norms (Görges and Nosenzo, 2020).[113] Different social norms and different types of social norms can depend on each other to develop their agency (e.g. Goldstein et al., 2008). Depending on the norm definition, this also includes moral norms or personal norms that may interfere with or support social norms. There are relatively few theoretical models on the interaction of several social norms (cf. subsection V.3.4.). In sociology, network theory has highlighted the interconnectedness of social components supporting social norms. Network theory in economics[114] is mainly concerned with the developments of groups and the value of the network linking its members together. In connection with sociological contributions modeling norms as a network or as components inside the current economic network theory could inspire a new, fruitful, economic perspective on social norms.

V.5 Critique and conclusion

We close this paper by answering the questions posed in the introduction and state a formal critique of the theoretical economic research on social norms over the last forty years.

[113] For additional empirical literature highlighting this issue cf. Ichino and Maggi (2000), Dana et al. (2007), Schultz et al. (2007), Thøgersen (2008), Kube and Traxler (2011), Helliwell et al. (2014), Raihani and McAuliffe (2014), Schram and Charness 2015, and Charles et al. (2018).

[114] For an introduction cf. Dutta and Jackson (2003).

Social norms have been introduced into economic analysis to account for observed human behavior inconsistent with homo oeconomicus. Behavioral economics are not in conflict with neoclassical economics, but rather constitute a necessary development in economic research. All sciences have to carefully modify theoretical models if they do not fit empirical data (Dhami, 2016). The models reviewed in this paper have concerned themselves with different topics in relation to norms. However, some modeling themes were prevalent, namely reputation concerns, norms as thresholds, and additive utility functions.

One central building block since Akerlof (1980) and Bernheim (1994) has been the assumption that people care about their reputation and the impression others have of them. This is the main content of norms referred to in the literature. The strength of the norm was equalized to the number of people adhering to the norm, which made the severity of punishment for transgression a function of the share of norm followers. As a second building block, norms have often been considered as a threshold for actions. People had to perform at least at some level to avoid social punishment. Taken together, this constitutes a relatively simple model of social norms. More sophisticated papers usually combine multiple factors, some of them modeled according to above blueprint for a social norm in economics. The main modeling choice to introduce social norms into economic models has been to extend the utility function to additively accommodate the conformity effect of the norm. This is potentially problematic. According to Kliemt (2020) economists are reluctant to give up their outcome dependent maximization models when analyzing social norms, despite the fact that social norms are strategy dependent. It might not be adequate to model social norms this way, since like this they remain anchored in outcome focused models, while aspects exactly outside the outcome space might be relevant to explain human behavior modified by social norms. Economists have taken one step away from neoclassical modeling by introducing norms as dependent on strategy, only to reintroduce the outcome as the decisive parameter for decision making by plugging social norms back into a utility function. The consequences of trying to model social norms have not been taken fully into account yet (Kliemt, 2020).

With respect to the general understanding of the theoretical concept social norms, economists also identified the equilibria of their games with social norms. They contribute to our understanding of social norms by showing that many different equilibria and therefore different social norms can be reached. Oftentimes, the analysis boils down to a 2 × 2 coordination game, where the two pure strategy Nash equilibria are candidates for stable social norms, while the mixed strategy Nash equilibrium serves as a tipping point. Social norms have long been considered as a coordination device to reduce coordination costs in economics (Ullmann-Margalit, [1977] (2015)). This reduction to norms as coordination device risks becoming a form of implicit functionalism. Norms in this perspective are wont to be seen to exist only due to the positive function they perform in society. This ignores social norms that make everybody worse off (Elster, 1989; Festré, 2010). It is not efficiency concerns why a norm of tipping exists (Azar, 2005) and it is not efficiency concerns that drive Indian tailors to refrain from mutually beneficial contract renegotiations (Iyer and Schoar, 2015).

Furthermore, the reviewed social norm models, outside of evolutionary models and norm entrepreneurship, usually postulate a norm to exist and people care (to some extent) about complying with the norm prescribed behavior, which might be in conflict with what they would do if there was no norm or they did not care about the norm. If used naively this theoretical approach could be used to explain any sort of behavior by postulating a correspondent norm (Basu, 2001). This would leave social norm analysis arbitrary and useless as a research tool. In order to avoid this fallacy, social norm analysis has to always be closely linked to social norm analysis in other social sciences. Only by making sure that a social norm is to some extent relevant in a specific situation can economic social norm analysis contribute to improve understanding of human behavior (Postlewaite, 1998, 2011; Rege, 2004). Grounding economic theory by employing insights from other social sciences can be an advantage. However, it also risks importing some of the problems of these other branches of the social sciences, for example, a lack of a uniform definition of social norms in sociology.

Additionally, social norm analysis is prone to focus on a single identified norm and its impact on behavior, usually in a partial equilibrium analysis. This approach might be insufficient in some cases. The call (Schultz et al., 2007) for models that are capable to illustrate the interaction between several social norms or between social norms and other social institutions has so far only been answered insufficiently. Exceptions and attempts to remedy this shortcoming are the signaling models of Bénabou and Tirole (2006, 2011) and Ali and Bénabou (2020), the model by Huck et al. (2012), and the models in subsection V.3.4, where also conflicts and struggles over correct interpretations of the contents of social norms and their applicability were considered. It is especially important to know how social norms support each other in order to be able to understand and potentially change them. A social norms is no abstract entity that just exists. They are always related to the society they belong to and other social factors relevant in a specific society. This interdependence of social norms has long been an important aspect of research in sociology, for example in the works of Foucault (1977) and Bourdieu (1979). These classical sociological contributions refer to the role played by social power as an important factor in the process that creates, stabilizes, and changes the interrelations of social norms. The two economic approaches reviewed here that consider the inception of norms either propose no social power (evolutionary game theory) or almost absolute social power (norm entrepreneurship). Both are lacking in modeling interactions of social norms. The models that attempted to model interactions, where not concerned with power. Admittedly, modeling power is difficult in economic models. The interactions and reciprocal stabilization of social norms remains an open space for future economic research. Lack of concepts for how social norms emerge and develop is a shortcoming of the empirical literature as well (Gneezy et al. 2016).

Nevertheless, economic research has opened itself up to introduce social norms in their theoretical models. This research agenda has improved economic knowledge of social norms and their effects in society. For example, economic social norm analysis has shown that norms can be differently strong or that different groups may have different social norms. Furthermore, it has been described that social norms are not always clear in their prescriptions and that people have some room for maneuver to pretend to be something they are not or to feign ignorance of the norm. Norms do function as coordination devices and can increase welfare in some circumstances. Harmful norms are less often researched but their continued existence is compatible with economic analysis. Economists have extensively shown that cooperative behavior can be a stable social outcome. Evolutionary game theory has highlighted that social norms are most likely the outcome of many small social interactions in a group. Other empirical facts of social norms have also been described with the help of evolutionary game theory (cf. subsection V.3.7). Similarly, the role of historical accidents, be it in form of shocks or norm entrepreneurs, has been highlighted. This constitutes an already long but still not exhaustive list of results of the above contributions. However, since economists have only comparatively recently become interested in social norms, other behavioral social scientists are mostly familiar with the above list of characteristics of social norms. The contribution so far has mainly consisted in showing that economic models can be designed to reproduce known empirical facts on social norms. From an economic perspective, social norm analysis has contributed much to our understanding of social norms. From a general social science perspective, the contribution has been moderate.

Finally, the advances in the modeling of social norms have been modest to promising. Most contributions rely on a modified version of Akerlof (1980) or posit a simple threshold model. However, researchers' main goal in these contributions is seldom the advancement of the modeling of social norms. They are interested in the effects of a specific norm under consideration. While, simplicity is not a bad modeling choice, in case of social norms it might be inadequate to capture the finer nuances that influence human behavior. Sophistication in this area of research does not necessarily require elaborate, complex models but sensitivity for the difficulty and an appreciation of the contributions of other social sciences. Such interesting, promising, sophisticated models do indeed exist, namely signaling models inspired by Bénabou and Tirole (2006), the model by Michaeli and Spiro (2015) from subsection V.3.5, the evolutionary models that employ advanced learning mechanisms, or the model by Acemoglu and Jackson (2015) on norm entrepreneurs. These models promise to carry economic social norm research beyond replicating the results of other social sciences.

VI Conclusion

This dissertation focused on motivational factors for human behavior. Incongruences between theoretically forecast behavior and actually observed behavior have led to the development of behavioral economics. Behavioral economics have suggested that, in addition to self-interest, psychological and social motivators influence human agency. This thesis has focused on two social motivators. The first of these was social preferences.

In chapters II through IV, a simple agency model with social preferences, specifically, socially attentive preferences, was presented and analyzed. It delivered empirically relevant predictions on wage posting, inefficient surveillance by employers, and non-monetary incentives.

The second social motivator considered in this thesis was the concept of social norms. Chapter V provided a review of the theoretical economic approaches with regard to social norms. The results of the last forty years of research are mixed. While the economic literature has been successful in replicating the findings of other social sciences with regard to norms and generally found ways of modeling that complied with the respective research interest, the overall sophistication of the approach, with some noteworthy exceptions, remains unsatisfactory. The models are hardly capable to reflect or analyze the complexities of social interactions. Here, identity economics could form a way for future research. Up until now, however, the approach has not proved its superiority for modeling complex social systems.

While psychological motivators have found their way into economic policy consulting (e.g. the Behavioral Insights Team, a.k.a the ”Nudge Unit” (Quinn, 2018), in the UK), the research on social motivators has thus far not had the same real world impact. It remains largely confined to the laboratory, field studies, and theoretical research. This may well be because these social concepts are difficult to nail down and to determine once an for all. They are fluid concepts that change in the same rhythm as the societal context around them. This is a challenge and should invite behavioral economists to engage with other behavioral social sciences that find it easier to model these flowing concepts, for example, system theory in sociology (Luhmann, 2018).

Appendix A: Pooling contract: Proof of Proposition 4

Suppose, contrary to our claim, that the principal optimally sets a menu of contracts $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$, where $m \geq 2$. We first examine the case where m is finite and later describe how the proof has to be adjusted in case m is infinite. It is without loss of generality to suppose that every contract in this menu is chosen by at least one type, since every contract that is not chosen by any type is redundant. We denote the payment spread of contract $\left(t_1^j, t_2^j\right)$ by Δt^j, i.e., $\Delta t^j := t_2^j - t_1^j$. Observe that all contracts in the menu have to provide different payment spreads. This holds true, because all types of agents, given the choice between contracts with identical payment spreads, would choose the contract that offered the highest payments. We denote the set of types that select the contract $\left(t_1^j, t_2^j\right)$ from the menu by $\boldsymbol{\beta^j}$. We have to distinguish between two cases.

Case I: the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$ contains a contract $\left(t_1^j, t_2^j\right)$ with $0 < \Delta t^j < \Delta R$. Denote the contract that provides the maximal payment spread among all contracts that satisfy $0 < \Delta t^j < \Delta R$ by $\left(t_1^k, t_2^k\right)$. If $t_1^k > \underline{t}$, i.e., if the payment in case of failure is larger than the minimal possible payment, we lower both payments t_1^k and t_2^k by $t_1^k - \underline{t}$. Denominate this contract by $\left(t_1^l, t_2^l\right)$. We next show that the principal is better off with the single contract $\left(t_1^l, t_2^l\right)$ than with the menu of contracts $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$. Suppose the principal offers the menu of contracts.

First, consider all contracts for which $\Delta t^j < \Delta R$ and $j \neq k$. Since an agent of type $\beta \in \boldsymbol{\beta^j}$ chooses the contract that maximizes his expected utility, the following incentive-selection constraint must be satisfied:

$$E\left[u_A|\beta, (t_1^j, t_2^j), e(\Delta t^j, \beta)\right] \geq E\left[u_A|\beta, (t_1^k, t_2^k), e\right] \tag{139}$$

for all efforts e and all types $\beta \in \boldsymbol{\beta^j}$. This, in particular, implies that

$$E\left[t|(t_1^j, t_2^j), e(\Delta t^j, \beta)\right] \geq E\left[t|(t_1^k, t_2^k), e(\Delta t^j, \beta)\right]. \tag{140}$$

That is, if the effort level is kept constant, the expected payment to the agent of type $\beta \in \boldsymbol{\beta^j}$ is weakly lower under the contract $\left(t_1^k, t_2^k\right)$ than under the contract $\left(t_1^j, t_2^j\right)$. Furthermore, since the contract $\left(t_1^l, t_2^l\right)$ provides weakly lower payments than the contract $\left(t_1^k, t_2^k\right)$, it holds that

$$E\left[t|(t_1^j, t_2^j), e(\Delta t^j, \beta)\right] \geq E\left[t|(t_1^k, t_2^k), e(\Delta t^j, \beta)\right] \geq E\left[t|(t_1^l, t_2^l), e(\Delta t^j, \beta)\right]. \tag{141}$$

Thus, if we artificially keep the effort level constant, the principal's expected payment to all types of agents who select contracts with $\Delta t^j < \Delta R$ and $j \neq k$ is at least weakly lower if the principal replaces the menu of contracts by the single contract $\left(t_1^l, t_2^l\right)$.

We have to examine next how the agents adjust their effort levels. By definition of the contract $\left(t_1^k, t_2^k\right)$, for all contracts with $\Delta t^j < \Delta R$ and $j \neq k$ it holds that $\Delta t^j < \Delta t^k = \Delta t^l$. Consequently, all types of agents who would select contracts with $\Delta t^j < \Delta R$ and $j \neq k$, if the menu is offered, increase their efforts if the principal replaces the menu of contracts by the single contract $\left(t_1^l, t_2^l\right)$. Since $\Delta t^l < \Delta R$, the principal benefits from the agents' higher efforts, $\frac{dE[u_P]}{de} > 0$. Together with (141), we thus have that

$$E\left[u_P|(t_1^l, t_2^l), e(\Delta t^l, \beta)\right] > E\left[u_P|(t_1^j, t_2^j), e(\Delta t^j, \beta)\right] \tag{142}$$

for all contracts with $\Delta t^j < \Delta R$ and $j \neq k$ and all types $\beta \in \boldsymbol{\beta^j}$.

Second, consider the contracts for which $\Delta t^j \geq \Delta R$. The principal's expected utility for contracts with $\Delta t^j \geq \Delta R$ is $E\left[u_P|(t_1^j, t_2^j), e(\Delta t^j, \beta)\right] \leq R_1 - \underline{t}$. Instead, with the contract $\left(t_1^l, t_2^l\right)$ her expected payoff is $E\left[u_P|(t_1^l, t_2^l), e(\Delta t^l, \beta)\right] > R_1 - \underline{t}$, since $e(\Delta t^l, \beta) > 0$. Hence, the principal's expected utility is higher with the contract $\left(t_1^l, t_2^l\right)$ than with the menu of contracts for all types who would select contracts with $\Delta t^j \geq \Delta R$.

Third, consider the contract $\left(t_1^k, t_2^k\right)$. When the principal replaces the menu of contracts with the single contract $\left(t_1^l, t_2^l\right)$, the agents who would select $\left(t_1^k, t_2^k\right)$ continue to provide the same effort since $\Delta t^l = \Delta t^k$. As the payments are weakly lower with the contract $\left(t_1^l, t_2^l\right)$ compared to the contract $\left(t_1^k, t_2^k\right)$, the principal's expected utility is weakly higher with the single contract $\left(t_1^l, t_2^l\right)$ than with the menu for all types who would select the contract $\left(t_1^k, t_2^k\right)$.

Summarizing, if the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$ contains a contract $\left(t_1^j, t_2^j\right)$ with $0 < \Delta t^j < \Delta R$, the principal's expected utility if she offers the single contract $\left(t_1^l, t_2^l\right)$ instead of the menu, is weakly higher for the types of agents who would have chosen contract $\left(t_1^k, t_2^k\right)$ from the menu and strictly higher for all other types who would have chosen any other contract from the menu. Note that there exists at least one type for whom the principal's expected utility is strictly higher.

Case II: the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$ does not contain a contract $\left(t_1^j, t_2^j\right)$ with $0 < \Delta t^j < \Delta R$. This implies that, for all contracts in the menu, either $\Delta t^j \leq 0$ or $\Delta t^j \geq \Delta R$. We next show that the principal's utility improves if she replaces the menu with the single contract $\left(t_1^l = \underline{t}, t_2^l = \underline{t}\right)$.

First, consider contracts $\left(t_1^j, t_2^j\right)$ with $\Delta t^j \leq 0$. Then $\Delta t^j \leq \Delta t^l$ and $t_i^j \geq t_i^l = \underline{t}$ for $i \in \{1,2\}$. Thus, the principal implements a weakly higher effort for a weakly lower expected payment. Accordingly, the principal's expected utility is weakly higher with the single contract $\left(t_1^l, t_2^l\right)$ than with any contract $\left(t_1^j, t_2^j\right)$ with $\Delta t^j \leq 0$ for all types who would have chosen these contracts.

Second, consider contracts $\left(t_1^j, t_2^j\right)$ with $\Delta t^j \geq \Delta R$. Then,

$$E[u_P|(t_1^j, t_2^j), e(\Delta t^j, \beta)] = R_1 - t_1^j + p(\cdot)(\Delta R - \Delta t^j) \leq R_1 - \underline{t}, \tag{143}$$

while

$$E[u_P|(t_1^l = \underline{t}, t_2^l = \underline{t}), e(\Delta t^l, \beta)] = R_1 - t_1^l + p(\cdot)(\Delta R - \Delta t^l) = R_1 - \underline{t} + p(\cdot)\Delta R \geq R_1 - \underline{t}, \quad (144)$$

such that the principal's expected profit again improves.

Further observe that, in order for the menu to be incentive-compatible, there must exist at least one contract for which $t_1^j > \underline{t}$, independently of whether $\Delta t^j \leq 0$ or $\Delta t^j \geq \Delta R$. This holds since if all contracts were characterized by $t_1^j = \underline{t}$, all types would choose the contract with the highest t_2^j, such that there would not exist a menu where different contracts were chosen. For all contracts with $t_1^j > \underline{t}$,

$$E\left[u_P|(t_1^j, t_2^j), e(\Delta t^j, \beta)\right] < E\left[u_P|(t_1^l, t_2^l), e(\Delta t^l, \beta)\right], \quad (145)$$

since

$$E[t|(t_1^j, t_2^j), e(\Delta t^j, \beta)] > E[t|(t_1^l = \underline{t}, t_2^l = \underline{t}), e(\Delta t^l, \beta)] = \underline{t} \quad (146)$$

and $e(\Delta t^j, \beta) \leq e(\Delta t^l, \beta)$ in case $\Delta t^j \leq 0$, and

$$E[u_P|(t_1^j, t_2^j), e(\Delta t^j, \beta)] = R_1 - t_1^j + p(\cdot)(\Delta R - \Delta t^j) < R_1 - \underline{t} \leq E[u_P|(t_1^l, t_2^l), e(\Delta t^l, \beta)] \quad (147)$$

in case $\Delta t^j \geq \Delta R$. Therefore, if the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$ does not contain a contract $\left(t_1^j, t_2^j\right)$ with $0 < \Delta t^j < \Delta R$, the principal's expected utility increases weakly for all types and strictly for some types when the principal replaces the menu of contracts by the single contract (t_1^l, t_2^l). At least one type of agent for which the principal's expected utility is strictly higher always exists.

In case m is infinite, the previous proof applies except that in Case I – i.e., when the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$ contains a contract $\left(t_1^j, t_2^j\right)$ with $0 < \Delta t^j < \Delta R$ – one has to use the supremum instead of the maximum operator. Formally, the contract $t_1^l = \underline{t}$, $t_2^l = \underline{t} + \sup\left\{\Delta t^j : 0 < \Delta t^j < \Delta R\right\}$ is used to replace the menu $\left\{\left(t_1^j, t_2^j\right)\right\}_{j=1}^m$. If $t_2^l = \underline{t} + \Delta R$, a final step has to be added to guarantee that the principal's expected utility strictly increases: the contract $\left(t_1^l = \underline{t}, t_2^l = \underline{t} + \Delta R\right)$ is further replaced by the contract $\left(t_1^{\tilde{l}} = \underline{t}, t_2^{\tilde{l}} = \underline{t} + \Delta R/2\right)$.
□

Appendix B: Complex screening contracts

Consider the model with moral hazard and adverse selection in chapter II. The agent's type β is drawn from the c.d.f. F with support $[0,\bar{\beta}]$, and corresponding probability density function f. We next argue that the construction of menus of complex contracts, which may be non binary or not reward outcome n, is not optimal. Suppose the principal designs a menu of contracts, where $\left(t_{1,\beta},\ldots,t_{n,\beta}\right)$ denotes the contract designated to type β. The principal maximize her expected utility subject to

(i) the incentive constraints that the effort level implemented for type β is a solution of agent β's maximization problem, i.e., that $\hat{e}_\beta \in \operatorname{argmax} E[u_A|\beta]$ for all types β,

(ii) the global selection constraints, according to which no type of agent can improve his expected utility by choosing a contract not designated for him; formally, it has to hold for all types β and $\hat{\beta}$ and all efforts e that

$$\begin{aligned}(1-\beta)\sum_{i=1}^{n} p_i(e_\beta)\,t_{i,\beta} + \beta\sum_{i=1}^{n} p_i(e_\beta)(R_i+V_i) - c(e_\beta)&\\ \geq (1-\beta)\sum_{i=1}^{n} p_i(e)\,t_{i,\hat{\beta}} + \beta\sum_{i=1}^{n} p_i(e)(R_i+V_i) - c(e),&\end{aligned}$$

(iii) the limited liability constraints $t_{i,\beta} \geq \underline{t}$ for all outcomes $i \in \{1,\ldots,n\}$ and all β.

We replace the global incentive constraint by the local incentive constraint and the global selection constraint by the relaxed selection constraint that

$$\sum_{i=1}^{n} p_i(e_\beta)\,t_{i,\beta} \geq \sum_{i=1}^{n} p_i(e_\beta)\,t_{i,\tilde{\beta}}$$

for all close types $\tilde{\beta} \in \tilde{\boldsymbol{\beta}}$, where closeness is defined with respect to efforts, i.e., $\tilde{\boldsymbol{\beta}}$ denotes the set of types for which $e_\beta - e_{\tilde{\beta}} \in [-\varepsilon,\varepsilon]$ holds, where ε is small and positive. The relaxed selection constraint is necessary for the global selection constraint and expresses that an agent of type β must receive an expected payment with the contract designed for his type that does not fall short of the expected payments associated with the contracts designed for types that are close, when keeping the effort level constant.

The Lagrangian for this problem writes as

$$\begin{aligned}\mathscr{L} = &\int_{\beta=0}^{1} f(\beta)\left(\sum_{i=1}^{n} p_i(e_\beta)R_i - \sum_{i=1}^{n} p_i(e_\beta)t_{i,\beta}\right)d\beta \\ &+\int_{\beta=0}^{1} \mu_\beta\left((1-\beta)\sum_{i=1}^{n} p_i'(e_\beta)t_{i,\beta} + \beta\sum_{i=1}^{n} p_i'(e_\beta)(R_i+V_i) - c'(e_\beta)\right)d\beta \\ &+\int_{\beta=0}^{1}\int_{\tilde{\boldsymbol{\beta}}} \lambda_{\beta,\tilde\beta}\left(\sum_{i=1}^{n} p_i(e_\beta)t_{i,\beta} - \sum_{i=1}^{n} p_i(e_\beta)t_{i,\tilde\beta}\right)d\tilde\beta d\beta \\ &+\int_{\beta=0}^{1}\sum_{i=1}^{n}\rho_{i,\beta}(t_{i,\beta}-\underline{t})d\beta.\end{aligned} \tag{148}$$

Hence, for (almost) all types β and outcomes i it has to hold that

$$\frac{\partial\mathscr{L}}{\partial t_{i,\beta}} = -f(\beta)p_i(e_\beta) + \mu_\beta(1-\beta)p_i'(e_\beta) + \int_{\tilde{\boldsymbol{\beta}}}\lambda_{\beta,\tilde\beta}p_i(e_\beta)d\tilde\beta - \int_{\tilde{\boldsymbol{\beta}}}\lambda_{\tilde\beta,\beta}p_i(e_{\tilde\beta})d\tilde\beta + \rho_{i,\beta} = 0,$$

where we used that type β is close to types $\tilde\beta$ and vice versa. Dividing by $p_i(e_\beta)$ yields

$$\frac{\partial\mathscr{L}}{\partial t_{i,\beta}} = -f(\beta) + \mu_\beta(1-\beta)\frac{p_i'(e_\beta)}{p_i(e_\beta)} + \int_{\tilde{\boldsymbol{\beta}}}\lambda_{\beta,\tilde\beta}d\tilde\beta - \int_{\tilde{\boldsymbol{\beta}}}\lambda_{\tilde\beta,\beta}\frac{p_i(e_{\tilde\beta})}{p_i(e_\beta)}d\tilde\beta + \frac{\rho_{i,\beta}}{p_i(e_\beta)} = 0. \tag{149}$$

From (149) we can infer that only outcome n – i.e., the outcome with the highest likelihood ratio – could be rewarded with a payment above the limited liability threshold $\underline{t}$. The proof is by contradiction. If, contrary to our claim, $t_{j,\beta} > \underline{t}$ for some $j \neq n$, then $\frac{\partial\mathscr{L}}{\partial t_{j,\beta}} = 0$ has to hold. But then $\frac{\partial\mathscr{L}}{\partial t_{n,\beta}} = 0$ cannot be satisfied, since (i) the Lagrange parameter of the incentive constraint μ_β is positive by standard arguments,[115] (ii) the likelihood ratio is smaller for outcome j than for outcome n, $\frac{p_j'(e_\beta)}{p_j(e_\beta)} < \frac{p_n'(e_\beta)}{p_n(e_\beta)}$, by definition, (iii) $\frac{p_i(e_{\tilde\beta})}{p_i(e_\beta)} \approx 1$ for $i = j, n$ because β and $\tilde\beta$ are close by definition, and (iv) the limited liability constraints are inequity constraints such that $\rho_{n,\beta} \geq 0$, while by complementary slackness $\rho_{j,\beta} = 0$. It must hence hold that $t_{i,\beta} = \underline{t}$ for all outcomes $i \neq n$ and $t_{n,\beta} \geq \underline{t}$.

Given this structure of contracts, the relaxed selection constraints require that $t_{n,\beta} = t_{n,\beta'}$ for all β and β' that are close and thus by induction also $t_{n,\beta} = t_{n,\beta'}$ for all β and β'. Therefore, in the optimum, each type of agent must receive the same payment in case outcome n realizes. Observe that this solution does not only satisfy the relaxed, local constraints, but also the global constraints such that the solution of the relaxed problem is also the solution of the full problem. We can thus conclude that the principal optimally offers the same binary contract $t_{i,\beta} = \underline{t}$ and $t_{n,\beta} = t_n \geq \underline{t}$ to all types.

[115] If the Lagrange multiplier where negative, the solution would specify that only outcome 1 – i.e., the outcome with the lowest and negative likelihood ratio – could be rewarded. But such contracts implement a weakly lower effort to strictly higher costs than the contract $t_{i,\beta} = \underline{t}$ and are thus not optimal.

The result is intuitive: We know from the main part of the paper that providing a menu of different binary contracts that reward outcome n is not optimal, since the principal can always improve by replacing the menu with a single contract. Menus of contracts that are more complex are not optimal either, since they additionally reward outcomes with relatively low likelihood ratios. Such contracts are not cost effective, because the incentive effect of rewarding outcomes with low likelihood ratios is small relative to the costs.

Appendix C: Contractible Effort: Proof of Proposition 13

Given the contract (w, e), the principal's expected utility writes

$$E[u_P] = p(e)R - w \tag{150}$$

and the agent's expected utility is

$$E[u_A] = w - c(e) + \beta\big(p(e)(R+V) - w\big). \tag{151}$$

The principal's problem is to maximize her expected utility subject to (i) the limited liability constraint $w \geq 0$ and (ii) the agent's participation constraint $E[u_A] \geq 0$.

Because, all else equal, the principal's expected utility is increasing in effort, it could never be optimal for the principal to require an effort level such that the agent's expected utility is positive. Accordingly, we can restrict our attention to the case where $E[u_A] = 0$. Solving $E[u_A] = 0$ for the wage payment w yields

$$w = \frac{c(e) - \beta p(e)(R+V)}{1-\beta}. \tag{152}$$

We are hence able to rewrite the constraint $w \geq 0$ as

$$c(e) - \beta p(e)(R+V) \geq 0. \tag{153}$$

The principal's problem can thus be reformulated. She maximizes her expected utility subject to (152) and (153).

We now take a closer look at constraint (153). Consider first the case where $\beta > 0$. Then (153) holds with equality for $e = 0$ as well as for exactly one positive effort level, namely

$$\underline{e} := \max\{e | c(e) - \beta p(e)(R+V) = 0\}. \tag{154}$$

See Figure 12.

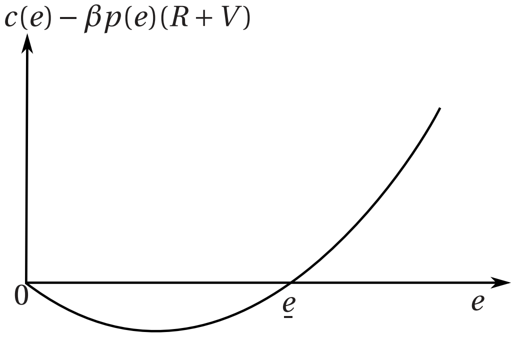

Figure 12: The constraint (153) and the construction of $\underline{e}$.

Because $c(e) - \beta p(e)(R+V)$ is convex in e, constraint (153) holds if either $e = 0$ or

$e \geq \underline{e}$. Since implementing $e = 0$ is never optimal for the principal,[116] we can replace constraint (153) by the simpler constraint $e \geq \underline{e}$.

Consider next the remaining case $\beta = 0$. Then (153) holds if and only if $e \geq \underline{e} = 0$, We can thus again replace the constraint (153) by the constraint $e \geq \underline{e}$. This allows us to write the principal's problem as

$$\max_{w,e} \; E[u_P] \;\text{ subject to }\; (152) \text{ and } e \geq \underline{e}. \tag{P3}$$

Plugging (152) into the principal's objective function yields

$$E[u_P] = p(e)R - \frac{c(e) - \beta p(e)(R+V)}{1-\beta}. \tag{155}$$

After multiplying by $(1-\beta)$,[117] we can rewrite the problem (P3) as

$$\max_{e} \; p(e)R - c(e) + \beta p(e)V \;\text{ subject to }\; e \geq \underline{e}. \tag{P4}$$

Because the objective function is concave, the principal optimally either implements the unconstrained maximizer of her objective function

$$\begin{aligned} e^{\text{uncon}} &:= \operatorname*{argmax}_{e} \; p(e)R - c(e) + \beta p(e)V \qquad (156) \\ \Longleftrightarrow e^{\text{uncon}} &\;\text{ solves }\; p'(e)(R+\beta V) - c'(e) = 0, \end{aligned}$$

or, if this is not possible due to $e^{\text{uncon}} < \underline{e}$, the effort level $\underline{e}$. Denoting the effort level the principal optimally implements in case of contractible effort by $e^{\text{contractible}}$, we thus have

$$e^{\text{contractible}} = \max\left\{e^{\text{uncon}}, \underline{e}\right\}. \tag{157}$$

The optimal wage payment provided by the principal is, by (152),

$$w^{\text{contractible}} = \frac{c\left(e^{\text{contractible}}\right) - \beta p\left(e^{\text{contractible}}\right)(R+V)}{1-\beta}. \tag{158}$$

This proves the first part of Proposition 13.

To compare the effort level the principal implements, $e^{\text{contractible}}$, to the efficient effort $e^{\text{efficient}}$, the following properties are useful and straightforward to derive. First, recall that $e^{\text{efficient}}$ solves (39), i.e., the first-order condition $\partial E[s]/\partial e = p'(e)(R+V) - c'(e) = 0$. We directly see that $e^{\text{efficient}}$ is positive and independent of β. Second, for effort $\underline{e}$, the

[116] When implementing effort zero, the principal would yield $E[u_P] \leq 0$, while she could obtain $E[u_P] > 0$ if she implements effort $e = \varepsilon$, where ε is small and positive. Formally, the contract $w = c(\varepsilon)$ and $e = \varepsilon$ guarantees the agent's participation and yields the principal an expected utility $E[u_P] = p(\varepsilon)R - c(\varepsilon)$, which is positive since ε is small.

[117] This multiplication is unproblematic, since it constitutes a positive monotone transformation of the principal's objective function.

following holds:

$$\underline{e}\big|_{\beta=0} = 0, \ \frac{\partial \underline{e}}{\partial \beta} > 0, \ \underline{e}\big|_{\beta=1} > e^{\text{efficient}}. \tag{159}$$

Third, for effort e^{uncon}, the following holds:

$$\text{if } V = 0, \text{ then } \ e^{\text{uncon}} > 0, \ e^{\text{uncon}} = e^{\text{efficient}}, \ \frac{\partial e^{\text{uncon}}}{\partial \beta} = 0; \tag{160}$$

$$\text{if } V > 0, \text{ then } \ e^{\text{uncon}} > 0, \ e^{\text{uncon}} < e^{\text{efficient}}, \ \frac{\partial e^{\text{uncon}}}{\partial \beta} > 0. \tag{161}$$

From the formulas (159)-(161) it directly follows that the effort level the principal optimally implements $e^{\text{contractible}} = \max\left\{e^{\text{uncon}}, \underline{e}\right\}$ is weakly increasing in β if $V = 0$, $\partial e^{\text{contractible}}/\partial\beta \geq 0$, and strictly increasing in β if $V > 0$, $\partial e^{\text{contractible}}/\partial\beta > 0$.

The threshold $\bar{\beta}$ is such that $\underline{e} = e^{\text{efficient}}$. Solving

$$c\left(e^{\text{efficient}}\right) - \bar{\beta} p\left(e^{\text{efficient}}\right)(R+V) = 0 \tag{162}$$

for $\bar{\beta}$ yields

$$\bar{\beta} = \frac{c\left(e^{\text{efficient}}\right)}{p\left(e^{\text{efficient}}\right)(R+V)}. \tag{163}$$

Note that, because at least for the effort level $e^{\text{efficient}}$ the expected surplus is positive, so that $p\left(e^{\text{efficient}}\right)(R+V) - c\left(e^{\text{efficient}}\right) > 0$, we have $\bar{\beta} \in (0,1)$. Figures 13 and 14 illustrate these properties and the second part of Proposition 13 summarizes them.

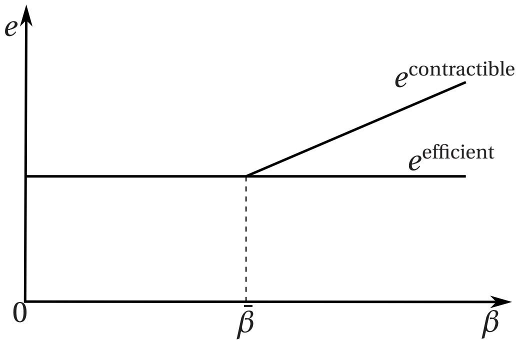

Figure 13: Comparison of $e^{\text{contractible}}$ and $e^{\text{efficient}}$ for the case $V = 0$.

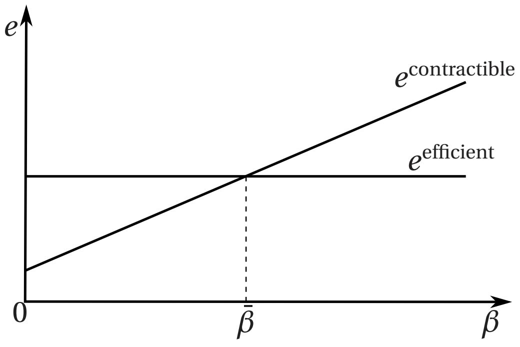

Figure 14: Comparison of $e^{\text{contractible}}$ and $e^{\text{efficient}}$ for the case $V > 0$.

Appendix D: Different weights on utilities

Consider the model with socially attentive agents in chapter II through IV. We assumed that the agent puts the same weight β on the principal's and the third party's utilities. The model is readily generalized to different weights. Denoting the weight on the principal's utility by β_P and the weight on the third party's utility by β_T, we can write the agent's expected utility as

$$E[u_A] = p(e)\,w_R + (1-p(e))\,w_0 - c(e) + \beta_P \big(p(e)(R-w_R) - (1-p(e))\,w_0\big) + \beta_T\, p(e) V. \quad (72')$$

Rescaling the third party's payoff to $\tilde{V} := \frac{\beta_T}{\beta_P} \times V$ and writing β for β_P allows us to reformulate (72') as

$$E[u_A] = p(e)\,w_R + (1-p(e))\,w_0 - c(e) + \beta\big(p(e)(R - w_R + \tilde{V}) - (1-p(e))\,w_0\big). \qquad (72")$$

Observe that – except for having $\tilde{V}$ instead of V, which is qualitatively inconsiderable – equation (72") is identical to the agent's objective function with equal weights, which we already know from the main text, see equation (72). Allowing the agent to put different weights on the principal's and the third party's utilities therefore has the same effect as variations of the third party's payoff V have. Due to this insight, and to keep the notation as compact as possible, we do not allow for different weights in the main text of the paper.

References

Abbink, Klaus, Lata Gangadharan, Toby Handfield, and John Thrasher (2017). Peer Punishment Promotes Enforcement of Bad Social Norms. *Nature Communications* 8, Article 609.

Abbott, Andrew, Shasikanta Nandeibam, and Lucy O'Shea (2013). Recycling: Social norms and warm-glow revisited. *Ecological Economics* 90, pp. 10-18.

Acemoglu, Daron, and Matthew O. Jackson (2015). History, Expectations, and Leadership in the Evolution of Social Norms. *The Review of Economic Studies* 82 (2), pp. 423-456.

Akerlof, George A. (1980). A Theory of Social Custom, of Which Unemployment May be One Consequence. *The Quarterly Journal of Economics* 94 (4), pp. 749-775.

Akerlof, George A. (1997). Social Distance and Social Decisions. *Econometrica* 65 (5), pp. 1005-1027.

Akerlof, George A., and William T. Dickens (1982). The Economic Consequences of Cognitive Dissonance. *The American Economic Review* 72 (3), pp. 307-319.

Akerlof, George A., and Rachel Kranton (2000). Economics and Identity. *The Quarterly Journal of Economics* 115 (3), pp. 715-753.

Akerlof, George A., and Rachel Kranton (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives* 19 (1), pp. 9-32.

Alger, Ingela, and Jörgen W. Weibull (2013). Homo Moralis - Preference Evolution Under Incomplete Information and Assortative Matching. *Econometrica* 81 (6), pp. 2269-2302.

Ali, S. Nageeb, and Roland Bénabou (2020). Image versus Information: Changing Societal Norms and Optimal Privacy. *American Economic Journal: Microeconomics* 12 (3), pp. 116-164.

Allcott, Hunt (2011). Social Norms and Energy Conservation. *Journal of Public Economics* 95 (9-10), pp. 1082-1095.

Aloud, Monira Essa, Sara Al-Rashood, Ina Ganguli, and Basit Zafar (2020). Information and Social Norms: Experimental Evidence on the Labor Market Aspirations of Saudi Women. *NBER Working Paper Series* No. 26693.

Alpizar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenmand (2008). Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a national Park in Costa Rica. *Journal of Public Economics* 92, pp. 1047-1060.

Alpman, Anil (2013). The Relevance of Social Norms for Economic Efficiency: Theory and its Empirical Test. *Centre d'Economie de la Sorbonne Working Paper* No. 2013.38R.

Andreoni, James (1990). Impure Altruism and Donations to Public Goods: A Theory of

Warm-Glow Giving. *The Economic Journal* 100 (401), pp. 464-477.

Andreoni, James, and B. Douglas Bernheim (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica* 77 (5), pp. 1607-1636.

Andreoni, James, William T. Harbaugh, and Lise Vesterlund (2007). Altruism in Experiments. In: Durlauf, Steven N., and Lawrence E. Blume (eds.) *New Palgrave Dictionary of Economics*, 2nd edition. Palgrave Macmillan, London, pp. 265-271.

Andreoni, James, and John Miller (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica* 70 (2), pp. 737-753.

Arce, Daniel G. (2013). Principal's Preferences for Agents with Social Preferences. *Journal of Economic Behavior and Organization* 90, pp. 154-163.

Arrow, Kenneth J. (1972). Some Mathematical Models of Race Discrimination in the Labor Market. In: Pascal, Anthony H. (ed.), *Racial Discrimination in Economic Life.* Lexington Books, Lexington, pp. 187-203.

Arrow, Kenneth J. (1994). Methodological Individualism and Social Knowledge. *The American Economic Review* 84 (2), pp. 1-9.

Ashby, W. Ross [1956] (2015). *An Introduction to Cybernetics.* Martino Publishing, Eastford. Reprint.

Ashraf, Nava, and Oriana Bandiera (2017). Altruistic Capital. *American Economic Review* 107 (5), pp. 70-75.

Azar, Ofer H. (2004). What Sustains Social Norms and how they Evolve? The Case of Tipping. *Journal of Economic Behavior and Organization* 54, pp. 49-64.

Balafoutas, Loukas, and Nikos Nikiforakis (2012). Norm Enforcement in the City: A Natural Field Experiment. *European Economic Review* 56 (8), pp. 1773-1785.

Balafoutas, Loukas, Nikos Nikiforakis, and Bettina Rockenbach (2016). Altruistic Punishment does not Increase with the Severity of Norm Violations in the Field. *Nature Communications* 7, Article No. 13327.

Balasubramanian, Anirudha (2015). On Weighted Utilitarianism and an Application. *Social Choice and Welfare* 44, pp. 745-763.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005). Social Preferences and the Response to Incentives: Evidence From Personnel Data. *Quarterly Journal of Economics* 120, pp. 917-962.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2009). Social Connections and Incentives in the Workplace: Evidence From Personnel Data. *Econometrica* 77 (4), pp. 1047-1094.

Bardsley, Nicholas, and Rupert Sausgruber (2005). Conformity and Reciprocity in Public Good Provision. *Journal of Economic Psychology* 26, pp. 664-681.

Barr, Abigail, Tom Lane, and Daniele Nosenzo (2018). On the Social Inappropriateness of Discrimination. *Journal of Public Economics* 164, pp. 153-164.

Bart, Christopher Kenneth, and Mark C. Baetz (1998). The Relationship Between Mission Statements and Firm Performance: An Exploratory Study. *Journal of Management Studies* 35(6), pp. 823-853.

Bart, Christopher Kenneth, Nick Bontis, and Simon Taggar (2001). A Model of the Impact of Mission Statements on Firm Performance. *Management Decision* 39 (1), pp. 19-35.

Bartkus, Barbara R., and Myron Glassman (2008). Do Firms Practice What They Preach? The Relationship Between Mission Statements and Stakeholder Management. *Journal of Business Ethics* 83, pp. 207-216.

Bartling, Björn (2011). Relative Performance or Team Evaluation? Optimal Contracts for Other-Regarding Agents. *Journal of Economic Behavior and Organization* 79 (3), pp. 183-193.

Bartling, Björn, and Klaus M. Schmidt (2015). Reference Points, Social Norms, and Fairness in Contract Renegotiations. *Journal of the European Economic Association* 13 (1), pp. 98-129.

Basu, Kaushik (2001). The Role of Norms and Law in Economics: An Essay in Political Economy. In: Scott, Joan W., and Debra Keates (eds.), *Schools of Thought: Twenty-Five Years of Interpretive Social Science*. Princeton University Press, Princeton.

Bell, Brian, and John Van Reenen (2013). Bankers and their Bonuses. *Economic Journal* 124, F1-F21.

Bénabou, Roland, and Jean Tirole (2003). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies* 70 (3), pp. 489-520.

Bénabou, Roland, and Jean Tirole (2006). Incentives and Prosocial Behavior. *The American Economic Review* 96 (5), pp. 1652-1678.

Bénabou, Roland, and Jean Tirole (2011). Laws and Norms. *NBER Working Paper* No. 17579.

Bénabou, Roland, Armin Falk, and Jean Tirole (2018). Narratives, Imperatives, and Moral Reasoning. *NBER Working Paper*, No. 24798.

Benjamin, Daniel J., James J. Choi, and A. Joshua Strickland (2010). Social Identity and Preferences. *The American Economic Review* 100 (4), pp. 1913-1928.

Berger, Peter L., and Thomas Luckmann (1966). *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*. Anchor Books, Garden City, New York.

Bernheim, B. Douglas (1994). A Theory of Conformity. *Journal of Political Economy* 102 (5), pp. 841-877.

Besley, Timothy, and Maitreesh Ghatak (2005). Competition and Incentives with Motivated Agents. *American Economic Review* 95, pp. 616-636.

Besley, Timothy, Anders Jensen, and Torsten Person (2019). Norms, Enforcement, and Tax Evasion. *NBER Working Paper Series* No. 25575.

Bicchieri, Christina (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge University Press. Cambridge.

Bicchieri, Christina, and Alex Chavez (2010). Behaving as Expected: Public Information and Fairness Norms. *Journal of Behavioral Decision Making* 23, pp. 161-178.

Bicchieri, Christina, and Alex Chavez (2013). Norm Manipulation, Norm Evasion: Experimental Evidence. *Economics and Philosophy* 29, pp. 175-198.

Bicchieri, Christina, Eugen Dimant, and Erte Xiao (2021). Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment. *CESifo Working Papers* No. 9067.

Bicchieri, Christina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior* 132, pp. 59-72.

Bidner, Chris, and Patrick Francois (2011). Cultivating Trust: Norms, Institutions and the Implications of Scale. *The Economic Journal* 121 (555), pp. 1097-1129.

Biel, Anders, and John Thøgersen (2007). Activation of Social Norms in Social Dilemmas: A Review of the Evidence and Reflections on the Implications for Environmental Behaviour. *Journal of Economic Psychology* 28, pp. 93-112.

Bigoni, Maria, Matteo Ploner, and Thi-Thanh-Tam Vu (2021). The Right Person for the Right Job: Workers' Prosociality as a Screening Device. *IZA Discussion Paper Series* No. 14779.

Binmore, Ken (2006). Why Do People Cooperate? *Politics, Philosophy and Economics* 5 (1), pp. 81-96.

Binmore, Ken, and Larry Samuelson (1994). An Economist's Perspective on the Evolution of Norms. *Journal of Institutional and Theoretical Economics* 150 (1), pp. 45-63.

Binmore, Ken, and Avner Shaked (2010). Experimental Economics: Where Next? *Journal of Economic Behavior and Organization* 73, pp. 87-100.

Bird, Edward J. (1999). Can Welfare Policy Make Use of Social Norms? *Rationality and Society* 11 (3), pp. 343-365.

Bisin, Alberto, and Thierry Verdier (2001). The Economics of Cultural Transmission and the Dynamics of Preferences. *Journal of Economic Theory* 97, pp. 298-319.

Bisin, Alberto, and Thierry Verdier (2008). Cultural Transmission. In: Durlauf, Steven N., Lawrence E. Blume (eds.), *The New Palgrave Dictionary of Economics.* 2nd edition. Palgrave Macmillan, London, pp. 1225-1229.

Bolton, Gary E., and Axel Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and

Competition. *The American Economic Review* 90 (1), pp. 166-193.

Boonmanunt, Suparee, Agne Kajackaite, and Stephan Meier (2020). Does Poverty Negate the Impact of Social Norms on Cheating? *Games and Economic Behavior* 124, pp. 569-578.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2013). Salience and Consumer Choice. *Journal of Political Economy* 121(5), 803-843.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2022). Salience. *Annual Review of Economics* 14, forthcoming.

Booth, Alison L. (1985). The Free Rider Problem and a Social Custom Model of Trade Union Membership. *The Quarterly Journal of Economics* 100 (1), pp. 253-261.

Bourdieu, Pierre (1979). *La Distinction. Critique Sociale du Jugement.* Les Editions de Minuit, Paris.

Bowles, Samuel (1998). Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions. *Journal of Economic Literature* 46, pp. 75-111.

Boyd, Robert, and Peter J. Richerson (2002). Group Beneficial Norms Can Spread Rapidly in a Structured Population. *Journal of theoretical Biology* 215, pp. 287-296.

Brächer, Michael (2021). Ikea Frankreich soll Millionenstrafe zahlen. In: *Spiegel Online*, URL: https://www.spiegel.de/wirtschaft/unternehmen/ikea-staatsanwaltschaft-fordert-millionenstrafe-a-3e411c91-0316-47cb-bc4a-412e872338eb (22.06.2022).

Brenzel, Hanna, Hermann Gartner, and Claus Schnabel (2014). Wage Bargaining or Wage Posting? Evidence from the Employers' Side. *Labour Economics* 29, pp. 41-48.

Briscese, Guglielmo, Nick Feltovich, Robert Slonim (2021). Who Benefits from Corporate Social Responsibility? Reciprocity in the Presence of Social Incentives and Self-Selection. *IZA Discussion Papers* No. 14067.

Bryson, Alex, Richard Freeman, Claudio Lucifora, Michele Pellizzare, and Virgine Perotin (2012). Paying For Performance: Incentive Pay Schemes and Employees' Financial Participation. *CEP Discussion Paper* No. 1112.

Burke, Mary A, and H. Peyton Young (2011). Social Norms. In: Benhabib, Jess, Alberto Bisin, and Matthew O. Jackson (eds.), *Handbook of Social Economics.* Elsevier, Amsterdam, pp. 311-338.

Burks, Stephen V., and Erin L. Krupka (2012). A Multimethod Approach to Identifying Norms and Normative Expectations within a Corporate Hierarchy: Evidence from the Financial Services Industry. *Management Science* 58 (1), pp. 203-217.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin (2020a). From Extreme to Mainstream: The Erosion of Social Norms. *The American Economic Review* 110 (11), pp. 3522-3548.

Bursztyn, Leonardo, Alessandra L. Gonzales, and David Yanagizawa-Drott (2020b). Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *The American Economic Review* 110 (10), pp. 2997-3029.

Bursztyn, Leonardo, and Robert Jensen (2017). Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure. *Annual Review of Economics* 9, pp. 131-153.

Camerer, Colin F., and Ernst Fehr (2004). Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists. In: Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin F. Camerer, Ernst Fehr, Herbert Gintis (eds.), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford University Press, Oxford et al., pp. 55-95.

Carpenter, Jeffrey P., and Peter Hans Matthews (2010). Norm Enforcement: The Role of Third Parties. *Journal of Institutional and Theoretical Economics* 166 (2), pp. 239-258.

Case, Anne C., and Lawrence F. Katz (1991). The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths. *NBER Working Paper* No. 3705.

Cassar, Lea (2016). Optimal Contracting with Endogenous Project Mission. *CESifo Working Paper* No. 6181.

Cassar, Lea (2019). Job Mission as a Substitute for Monetary Incentives: Benefits and Limits. *Management Science* 65 (2), pp. 896-912.

Cassar, Lea, and Jesper Armouti-Hansen (2020). Optimal Contracting with Endogenous Project Mission. *Journal of the European Economic Association* 18 (5), pp. 2647-2676.

Cassar, Lea, and Stephan Meier (2018). Nonmonetary Incentives and the Implications of Work as a Source of Meaning. *Journal of Economic Perspectives* 32(3), pp. 215-238.

Castro, Silvia, and Kristina Czura (2021). Social Norms and Misinformation: Experimental Evidence on Learning about Menstrual Health Management in Rural Bangladesh. *CESifo Working Papers* No. 9081.

Chang, Daphne, Roy Chen, and Erin Krupka (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior* 116, pp. 158-178.

Charles, Kerwin Kofi, Jonathan Guryan, and Jessica Pan (2018). The Effects of Sexism on American Women: The Role of Norms vs. Discrimination. *NBER Working Paper* No. 24904.

Charness, Gary, and Matthew Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117 (3), pp. 817-869.

Che, Yeon-Koo, and Seung-Weon Yoo (2001). Optimal Incentives for Teams. *American Economic Review* 91 (3), pp. 525-541.

Chen, Ying, Navin Kartik, and Joel Sobel (2008). Selecting Cheap-Talk Equilibria. *Econo-*

metrica 76 (1), pp. 117-136.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology* 58 (6), pp. 1015-1026.

Cialdini, Robert B., Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett, Kelton Rhoads, and Patricia L. Winter (2006). Managing Social Norms for Persuasive Impact. *Social Influence* 1 (1), pp. 3-15.

Clark, Andrew E. (2003). Unemployment as a Social Norm: Psychological Evidence from Panel Data. *Journal of Labor Economics* 21 (2), pp. 323-351.

Cole, Harold L., George J. Mailath, and Andrew Postlewaite (1992). Social Norms, Savings Behavior, and Growth. *The Journal of Political Economy* 100 (6), pp. 1092-1125.

Corneo, Giacomo (1995). Social Custom, Management Opposition, and Trade Union Membership. *European Economic Review* 39, pp. 275-292.

Corneo, Giacomo, and Oliver Jeanne (1997). Snobs, Bandwagons, and the Origin of Social Customs in Consumer Behavior. *Journal of Economic Behavior and Organization* 32, pp. 333-347.

Conlin, Michael, Michael Lynn, and Ted O'Donoghue (2003). The norm of restaurant tipping. *Journal of Economic Behavior and Organization* 52, pp. 297-321.

d'Adda, Giovanna, Martin Dufwenberg, Francesco Passarelli, and Guido Tabellini (2020). Social Norms with Private Values: Theory and Experiments. *Games and Economic Behavior* 124, pp. 288-304.

Dahl, Gordon B., Christina Felfe, Paul Frijters, and Helmut Rainer (2020). Caught between Cultures: Unintended Consequences of Improving Opportunity for Immigrant Girls. *CESifo Working Papers* No. 8045.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007). Exploiting Moral Wiggle Room: experiments demonstrating an illusory preference for fairness. *Economic Theory* 33, pp. 67-80.

Danilov, Anastasia, Kiryl Khalmetski, and Dirk Sliwka (2018) Norms and Guilt. *CESifo working Papers* No. 6999.

Dawes, Robyn M., and Richard H. Thaler (1988). Anomalies: Cooperation. *The Journal of Economic Perspectives* 2 (3), pp. 187-197.

Delfgaauw, Josse, and Robert Dur (2007). Signaling and Screening of Workers' Motivation. *Journal of Economic Behavior and Organization* 62, pp. 605-624.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2019). Estimating Social Preferences and Gift Exchange with a Piece-Rate Design. *Working Paper.*

Demiral, Elif E., and Johanna Mollerstrom (2018). The Entitlement Effect in the Ultimatum Game – Does it Even Exist? *DIW Discussion Papers* No. 1756.

Demougin, Dominique, and Claude Fluet (1998). Mechanism Sufficient Statistic in the Risk-Neutral Agency Problem. *Journal of Institutional and Theoretical Economics* 154(4), pp. 622-639.

Dhami, Sanjit (2016). *The Foundations of Behavioral Economic Analysis.* Oxford University Press, Oxford and New York.

Dimant, Eugen, and Tobias Gesche (2021). Nudging Enforcers: How Norm Perceptions and Motives for Lying Shape Sanctions. *CESifo Working Papers* No. 9385.

Dufwenberg, Martin, and Michael Lundholm (2001). Social Norms and Moral Hazard. *The Economic Journal* 111, pp. 506-525.

Dutta, Bhaskar, Matthew O. Jackson (eds.) (2003). *Networks and Groups. Models of Strategic Formation*, Springer, Berlin, Heidelberg, New York.

Ellickson, Robert C. (2001). The Market for Social Norms. *American Law and Economics Review* 3 (1), pp. 1-49.

Ellingsen, Tore, and Magnus Johannesson (2008). Pride and Prejudice: The Human Side of Incentive Theory. *The American Economic Review* 98 (3), pp. 990-1008.

Elster, Jon (1989). Social Norms and Economic Theory. *Journal of Economic Perspectives* 3 (4), pp. 99-117.

Engel, Christoph (2011). Dictator Games: A Meta Study. *Experimental Economics* 14, pp. 583-610.

Englmaier, Florian, and Achim Wambach (2010). Optimal Incentive Contracts under Inequity Aversion. *Games and Economic Behavior* 69(2), pp. 312-328.

Eriksson, Lina (2015). Social Norms Theory and Development Economics. *Policy Research Working Paper* 7450, World Bank.

Falk, Armin, Peter Andre, Teodora Boneva, and Felix Chopra (2021). Fighting Climate Change: The Role of Norms, Preferences and Moral Values. *CESifo Working Papers* No. 9175.

Falk, Armin, and Urs Fischbacher (2006). A theory of reciprocity. *Games and Economic Behavior* 54, pp. 293-315.

Farrow, Katherine, Gilles Grolleau, and Lisette Ibanez (2017). Social Norms and Proenvironmental Behavior: A Review of Evidence. *Ecological Economics* 140, pp. 1-13.

Farrow, Katherine, Gilles Grolleau, and Lisette Ibanez (2020). Designing More Effective Norm Interventions: The Role of Valence. *CEE-M Working Papers* hal-01954927.

Fehr, Ernst, and Urs Fischbacher (2004). Third-Party Punishment and Social Norms. *Evolution and Human Behavior* 25 (2), pp. 63-87.

Fehr, Ernst, Urs Fischbacher and Simon Gächter (2002). Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms. *Human Nature* 13 (1), pp. 1-25.

Fehr, Ernst, and Simon Gächter (2000a). Cooperation and Punishment in Public Goods

Experiments. *The American Economic Review* 90 (4), pp. 980-994.

Fehr, Ernst, and Simon Gächter (2000b). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* 14 (3), pp. 159-181.

Fehr, Ernst, and Simon Gächter (2002). Altruistic Punishment in Humans. *Nature* 415, pp 137-140.

Fehr, Ernst, and Klaus M. Schmidt (1999). A Theory of Fairness, Competition and Cooperation. *The Quarterly Journal of Economics* 114 (3), pp. 817-868.

Fehr, Ernst, and Ivo Schurtenberger (2018). Normative Foundations of Human Cooperation. *Nature Human Behavior* 2, pp. 458-468.

Fehr, Ernst, and Tony Williams (2018). Social Norms, Endogenous Sorting and the Culture of Cooperation. *CESifo Working Papers* No. 7003.

Fershtman, Chaim, Uri Gneezy, and John A. List (2012). Equity Aversion: Social Norms and the Desire to be Ahead. *American Economic Journal: Microeconomics* 4 (4), pp. 131-144.

Fershtman, Chaim, and Yoram Weiss (1998). Social Rewards, Externalities and Stable Preferences. *Journal of Public Economics* 70, pp. 53-73.

Festré, Agnès (2010). Incentives and Social Norms: A Motivation-Based Economic Analysis of Social Norms. *Journal of Economic Surveys* 24 (3), pp. 511-538.

Fischer, Paul, and Steven Huddart (2008). Optimal Contracting with Endogenous Social Norms. *The American Economic Review* 98 (4), pp. 1459-1475.

Fisman, Raymond, and Edward Miguel (2007). Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets. *Journal of Political Economy* 115 (6), pp. 1020-1048.

Foucault, Michel (1977). *Discipline and Punish: The Birth of the Prison.* Pantheon Books, New York.

Frey, Bruno S., and Stephan Meier (2004). Social Comparisons and Pro-social Behavior: Testing ”Conditional Cooperation” in a Field Experiment. *The American Economic Review* 94 (5), pp. 1717-1722.

Fudenberg, Drew, and Jean Tirole (1991). *Game Theory.* The MIT Press, Cambridge, London.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton (2013). Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences? *Journal of the European Economic Association* 11 (3), pp. 548-573.

Gächter, Simon, Leonie Gerhards, and Daniele Nosenzo (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review* 97, 72-86.

Gerber, Alan S., Todd Rogers (2009). Descriptive Social Norms and Motivation to Vote: Everybody’s Voting and so Should You. *The Journal of Politics* 71 (1), pp. 178-191.

Gintis, Herbert (2003). The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology* 220, pp. 407-418.

Gintis, Herbert (2009). *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction.* 2nd edition. Princeton University Press, Princeton and Woodstock.

Gittleman, Maury, and Brooks Pierce (2013). How Prevalent is Performance-Related Pay in the United States? Current Incidence and Recent Trends. *National Institute Economic Review* 226, R4-R16.

Gneezy, Uri, Andreas Leibbrandt, and John A. List (2016). Ode to the Sea: Workplace Organization and Norms of Cooperation. *The Economic Journal* 126 (595), pp. 1856-1883.

Gneezy, Uri, and Aldo Rusticini (2000). A Fine is a Price. *The Journal of Legal Studies* 29 (1), pp. 1-17.

Goette, Lorenz, David Huffman, and Stephan Meier (2006). The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups. *The American Economic Review* 96 (2), pp. 212-216.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius (2008). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research* 35, pp. 472-482.

Görges, Luise, and Daniele Nosenzo (2020). Social Norms and the Labor Market. In: Zimmerman, Klaus F. (ed.), *Handbook of Labor, Human Resources and Population Economics.* Springer, Cham.

Grewenig, Elisabeth, Philipp Lergetporer, and Katharina Werner (2020). Gender Norms and Labor-Supply Expectations: Experimental Evidence from Adolescents. *CESifo Working Papers* No. 8611.

Grossman, Gene M., and Elhanan Helpman (2021). Identity Politics and Trade Policy. *Review of Economic Studies* 88, pp. 1101-1126.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3 (4), pp. 367-388.

Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev (2017). The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance. *Journal of Public Economics* 148, pp. 14-31.

Harris, Donna, Benedikt Herrmann, Andreas Kontoleon, and Jonathan Newton (2015). Is it a Norm to Favour Your Own Group? *Experimental Economics* 18, pp. 491-521.

Hauge, Karen Evelyn (2016). Generosity and Guilt: The Role of Beliefs and Moral Standards of Others. *Journal of Economic Psychology* 54, pp. 35-43.

Helliwell, John F., Shun Wang, and Jinwen Xu (2014). How Durable are Social Norms? Immigrant Trust and Generosity in 132 Countries. *NBER Working Paper* No. 19855.

Herweg, Fabian, Daniel Müller, and Philipp Weinschenk (2018). Salience in Markets. In: Tremblay, Victor J., Roland Eisenhuth, Elizabeth Schroeder, Carol Horton Tremblay *Handbook of Behavioral Industrial Organization.* Edward Elgar Publishing, Cheltenham, Northhampton, pp. 75-113.

Hoff, Karla, Mayuresh Kshetramade, and Ernst Fehr (2011). Caste and Punishment: The Legacy of Caste Culture in Norm Enforcement. *The Economic Journal* 121, F449-F475.

Hoffmann, Elisabeth, Kevin McCabe, and Vernon L. Smith (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *The American Economic Review* 86 (3), pp. 653-660.

Holmström, Bengt (1979). Moral Hazard and Observability. *Bell Journal of Economics* 10(1), pp. 74-91.

Holmström, Bengt, and Paul Milgrom (1990). Regulating Trade Among Agents. *Journal of Institutional and Theoretical Economics* 146(1), pp. 85-105.

Holmström, Bengt, and Paul Milgrom (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization* 7 (Special Issue), pp. 24-52.

Huck, Steffen, Dorothea Kübler, and Jörgen Weibull (2012). Social Norms and Economic Incentives in Firms. *Journal of Economic Behavior and Organization* 83, pp. 173-185.

Ichino, Andrea, and Giovanni Maggi (2000). Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. *The Quarterly Journal of Economics* 115 (3), pp. 1057-1090.

Itoh, Hideshi (2004). Moral Hazard and Other-Regarding Preferences. *Japanese Economic Review* 55, pp. 18-45.

Iyer, Rajkamal, and Antoinette Schoar (2015). Ex Post (In) Efficient Negotiation and Breakdown of Trade. *American Economic Review: Papers and Proceedings* 105 (5), pp. 291-294.

Janssen, Maarten C.W., and Ewa Mendys-Kamphorst (2004). The Price of a Price: On the Crowding Out and In of Social Norms. *Journal of Economic Behavior and Organization* 55, pp. 377-395.

Jayachandran, Seema (2020). Social Norms as a Barrier to Women's Employment in Developing Countries. *NBER Working Paper* No. 27449.

Kahn, Muhammad Yasi, (2021). Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan. *Working Paper.*

Kahnemann, Daniel (1992). Reference Points, Anchors, Norms, and Mixed Feelings. *Organizational Behavior and Human Decision Processes* 51, pp. 296-312.

Kahnemann, Daniel, and Amos Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47 (2), pp. 263-292.

Kallgren, Carl A., Raymond R. Reno, and Robert B. Cialdini (2000). A Focus Theory of Normative Conduct: When Norms Do and Do Not Affect Behavior. *Personality and Social Psychology Bulletin* 26 (8), pp. 1002-1012.

Kandori, Michihiro (1992). Social Norms and Community Enforcement. *Review of Economic Studies* 59 (1), pp. 63-80.

Kassas, Bachir Mohamad (2018). *On the Social and Psychological Determinants of Cooperative and Risk-Taking Behavior.* Doctoral Dissertation, Texas A&M University.

Kessler, Judd B., and Stephen Leider (2012). Norms and Contracting. *Management Science* 58(1), pp. 62-77.

Kimbrough, Erik O., and Alexander Vostroknutov (2016). Norms Make Preferences Social. *Journal of the European Economic Association* 14 (3), pp. 608-638.

Kimbrough, Erik O., and Alexander Vostroknutov (2018). A Portable Method of Eliciting Respect for Social Norms. *Economics Letters* 168, pp. 147-150.

Kliemt, Hartmut (2020). Economic and Sociological Accounts of Social Norms. *Analyse und Kritik* 42 (1), pp. 41-95.

Koch, Simon, and Philipp Weinschenk (2021). Contract Design with Socially Attentive Preferences. *Games and Economic Behavior* 130, pp. 591-601.

Konow, James (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *The American Economic Review* 90 (4), pp. 1072-1091.

Köszegi, Botond (2014). Behavioral Contract Theory. *Journal of Economic Literature* 52 (4), pp. 1075-1118.

Kräkel, Matthias (2016). Peer Effects and Incentives. *Games and Economic Behavior* 97, pp. 120-127.

Krueger, Joachim I., Adam L. Massey, and Theresa E. DiDonato (2008). A Matter of Trust: From Social Preferences to the Strategic Adherence to Social Norms. *Negotiation and Conflict Management* 1 (1), pp. 31-52.

Krupka, Erin L., and Roberto A. Weber (2009). The Focusing and Informational Effects of Norm on Pro-social Behavior. *Journal of Economic Psychology* 30, pp. 307-320.

Krupka, Erin L., and Roberto A. Weber (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of the European Economic Association* 11(3), pp. 495-524.

Krupka, Erin L., Stephen Leider, and Ming Jiang (2017). A Meeting Of The Minds: Informal Agreements and Social Norms. *Management Science* 63 (6), pp. 1657-2048.

Kube, Sebastian, and Christian Traxler (2011). The Interaction of Legal and Social Norm Enforcement. *Journal of Public Economy Theory* 13 (5), pp. 639-660.

Kübler, Dorothea (2001). On the Regulation of Social Norms. *Journal of Law, Economics, and Organization* 17 (2), pp. 449-476.

Laffont, Jean-Jacques, and David Martimort (2001). *The Theory of Incentives: The Principal-Agent Model.* Princeton University Press, Princeton, New Jersey.

Lai, Ching-Chong, Chih-Yu Yang, and Juin-Jen Chang (2003). Environmental Regulations And Social Norms. *International Tax and Public Finance* 10, pp. 63-75.

Lemieux, Thomas, W. Bentley MacLeod, and Daniel Parent (2009). Performance Pay and Wage Inequality. *Quarterly Journal of Economics* 124, pp. 1-49.

Lindbeck, Assar (1995). Welfare State Disincentives with Endogenous Habits and Norms. *Scandinavian Journal of Economics* 97 (4), pp. 477-494.

Lindbeck, Assar (1997). Incentives and social norms in Household Behavior. *The American Economic Review* 87 (2), pp. 370-377.

Lindbeck, Assar, and Sten Nyberg (2006). Raising Children to Work Hard: Altruism, Work Norms, and Social Insurance. *The Quarterly Journal of Economics* 121 (4), pp 1473-1503.

Lindbeck, Assar, Sten Nyberg, and Jörgen W. Weibull (1999). Social Norms and Economic Incentives in the Welfare State. *The Quarterly Journal of Economics* 114 (1), pp. 1-35.

Lindbeck, Assar, Sten Nyberg, and Jörgen W. Weibull (2003). Social Norms and Welfare State Dynamics. *Journal of the European Economic Association* 1 (2/3), Papers and Proceedings of the Seventeenth Annual Congress of the European Economic Association, pp. 533-542.

Lopez-Perez, Raul (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior* 64, pp. 237-267.

Luhmann, Niklas (1998). *Die Gesellschaft der Gesellschaft.* 10th edition. Suhrkamp, Frankfurt am Main.

Mas-Colell, Andreu, Michael D. Whinston and Jerry R. Green (1995). *Microeconomic Theory.* Oxford University Press. Oxford, New York et al..

Matsui, Akihiko (1996). On Cultural Evolution: Social Norms, Rational Behavior, and Evolutionary Game Theory. *Journal of the Japanese and International Economics* 10, pp. 262-294.

Maynard Smith, J., and G.R. Price (1973). The Logic of Animal Conflict. *Nature* 246, pp. 15-18.

McAdams, Richard H., and Eric B. Rasmusen (2007). Norms and the Law. In: Polinsky, A. Mitchell, and Steven Shavell (eds.) *Handbook of Law and Economics, Volume 2.*

North Holland, Amsterdam, pp. 1573-1618.

Mengel, Friederike (2008). Matching Structure and the Cultural Transmission of Social Norms. *Journal of Economic Behavior and Organization* 67, pp. 608-623.

Michael, Robert T., and Gary S. Becker (1973). On the New Theory of Consumer Behavior. *The Swedish Journal of Economics* 75 (4), pp. 378-396.

Michaeli, Moti, and Daniel Spiro (2015). Norm Conformity Across Societies. *Journal of Public Economics* 132, pp. 51-65.

Munshi, Kaivan, and Jacques Myaux (2006). Social norms and the fertility transition. *Journal of Development Economics* 80, pp. 1-38.

Meyerson, Roger B., Gregory B. Pollock, and Jeroen M. Swinkels (1991). Viscous Population Equilibria. *Games and Economic Behavior* 3 (1), pp. 101-109.

Naylor, Robin (1989). Strikes, Free Riders, and Social Customs. *The Quarterly Journal of Economics* 104 (4), pp. 771-785.

Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevivius (2008). Normative Social Influence is Underdetected. *Personality and Social Psychology Bulletin* 34 (7), pp. 913-923.

Non, Arjan (2012). Gift-Exchange, Incentives, and Heterogeneous Workers. *Games and Economic Behavior* 75, pp. 319-336.

Non, Arjan, Ingrid Rohde, Andries de Grip, and Thomas Dohmen (2021). Mission of the Company, Prosocial Attitudes and Job Preferences: A Discrete Choice Experiment. *IZA Discussion Paper Series* No. 14836.

Nyborg, Karine, and Mari Rege (2003). On Social Norms: The Evolution of Considerate Smoking Behavior. *Journal of Economic Behavior and Organization* 52, pp. 323-40.

Nyborg, Karine, Richard B. Howarth, and Kjell Arne Brekke (2006). Green Consumers and Public Policy: On Socially Contingent Moral Motivation. *Resource and Energy Economics* 28, pp. 351-366.

Offiaeli, Kingsley, and Firat Yaman (2021). Social Norms as a Cost-Effective Measure of Managing Transport Demand. Evidence from an Experiment on the London Underground. *Transportation Research Part A: Policy and Practice* 145, pp. 63-80.

Okuno-Fujiwara, Masahiro, and Andrew Postlewaite (1995). Social Norms and Random Matching Games. *Games and Economic Behavior* 9, pp. 79-109.

Ostrom, Elinor (2000). Collective Action and the Evolution of Social Norms. *The Journal of Economic Perspectives* 14 (3), pp. 137-158.

Postlewaite, Andrew (1998). Social Status, Norms and Economic Performances. The Social Basis of Interdependent Preferences. *European Economic Review* 42, pp. 779-800.

Postlewaite, Andrew (2011). Social Norms and Social Assets. *Annual Review of Eco-*

nomics 3, pp. 239-259.

Quinn, Ben (2018). The 'Nudge Unit': the Experts that Became a Prime UK Export. In: *The Guardian*, URL: https://www.theguardian.com/politics/2018/nov/10/nudge-unit-pushed-way-private-sector-behavioural-insights-team (23.6.2022).

Raihani, Nichola J., and Katherine McAuliffe (2014). Dictator Game Giving: The Importance of Descriptive Versus Injunctive Norms. *PLoS ONE* 9 (12), e113826.

Romer, David (1984). The Theory of Social Custom: A Modification and Some Extensions. *The Quarterly Journal of Economics* 99 (4), pp. 717-727.

Rege, Mari (2004). Social Norms and Private Provision of Public Goods. *Journal of Public Economic Theory* 6 (1), pp. 65-77.

Reuben, Ernesto, and Arno Riedl (2013). Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations. *Games and Economic Behavior* 77, pp. 122-137.

Roos, Patrick, Michele Gelfand, Dana Nau, and Janetta Lun (2015). Societal Threat and Cultural Variation in the Strength of Social Norms: An Evolutionary Basis. *Organizational Behavior and Human Decision Processes* 127, pp. 14-23.

Salop, Steven C. (1979). Monopolistic Competition with Outside Goods. *The Bell Journal of Economics* 10 (1), pp. 141-156.

Schram, Arthur, and Gary Charness (2015). Inducing Social Norms in Laboratory Allocation Choices. *Management Science* 61 (7), pp. 1531-1546.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science* 18 (5), pp. 429-434.

Schurtenberger, Ivo. (2018). *Essays on Social Norms and Economic Behavior.* Doctoral Dissertation. Universität Zürich.

Sen, Amartya (1993). Capability and Well-Being. In: Nussbaum, Martha, and Amartya Sen (eds.), *The Quality of Life.* Clarendon Press, Oxford, pp. 30-53.

Sethi, Rajiv (1996). Evolutionary Stability and Social Norms. *Journal of Economic Behavior and Organization* 29, pp. 113-140.

Sethi, Rajiv, and E. Somanathan (1996). The Evolution of Social Norms in Common Property Resource Use. *The American Economic Review* 86 (4), pp. 766-788.

Sethi, Rajiv, and E. Somanathan (2003). Understanding reciprocity. *Journal of Economic Behavior and Organization* 50, pp. 1-27.

Sliwka, Dirk (2007). Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes. *The American Economic Review* 97 (3), pp. 999-1012.

Sontuoso, Alessandro (2013). A Dynamic Model of Belief-Dependant Conformity to Social Norms. *MPRA Paper* No. 53234.

Spiekermann, Kai, and Arne Weiss (2016). Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room'. *Games and Economic Behavior* 96, pp. 170-183.

Steers, Richard M., Richard T. Mowday, and Debra L. Shapiro (2004). The Future of Work Motivation Theory. *Academy of Management Review* 29 (3), pp. 379-387.

Stutzer, Alois, and Rafael Lalive (2004). The Role of Social Work Norms in Job Searching and Subjective Well-Being. *Journal of the European Economic Association* 2 (4), pp. 696-719.

Sugden, Robert [1986] (2005). *The Economics of Rights, Co-operation and Welfare.* 2nd edition. Palgrave Macmillan, New York.

Tabellini, Guido (2008). The scope of Cooperation: Values and Incentives. *The Quarterly Journal of Economics* 123 (3), pp. 905-950.

Thaler, Richard H. (2008). Mental Accounting and Consumer Choice. *Marketing Science* 27 (1), pp. 15-25.

Thøgersen, John (2008). Social Norms and Cooperation in Real-life Social Dilemmas. *Journal of Economic Psychology* 29 (4), pp. 458-472.

Traxler, Christian (2010). Social norms and conditional cooperative taxpayers. *European Journal of Political Economy* 26, pp. 89-103.

Tversky, Amos, and Daniel Kahneman (1981). The Framing of Decisions and the Psychology of Choice. *Science* 211 (4481), pp. 453-458.

Ullmann-Margalit, Edna [1977] (2015). *The Emergence of Norms.* 1st Publication in Paperback. Oxford University Press Oxford.

Viscusi, W. Kip, Joel Huber, and Jason Bell (2011). Promoting Recycling: Private Values, Social Norms, and Economic Incentives. *American Economic Review: Papers and Proceedings* 101 (3), pp. 65-70.

Voss, Thomas (2001): Game-theoretical Perspectives on the Emergence of Social Norms. In: Hechter, Michael, and Karl-Dieter Opp (Eds.), *Social Norms.* Russell Sage Foundation, New York, pp. 105-136.

Vostroknutov, Alexander (2020). Social Norms in Experimental Economics: Towards a Unified Theory of Normative Decision Making. *Analyse und Kritik* 42 (1), pp. 3-39.

Vu, Thi-Thanh-Tam (2019). *Identification, Signaling and Exploitation of Social Preferences. An experimental Analysis.* Doctoral Dissertation. University of Trento.

Walker, Peter (2021). Call Centre Staff to be Monitored via Webcam for Home-working 'Infractions'. In: *The Guardian,*
URL: https://www.theguardian.com/business/2021/mar/26/teleperformance-call-centre-staff-monitored-via-webcam-home-working-infractions (22.06.2022).

Weibull, Jörgen B., and Edgar Villa (2005). Crime, Punishment and Social Norms. *SSE/EFI*

Working Paper Series in Economics and Finance No. 610.

Wenzel, Michael (2005). Misperceptions of Social Norms about Tax Compliance: From Theory to Intervention. *Journal of Economic Psychology* 26, pp. 862-883.

Whitson, Jennifer, Cynthia S. Wang, Joongseo Kim, Jiyin Cao, and Ales Scimpshire (2015). Responses to Normative and Norm-Violating Behavior: Culture, Job Mobility, and Social Inclusion and Exclusion. *Organizational Behavior and Human Decision Processes* 127, pp. 24-35.

Williams, Linda Stallworth (2008). The Mission Statement. A Corporate Reporting Tool With a Past, Present, and Future. *Journal of Business Communication* 45 (2), pp. 94-119.

Williamson, Oliver (1985). *The Economic Institutions of Capitalism.* Free Press, New York.

Wilson, Bart J. (2010). Social Preferences aren't Preferences. *Journal of Economic Behavior and Organization* 73, pp. 77-82.

Xiao, Erte (2013). Profit-seeking Punishment Corrupts Norm Obedience. *Games and Economic Behavior* 77, pp. 321-344.

Yeomans, Mike, and David Herberich (2014). An Experimental Test of the Effect of Negative Social Norms on Energy-efficient Investments. *Journal of Economic Behavior and Organization* 108, pp. 187-197.

Young, H. Peyton (1993). The Evolution of Conventions. *Econometrica* 61 (1), pp. 57-84.

Young, H. Peyton (1998a). Social Norms and Economic Welfare. *European Economic Review* 42, pp. 821-830.

Young, H. Peyton (1998b). *Individual Strategy and Social Structure. An Evolutionary Theory of Institutions.* Princeton University Press, Princeton.

Young, H. Peyton (2015). The Evolution of Social Norms. *Annual Review of Economics* 7, pp. 359-387.

Curriculum Vitae
Simon Koch

Work Experience

11/2021-03/2022	Teaching Assignment for Game Theory University of Kaiserslautern
11/2015-10/2021	Research Associate Chair of Microeconomics Prof. Dr. Philipp Weinschenk University of Kaiserslautern

Education

10/2012-09/2014	Master of Arts (M.A.) University of Erfurt Governance and Public Policy
10/2008-05/2012	Bachelor of Arts (B.A.) University of Passau Governance and Public Policy
09/1998-06/2007	Abitur Gymnasium Weilheim i. OB.