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

DISCUSSION PAPER SERIES

IZA DP No. 12782

Incorporating Conditional Morality into Economic Decisions

David Masclet
David L. Dickinson

NOVEMBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12782

Incorporating Conditional Morality into Economic Decisions

David Masclet
Université de Rennes 1, CREM and CNRS

David L. Dickinson
Appalachian State University, IZA and ESI

NOVEMBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Incorporating Conditional Morality into Economic Decisions

We present a framework that incorporates both moral motivations and fairness considerations into utility. The main idea is that individuals face a preference trade-off between their material individual interest and their desire to follow moral norms. In our model, we assume that moral motivation is conditional and may be influenced by others' actions. Specifically, in our framework moral obligation is a combination of two main components: an autonomous component and a social influence component that captures the influence of others. Our framework is able to explain many stylized results in the literature and to improve theories of economic behavior.

JEL Classification: B3, D6, D9

Keywords: fairness, ethical decision making, moral motivation, behavioral economics

Corresponding author:
David L. Dickinson
Economics Department
Appalachian State University
Boone, NC 28608
USA

E-mail: dickinsondl@appstate.edu

1. Introduction

In the early history of the profession, it was common to include social concerns (e.g. Smith, 1759) and moral values (see for example Smith, 1759; Edgeworth, 1881). These authors pointed out that people often care about others, have moral ethics, and that this may have important economic consequences. However, most economists still routinely assume that people are motivated only by their own material self-interest and do not care about "social" considerations nor "moral" values. This sharply contrasts with the overwhelming empirical evidence, in particular from experiments, showing that individuals do have social preferences and care about others (see Fehr and Schmidt. 2006 for a discussion). Other studies have shown that a non-negligible proportion of individuals appear to have moral concerns that promote honesty even when the material gains from dishonesty outweigh the material incentives associated with honesty (e.g. Gneezy, 2005; Mazar et al. 2008b). In the same vein, some studies indicate that people tend to balance the competing forces of *Homo Economicus* and unconditional honesty by being only partially dishonest (Shalvi et al. 2011).

It is only relatively recently that a few papers have attempted to explicitly formalize the notions of fairness and social preferences in response to the observed behavior across different experiments within the rational choice framework (see for instance Arrow, 1981; Samuelson, 1993; Sen, 1995 and more recently Fehr and Schmidt 1999; Bolton and Ockenfels, 2000; Rabin, 1993).[1] In parallel, some authors have attempted to incorporate moral motivation into theoretical framework (e.g. Arrow, 1973; Laffont, 1975; Sen 1995; Nyborg, 2000; Brekke et al, 2003; Alger and Renault, 2006; Benabou and Tirole, 2011; Roemer, 2010; Figuieres et al, 2013; Alger and Weibull, 2013).

In this current paper, we present an original framework that attempts to combine both moral motivations and fairness considerations. Our model is based on two simple ideas. The first idea is that individuals face a trade off in their utility function between their material individual interest and their desire to follow moral norms.[2] On the one hand, individuals want to maximize their own material payoffs. On the other hand, they would like to "do the right

[1] In these models, preferences do not only depend on material payoffs but also on intentions (whether the behavior of others is fair or unfair). A reciprocal individual, as we define it here, responds to actions he perceives to be kind in a kind manner, and to actions he perceives to be hostile in a hostile manner.

[2] Moral norms may consist in the norm of equal sharing or the norm of "'everyday Kantianism' (do what would be best if everyone did the same)" such that any deviation from the norm may trigger guilt.

thing" by acting morally in reference to Kant's (1785) categorical imperative.[3] Individuals may attribute different weights to material payoffs and satisfaction derived from morality such that at one end of spectrum is *Homo Economicus* who acts so to maximize her own monetary payoff. At the opposite end is *Homo Kantis* who always hold to their (Kantian) moral imperative despite the possible costs of doing so. The second idea behind our model is that morality is conditional in the sense that it is influenced by the observation of others and by fairness considerations. Specifically, in our model moral obligation is a combination of both an autonomous moral imperative component and a social influence component. Social influence is embodied in a fairness function in the vein of Rabin's (1993) kindness function. In other words, individuals will not necessarily stick to their ideal moral target but rather are prone to revise it upward (downward) when they observe that others treat them kindly (badly). Consequently, our model is compatible with Rabin's (1993) intention-based theory according to which an individual who feels kind (hostile) others' intentions will be willing to reward (hurt) others. Our model is also close to Figuieres et al. (2013), who consider that moral motivation is weakened as one's judgment about right behavior (rooted in intrinsic moral ideals) is influenced by observed behavior of others. [4] A notable difference with Figuieres et al. (2013) is that our model incorporates the role of fairness considerations in the utility function. A recent review study (Abeler et al, 2019), offers support for both the notion that honest behavior observed across many experimental studies is due both a preference for honest as well as influences of others as one often wishes to be viewed as being honest.

The rest of the paper is organized as follow. In Section 2 we present our model of conditional morality coupled with fairness considerations. Section 3 applies this model to different games from experiments. Section 4 discusses potential extensions, objections to our model and compare our model with alternative approaches in the literature. Finally section 5 concludes.

2. A Simple Model of Conditional Moral Motivation

[3] The categorical imperative was introduced in Kant's 1785 Groundwork of the Metaphysics of Morals. It is the central philosophical concept in the deontological moral philosophy of Immanuel Kant. According to Kant, morality can be summed up in an imperative from which all obligations derive. A categorical imperative denotes an absolute, unconditional requirement that must be obeyed in all circumstances and is justified as an end in itself.
[4] Figuieres et al. (2013) develop a model that accounts for the decay of the average contribution observed in experiments on voluntary contributions to a public good.

Here, we present a framework that incorporates both moral motivations and fairness considerations into one's preferences. Consider the following utility function:

$$U(a) = b(a) - c(a) - v(a - \hat{a}) \quad (1)$$

Here, a is an action that generates both benefits, b, and costs, c, and belongs to the set $A_i = [a_i^{min}, a_i^{max}]$. The morality component of the utility function is captured by $v(a - \hat{a})$ where $\hat{a}$ describes one's moral imperative, $\hat{a}_i \in A_i$. Deviations of one's action from this moral imperative generate disutility (e.g., Nyborg, 2000; Brekke et al, 2003; Figuieres et al, 2013).[5] Both material benefits, $b(a)$ and costs $c(a)$ increase in the action, a, with $b(a)$ increasing at a decreasing rate but $c(a)$ increasing at an increasing rate: $b' > 0$, $c' > 0$, $b'' < 0$, $c'' > 0$. The disutility of deviations in either direction from one's moral ideal are captured by assuming $v'_a > 0$ if $a > \hat{a}$, $v'_a < 0$ if $a < \hat{a}$, and $v'_a = 0$ if $a = \hat{a}$. We also assume that $v''_{aa} > 0$ such that marginal disutility increases at an increasing rate as one's action gets further from the moral obligation, and $v''_{a\hat{a}} < 0$. The interpretation of this condition is that a small increase in one's moral obligation raises the marginal benefit to increased action (i.e., a reduction in moral disutility by moving one's action closer to the moral target).

This approach to formulating preferences as a function of one's "action" is quite general. Though we assume higher levels of the action generate material benefits and costs, the morally better action may, in general, be a higher or lower level of a. Note that our framework also implies that a change in one's moral imperative, $\hat{a}$, ceteris paribus, will increase or decrease utility as the imperative moves farther or closer to one's action, respectively. This is, in essence, the idea behind cognitive dissonance theory and how cognitive dissonance can be reduced by altering one's view of appropriate behavior.

[5]One may interpret this loss of utility attached to violation of moral norm in terms of guilt (See the typologies in Elster (2009) and Kandel and Lazear (1992) for a distinction between shame and guilt.). While shame is elicited by the presence of contempt in some observer, guilt does not depend on the fact of being observed. It is elicited when agents contemplate possible norm violations or when they remember past violations. In a sense guilt consists in the internalization of the observation by others. Interestingly, this moral ideal function may also be interpreted in relation to the literature on peer pressure (e;g. Kandel and Lazear, 1992) as well as literature on inequality aversion (Fehr and Schmidt, 1999). Indeed, by relaxing a few hypotheses and considering that the ideal moral relies on interpersonal comparisons, one may replace the moral ideal by the average effort of others (e.g. Kandel and Lazear, 1992). Alternatively, one may also relate this model to models of inequality aversion by assuming that $\hat{a}$ corresponds to another individual j's action and that v is quadratic in the action such that any deviations from $\hat{a}$ in either direction generate disutility.

Following Figueres et al (2013), we assume that the moral obligation component, $\hat{a}$, includes both a Kantian categorical imperative (Lafont, 1975; Harsanyi 1980), which we denote *K*, and a component that is a function of social influence and fairness considerations in the spirit of Rabin (1993), which we denote $F(a_j)$, where a_j is the action of others.[6] This moral obligation function can therefore be written as:

$$\hat{a_i} = \hat{a}_i\left(K_i, F_i(a_j)\right), j \neq i \tag{2}$$

This moral obligation function satisfies the following intuitive properties:

Assumption 1: $\frac{\partial \hat{a_i}}{\partial K_i} = \hat{a}'_K \geq 0$
(one's moral obligation is weakly increasing in the Kantian imperative)

Assumption 2: $\frac{\partial \hat{a_i}}{\partial F(a_j)} = \hat{a}'_F \geq 0$
(one's moral obligation is weakly increasing in the perceived fairness of others' behavior)

Essentially, Assumption 2 states that if the individual *i* observes a (un)kind action of individual *j*, she will (decrease) increase her moral imperative. This framework implies the worker will wish to reciprocate a kind action or mimic observed kind actions with higher effort to reduce the disutility of choosing effort below her revised higher moral obligation level. Here we endogenize the role played by social influence by incorporating fairness or moral conformity considerations into the $F(a_j)$function.

An example of a conditional moral motivation function is the following:

$$\hat{a}_i = (1-\theta_i)K_i + \theta_i F(a_j), \theta_i \in [0,1], \qquad \forall\, i,j, i \neq j$$

$$= K_i + \theta_i\left[F_i(a_j) - K_i\right] \tag{3}$$

The weight θ_i may be interpreted as the conditionality of individual *i*'s moral motivation. If $\theta_i = 0$ individual *i* has strong unconditional moral motivation: he never deviates from his ideal moral intrinsic target K_i whatever the observed action of others. This may be also the case when

[6] While Rabin (1993) considers one's "belief" of how kind someone is, beliefs may be replaced with an actual signal of kindness (see Dickinson, 2000) in the case of our workers who make effort choices with full knowledge of the wage offer.

the individual cannot observe others. At the other extreme, a player for whom θ_i is close to 1 is strongly influenced by others and will strongly revise her initial moral ideal through $F(a_j)$ whenever her current action differs from observed action of others or from one may be considered acceptable moral consensus. Assuming the above revision rule, both a strongly reciprocal player as well as an individual prone to mimicry are defined as those for whom θ_i is close to one, while those not susceptible to be influenced are defined by θ_i equal to zero. Finally a pure *homo economicus* will not be affected by moral motivation at all, such that (in this case, we also define $v(\hat{a}) = 0$).

Finally, the $F(a_j)$ function is close to Rabin's kindness function. Specifically, we define $F(a_j)$ as follows:

$$F(a_j) = \frac{(a_j - a_j^{min})}{(a_j^{max} - a_j^{min})}[a_i^{max} - a_i^{min}] + a_i^{min} \tag{4a}$$

Where a_j is individual *j*'s action in the set that contains all possible actions from minimal to maximal, $a_j \in A_j = [a_j^{min}, a_j^{max}]$. Similarly, a_i will be within the set $A_i = [a_i^{min}, a_i^{max}]$. The function $F(a_j)$ considers the location of player *j*'s action within the set of all possible actions for that individual. Because action sets between individuals *i* and *j* may differ, $A_i \neq A_j$, we normalize action location in equation (4a) by translating it into what a comparable level of action would be for individual *i* in order to mimic the action of individual *j*.[7]

If player *i* feels he is treated badly by player *j* in an interactive decision environment, such that $F(a_j) < K_i$, she will revise downward her moral ideal obligation to an extent determined by the weight θ_i. Alternatively, player *i* will positively reciprocate a fair action by upwardly revising her moral motivation when $F(a_j) > K_i$. Interestingly, here player *i* evaluates player *j*'s degree of (un)fairness in reference to her own moral motivation. This contrasts with Rabin (1993) where (un)fairness is interpreted in comparison to the average payoff as the focal point. Additionally, our approach to fairness or moral motivations may be applied to individual

[7] In the case where the direction of actions differs across players, we assume player *i* considers the mirror-definition for comparability within (4a). For example, in the case of a zero-sum game one party's demand, which increases her payoff, decreases the others monetary payoff. A high demand by the worker must then be interpreted by the employer as a low action within the set of actions that increase the employer's monetary payoff.

choice where others' actions provide the comparison for potential mimicry (or, the influence of peer pressure). Within this framework, one may even selectively choose which "others' actions" to consider such that individuals may be more or less resistant to revising their moral benchmark.

A particular case of this function in (4a) is the case where an individual's action set is the same as that of the comparable other individual(s), $A_i = A_j$. In this case (4a) simplifies to:

$$F(a_j) = a_j \tag{4b}$$

This simplified function corresponds to the case where social influence takes place via mimicry considerations because of the comparability of the action spaces across individuals within the environment being considered.

3. Predictions across decision environments

While not intended to be exhaustive, we use this section to illustrate applications of the theory. Both individual choice and interactive decision environments are considered, and we discuss the descriptive success of our framework in explaining stylized results from empirical and experimental studies.

3.1 Predictions in the context of unethical activities

A natural application of our model is the context of unethical activities. Unethical behaviors is a major concern of modern societies including cheating in exams, fare dodging, CV inflation or sabotage at work. For instance, dishonesty is common in the workplace and is at the heart of the principal-agent problem.[8] Unethical behavior often results in high costs for the entire society. It raises transaction costs, weaken social cohesion, harm firm performance by discouraging effort, and ultimately reduce the freedom of citizens and impoverish society. Governments and organizations spend considerable resources to detect dishonest behavior and

[8] Both moral hazard and adverse selection are examples of dishonesty on the workplace where one party may take advantage of information asymmetry to conceal the truth, at the expense of another party.

to implement coercive measures, which may be detrimental in terms of freedom and social welfare.

3.1.1.The cheating game

Economists have been working on the determinants of dishonesty for decades. In the so-called economics-of-crime approach, cheating activities result from a comparison of the expected benefits and costs of fraudulent actions (Becker, 1968). Standard economics assumes that people cheat when it is in their material interest to do so. It depends on the probability of being caught when it is possible to detect cheating and on the cost associated with this detection. In Becker's framework, an increase in the probability of apprehension or in the severity of punishment reduce the incentive to engage in the illegal market. The deterrent effect of punishment suggests that a marginal increase in expected punishment *ceteris paribus* reduces the propensity to engage in a unethical activity by increasing its relative price (Becker, 1968).[9]

Models of tax compliance rely on Becker's model where individual are rational acting on self-interest; they optimize their expected utility and choose illegal activity if rewards exceed the expected cost in term of probability of detection associated with the penalty. For example, in the domain of taxation Allingham and Sandmo (1972) and Yitzhaki (1974) have developed models in which the taxpayer will comply or evade taxes depending on the tax rate, the probability of audit and the amount of the fine in case of an audit. In these models the decision to cheat depends on the extrinsic costs associated with dishonesty related to the probability of being caught and the punishment resulting from it.

However, these standard models do a poor job in explaining dishonest behavior because they predict more cheating than what is usually estimated (see an overview provided by Abeler et al., 2019). Indeed empirical data sometimes seems at odds with this viewpoint. For example, most studies on tax compliance find higher compliance rates than predicted by models that are only based on material incentives like audit and penalty rates, and studies also find that social and institutional factors matter (Andreoni et al 1998; Torgler 2002). Evidence for honest behavior has also been shown in studies in the labor market using field studies and field experiments (Evans et al. 2001; Nagin et al. 2002; Grover and Hui 2005).

[9] In addition to the deterrence effect there also exists another effect, namely the incapacitation effect, i.e. the fact that removing criminals from the illegal market will mechanically reduce crimes. (Levitt, 1996 and Kessler and Levitt, 1999). In this paper we will not focus our attention on this effect.

This suggests that the decision to commit a dishonest action does not only depend on the extrinsic costs associated with cheating but also depends on many other determinants. In particular, intrinsic costs of cheating, (i.e. costs that are not based on strategic considerations) should also be incorporated in economic models to increase their validity. Such intrinsic costs may result from a pure distaste of cheating because of guilt aversion (Battigalli et al. 2013) or self-image (Bénabou and Tirole, 2006). For instance, it has been shown that most individuals do not fully exploit their opportunities of lying, probably because they are willing to preserve a good self- or social image (Mazar *et al*., 2008a, 2008b; Fischbacher and Föllner-Heusi, 2012). Intrinsic costs may also result from aversion to violate social norms (Elster, 1989). For example, tax compliance is influenced by peer effects and social conformity (Fortin et al, 2007), emotions (Corricelli et al., 2010).

Some individuals may incur high intrinsic cost of cheating such that they may always behave honestly, regardless of their material benefits from cheating ("ethical individuals").[10] At the other extreme, economic agents always cheat in the absence of extrinsic costs because their intrinsic cost of cheating is assumed to be null. Between these two extremes cases, the majority of individuals may be conditionally honest and only cheat if the benefits outweigh the associated intrinsic costs.

In this section we show that including moral motivation may explain such phenomena in the domain of dishonesty or illegal activity. Specifically, we assume that individual not only compare monetary costs and benefits associated to the illegal activities, but they are also influenced by moral values when engaging in the illegal choices. To keep matters simple, consider a rational and risk neutral individual. Let $B = ba$ be the illegal return or "benefit", where a represents the unethical activity and b is the marginal return on unethical or illegal activities. We assume that the disutility of effort $c(\cdot)$ is increasing and convex in effort level such that $c' > 0$ and $c'' > 0$. Assuming expected utility is separable in output and effort cost and linear in output, we have the following utility function for individual i who exerts an effort towards illegal activity a_i (suppressing subscripts i):

$$\begin{aligned} EU(a) &= p(ba - sa) + (1-p)(ba) - c(a) \\ &= ba - psa - c(a) \end{aligned} \tag{5}$$

[10] This is in line with the views of St. Augustine (354-430 AD) and Kant (1787) who advocated such an uncompromising approach to the analysis of cheating.

The first term of (5), ba, is the return rate of illegal activities for each unit of effort *a*, psa corresponds to the expected cost of being caught with the sentence (or sanction) $s \in [0,1]$ if the individual is caught with a probability $p \in [0,1]$. Assume that cost of effort is $c(a) = a^2$. The optimal choice requires that the choice for each agent (*i, j*) maximizes utility as given by equation (6) below. For agent *i*, the effort choice thus solves the following problem:

$$maxEU_{a_i}(ba_i - psa_i) - c(a_i) \quad i,j = 1,2;\ \forall i \neq j \tag{6}$$

We obtain the following first-order condition (FOC) with respect to effort level in the illegal market.

$$b - ps - c'(a) = b - ps - 2a = 0 \tag{7}$$

We can solve this first order condition to obtain the optimal level of illegal activity:

$$a^* = \frac{b-ps}{2} \tag{8}$$

Next consider the case where each agent *i*'s preferences include moral motivations:

$$EU(a):\ ba - psa - c(a) - v(\mathrm{a} - \hat{a}) \tag{9}$$

Here, the term, $v(\mathrm{a} - \hat{a})$, is agent *i*'s moral motivation where $\hat{a}$ stands for the moral motivation defined in equation (2), $\widehat{a_i} = \hat{a}_i\left(K_i, F_i(a_j)\right)$. The idea here is that observing others, *j*, who possibly behave immorally may influence individual *i*'s moral obligation, $\widehat{a_i}$. Indeed, while the standard economic perspective postulates that unethical activities vary with the 'price' of crime, which depends on the severity of punishment and the detection probability, our model is based on the idea that unethical activities are also influenced by social interaction phenomena, such as 'peer pressure' or 'neighborhood effects (Falk and Fischbacher, 2002).[11] This is consistent with empirical studies on tax compliance that have reported negative effect of observability—

[11] Glaeser et al. (1996), e.g., identify social interaction as an important determinant of criminal activity. Similarly, Ludwig et al. (2001) argue that the opportunity to move to lower-poverty neighborhoods reduces criminal behaviour by teens. Similarly, Case and Katz (1991) report that an individual's probability to be involved in crime varies positively with the proportion of other youths that are involved in crime.

seeing other group members cheating profusely may incite individuals to cheat more due to mimicry (see Fortin et al. 2007). Several studies provide evidence for the conditionality of unethical behavior. People who observe that others are violating a certain social norm or legitimate rule, are more likely to violate it themselves (Keizer et al. 2008). Using the strategy method, Falk and Fischbacher (2002) show that subjects' willingness to steal increases with others' stealing. Gino et al. (2009) show that observing someone cheating increases own cheating if the cheater is from the same university. Abeler et al. (2014) find that subjects who believe that others cheat are more likely to cheat. Based on a two-period setting, Kroher and Wolbring (2015) and Diekmann et al. (2015) find that notifying subjects participating in a die rolling task with others' unethical behavior after the first roll partly increases cheating in the second roll.

Let us also assume that the moral motivation is captured by quadratic function such that $v_i(a_i - \widehat{a_i}) = (a_i - \widehat{a_i})^2$. We obtain the following F.O.C. with respect to effort level in the illegal market.

$$b - ps - 4a_i + 2\widehat{a_i} = 0$$

Now we have the optimal level of illegal activity a_i^* as:

$$a_i^* = \frac{b - ps + 2\widehat{a_i}}{4} \tag{10}$$

We can see that the optimal activity level in the illegal market is now an increasing function of the moral obligation, $\frac{\partial a^*}{\partial \hat{a}} > 0$, which both depends on the autonomous moral component *K* and the component describing social influence through peers, $F_i(a_j)$. Thus, peers may negatively influence one's activity in the illegal market, which is consistent with the phenomenon of recidivism.

The cheating game is a particular case of the Becker's utility function where subjects face no threat of being caught individually. In the classical cheating game of Fischbacher and Föllmi-Heusi (2013) a participant rolls a six-sided die in private, and then reports the outcome to the experimenter. The participants' payoff depends on their report: they receive 1, 2, 3, 4, 5, 0 Swiss Francs for reporting 1, 2, 3, 4, 5, 6, respectively. In the game, the participant is told that if she would report a 5, she would receive $X; otherwise, she would earn $0. Fischbacher and

Föllmi-Heusi (2013) found that about 20% of inexperienced subjects report an outcome consistent with lying to the fullest extent possible while 39% of subjects appear fully honest. In addition, a high share of subjects consists of partial liars; these subjects report outcomes consistent with lying, but do not report the payoff-maximizing draw. Honesty has been studied in a variety of other experimental settings as well (e.g., Mazar et al, 2008b; Gneezy, 2005; Abeler et al., 2016; 2019) and in field data (e.g., Pruckner and Sausgruber, 2013; Gächter and Schulz, 2016). In general, findings are consistent with the fact that observed honesty is due, not surprisingly, to preferences for honesty or preferences that others view one as being honest (see recent meta-analysis in Abeler et al, 2019).

Our model is consistent with these findings regarding dishonest behaviors, even when there is no probability of detection. For some people, cheating or lying is intrinsically costly (i.e., a high moral ideal, *K*, in their moral obligation function) and therefore they prefer honesty over dishonesty. In principle, based on our model we can distinguish between three types of people according to their moral motivation. Some people may be unwilling to tell a lie, regardless of their benefit from it ("ethical type"). At one extreme are people with a zero moral motivation ($\widehat{a_i} = 0$, classic "homo economicus"). At the other extreme are people who have high moral motivations. Among those are some with high moral imperatives and who are not influenced by others (high *K*, $F_i(a_j) = 0$). Such individuals may never cheat. Finally, we have the conditional cheaters who will revise their moral ideal based on the observation of others high (*K>0*, $F_i(a_j) > 0$).

In the cheating game, with no disutility of effort and no probability of detection, we can simplify the utility function as follows:

$$U(a) = ba_i - v(a_i - \hat{a}_i)$$

with the FOC:

$$b - 2a_i + 2\widehat{a_i} = 0 \tag{11}$$

We then have optimal effort given by:

$$a_i^* = \frac{b+2\widehat{a_i}}{2} \tag{12}$$

If $\widehat{a_i} = 0$, i.e. if the individual is intrinsically honest, then $a_i^* = \frac{b}{2}$. This suggests a greater than zero level of optimal dishonestly when there are positive benefits to dishonesty, b. An individual who is predicted to never be dishonest must therefore perceive zero benefits to dishonesty and/or have a strong moral obligation of honesty in our model.[12]

3.1.2 The money burning game

In this section we apply our model to unethical activities that involve no monetary benefits nor probability of being caught. Unethical behavior within organizations is not rare and often results in high costs for the entire society. In economics, studies focusing on the antisocial dimension of behaviour include the seminal studies by Zizzo and Oswald (2001) and Zizzo (2004), whose results show that many subjects are willing to incur a real cost in order to reduce other's payoffs—"money burning". Money burning games allow researchers to test whether subjects are willing to pay for reducing other people's income in the context of laboratory money burning experiments. In a classical money burning experiment, each participant is randomly matched with another participant and has the opportunity to reduce the other's payoff. Most of the time this decision is costly for the burner.[13] The seminal studies by Zizzo and Oswald (2001) and Zizzo (2004), find that many subjects are willing to pay for reducing other people's income, mainly to close a disadvantageous income gap. More recently, Abbink and Sadrieh (2009) and Abbink and Herrmann (2011) remove the inequality aversion motive from their joy-of-destruction game, and still find destruction frequencies of up to 40 per cent.

For simplicity we assume here that burning decisions are costly both to the target and the decider. The individual has utility defined by the cost of the burning decision and the moral obligations to burn (or not burn) money. Utility for player *i* is defined as:

[12] Even with positive material benefits to cheating, one may have strong enough moral standard to not cheat that the option action is to not cheat. This would be the case if $a_i^* < 0$. In this game, the implication would be a "negative" cheating moral obligation that results in a corner solution of $a_i^* = 0$. Intuitively, if one's standard is so strongly against cheating that one would *negatively* cheat if possible, then the individual will likely not cheat even in the presence of at least some level of material benefits to cheating.

[13] Note that the money burning game could be considered as a modified version of the cheating game in absence of benefit $b \cdot a = 0$ and without monitoring nor punishment ($(psa = 0)$.

$$U_i = -c(b) - v\left(b - \hat{b}\right) \tag{13}$$

where $c(b)$ is the cost for individual *i* of burning resources of individual *j*. The last term represents moral loss of utility due to deviating from one's moral idea in terms of money burning, $\hat{b}$. As before, we assume $\hat{b}$ is a moral obligation function that is a combination of an autonomous Kantian categorical imperative component, K, and a social component $F(c)$, $\hat{b}_{ij} = \hat{b}_{ij}\left(K_i, F_i(c)\right)$. For simplicity, assume that the Kantian component is to not burn resources such that K=0.

Each individual chooses her optimal money burning level by maximizing (13) which yields the following FOC:

$$\frac{\partial U}{\partial b}: \ -c'(b) - v'\left(b - \hat{b}\right) = 0 \tag{14}$$

The first term of each first-order condition is negative and corresponds to the marginal cost of burning money, while the second term corresponds to the disutility from deviation from one's moral ideal in terms of burning. Again we assume the moral motivation function is quadratic such that $v_i\left(b_i - \widehat{b_\iota}\right) = \left(b_i - \widehat{b_\iota}\right)^2$. We can rewrite the FOC as follows:

$$\frac{\partial U}{\partial b}: \ -c'(b) - 2b + 2\hat{b} = 0 \tag{15}$$

If $\hat{b} = 0$ we get a corner optimum such that $b^*(\hat{b}) = 0$.

Interestingly, our model allows us to account for pure nastiness in money burning games, which would mean that nasty individuals have $\hat{b}_{ij} > 0$ resulting either from a positive categorical imperative to burn resources ($K > 0$) or due to an unkind act of others $F(c) > 0$ that triggers negative reciprocity (e.g., Abbink and Sadrieh, 2009). In this case we get:

$$b^*\left(\hat{b}\right) = \frac{-c'}{2} + \hat{b} \tag{16}$$

This means that for sufficiently high values of $\hat{b}$ such that $\hat{b} > \frac{c'}{2}$ (i.e., the moral obligation must sufficiently outweigh the marginal cost of burning resources), individuals may be incited to burn others' resources.

3.2 Predictions in the context of ethical activities and social dilemma

After investigating the more dark side of human dishonesty in the previous section, we now turn our focus to games in which participants can express their willingness to redistribute, cooperate or behave in a kind or ethical manner.

3.2.1 The Dictator Game

The dictator game is a popular game in experimental economics, though it has been applied outside of economics as well. The game is a derivative of the ultimatum game (Güth et al, 1982) first studied in Forsythe et al, (1994). The term "game" is a misnomer because it captures a decision by a single player: the dictator who can decide to send money to another player or not. In absence of moral concerns, player *i* should keep all her endowment for herself. The experimental results indicate, however, that a non negligible number of players choose to send money, which undermines the assumption of narrow self-interest. The give rate varies between 20 and 30% of the endowment (see Engel, 2011 for a meta study). Our theoretical model is consistent with such findings.

Suppose that individual *i* with endowment ω chooses her action, a, which is how much to give to the other player. We can write player *i*'s utility function with moral motivation as follows:

$$U(a) = b(a) - c(a) - v(a - \hat{a})$$

$$= (\omega - a) - v(a - \hat{a}) \qquad (17)$$

Here, $(\omega - a)$ describes the monetary payoff (net benefit) to individual i of each possible action, $a \in [0, \omega]$. For simplicity, we again assume that *i*'s moral motivation is captured by a quadratic function such that $v_i(a - \hat{a}_i) = (a - \hat{a}_i)^2$

The FOC is (suppressing all *i* susbcripts):

$$\frac{\partial U}{\partial a} = -1 - 2a + 2\hat{a} = 0 \qquad \text{such that}$$

$$a^*(\hat{a}) = \hat{a} - \frac{1}{2} \qquad (18)$$

From (18) we can easily see that, in the absence of moral motivations (v=0), or if one's moral obligation is to offer nothing (i.e. $\hat{a} = 0$), we have a corner solution and the individual should keep all her endowment for herself. The amount sent can only be positive only if $\hat{a} > 0$, and the optimal offer (action) is increasing with one's moral obligation. For instance, if the ideal moral obligation $\hat{a}$ is the "the norm of payoff equality" then, for a \$10 pie we have $\hat{a} = \$5$ and the optimal offer is $a^* = \$4.50$. That is, the dictator offers a bit less than her moral obligation due to the monetary payoff costs that enter into the decision.[14]

3.2.2. Gift Exchange Game

The gift exchange game of Akerlof (1982) was first studied experimentally by Fehr et al (1993). The gift exchange game is a two-player sequential move game that consists of two stages. In the first stage, a "firm" offers a wage, $w \in [20,120]$ to her ``worker''. In the second stage, the worker has to choose an "effort level" $e_i \in [0.1,1]$. The higher the effort level, the higher are the associated effort costs, c(e). A high wage "gift" is presumably reciprocated by the worker in the form of higher than minimal effort.

In absence of moral motivation, and under the assumption of common knowledge of rationality and selfishness, standard theoretical prediction are straightforward. Backwards induction dictates minimal effort in stage 2 given that effort is costly. Firms anticipate this and therefore offer the lowest wage possible in stage 1. The equilibrium of this game with selfish and rational players is a minimum wage – minimum effort pair of decisions, [w=20, e=0.1]. Despite the predictions of standard theory, a large body of experimental evidence in support of reciprocity has been reported in the past two decades. General finding that effort (either a monetary transfer or real effort on a task) is positively correlated with the size of the wage. One of the first experiments to test this assumption is Fehr et al. (1993), who constructed a market

[14] A feature of the predicted outcomes for dictator games is that, all else held equal, one's offer will converge upon one's moral obligation as the stakes of the game rise. This prediction may seem less intuitive given that it predicts that a dictator with $\hat{a}$ fixed at 50% of the pie will offer 25% of a \$2 pie, but 45% of a \$5 pie (49.5% of a \$100 pie) given equation (18). However, this results from a fixed percentage of the pie carrying a relatively more important monetary utility loss for smaller pie sizes. Alternatively, one's moral obligation may quite naturally be a function of the size of the pie. For example, Dickinson (2000) notes that for higher stakes *ultimatum* games, which differ by giving the recipient the opportunity to reject an offer and generate a zero payoff to both, the recipient is willing to accept lower offers as the stakes of the game grow (this result is derived as an application of Rabin (1993) to the ultimatum game). Thus, dictators may feel a reduced moral obligation, in terms of *percentage* offer, even when rejection is not possible because there may be a sense that lesser percentage offers are morally acceptability for higher stakes games.

with excess supply of labor, ensuring a low equilibrium wage. The authors found that most employers attempted to induce employees to invest greater effort by offering them higher (at times by more than 100%) than market-clearing wages. On average, this high wage was reciprocated by greater employee effort, making it profitable for employers to offer high wage contracts. Subsequent laboratory exercises have largely led to similar conclusions (Fehr et al, 1993; Fehr et al, 1998). Another important finding is that effort level remains positive even for low wages (Brüggen and Strobel, 2007, Gneezy and List, 2006). Several empirical studies including field and lab experiments have shown that, despite the absence of any penalty for shirking, workers do not hesitate to exert a positive effort under a flat wage scheme (Falk and Ichino, 2006; Mas and Moretti, 2009; Dohmen and Falk, 2010; Armentier and Boly, 2011; Greiner et al. 2011; Kuhnen and Tymula, 2012; Charness et al. 2014). While some of the evidence qualifies more directly testing gift exchange is mixed (see Dickinson, *forthcoming*) the validity of gift exchange as a product of reciprocity is still the prevailing wisdom in most instances.

Consistent with the hypothesis of intrinsic motivation, these findings suggest that individuals derive some utility from exerting effort. Intrinsic motivation includes self-esteem, interest and pride in one's work, an innate sense of duty to honor contractual obligations (Baron, 1988; Kreps, 1997; James, 2005; Ellingsen and Johanesson, 2008), or a sense of fulfillment (Deci, 1975; Kuhnen and Tymula, 2012). Our theoretical model attempts to include these two dimensions, namely intrinsic motivation and reciprocity.[15]

Let us consider the worker's payoff function in the gift-exchange game with moral concerns:

$$U_i(e_i, w_{ij}) = w_{ij} - c(e_i) - FC - v_i(e_i - \widehat{e_i}) \tag{19}$$

Here, w_{ij} is the wage employer *j* offers worker *i*, $c(e_i)$ is worker *i*'s cost of effort function (where $c' > 0$ and $c'' < 0$) and *FC* are fixed costs (if desired). To keep matters simple, we can specify the cost function by considering a simple disutility function: $c(e_i) = e_i^2$. In equation (7) $v_i(e_i - \widehat{e_i})$ is one's "moral obligation" function that generates disuility when effort differs from one's personal moral ideal, $\hat{e}_i$ (e.g., Nyborg, 2000; Brekke et al, 2003; Figueres et al, 2013). According to (2) above, this moral obligation is a function of both a Kantian imperative,

[15]See also Fehr et al (1997) on the benefits of reciprocity in markets to increase gains from trade.

K_i, and a fairness component that depends on the wage received by the employer, $\hat{e}_i = \hat{e}_i\left(K_i, F_i(w_{ij})\right)$. This fairness component $F_i(w_{ij})$ can be interpreted as worker *i*'s perception regarding employer *j*'s fairness, in the spirit of Rabin (1993), where a high wage is perceived as an act of kindness, such that $\frac{\partial \hat{e}}{\partial w} > 0$.

Let us now consider the firm's simplified profit function with moral motivation following Gächter and Falk (2002):

$$\pi(w, e) = (Q - w)e - v(w - \widehat{w}) \tag{20}$$

where $\widehat{w}$ is the moral ideal for the wage offered to worker *i*. Q represents an exogenously given value of the worker's marginal product to the firm. In Gächter and Falk (2002) the firm's redemption value from each unit of worker effort was Q=120.

We proceed to solve the game predictions by backwards induction. In stage 2, worker *i* chooses effort level e_i to maximize:

$$max_{e_i}\ w_{ij} - c(e_i) - FC - v_i(e_i - \widehat{e_i}), \text{with } \hat{e}_i = \hat{e}_i(K_i, w_{ij}) \tag{21}$$

The first order condition for worker *i* is:

$$\frac{\partial U}{\partial e_i}: -c_e'(e_i) - v_e'(e_i - \widehat{e_i}) = 0 \tag{22}$$

In (22), the first term is negative and the sign of the second term depends, by assumption, on whether one's effort is above or below her moral obligation. Starting from a situation where worker *i* exerts an effort lower than her moral obligation, a marginal increase in effort reduces her loss of utility. This FOC can be solved to obtain Nash equilibrium effort level $e_i^* = e^*(w_{ij})$, where wage influences through the $F_i(w_{ij})$ function, such that the following identity holds when substituting optimal effort and the moral obligation function back into (22):

$$-c'_e\left(e^*(w_{ij})\right) - v_e{}'\left(e^*(w_{ij}) - \hat{e}_i(K_i, w_{ij})\right) \equiv 0 \tag{23a}$$

By differentiating both sides of this identity we get:

$$(-c''_{ee} - v''_{ee})\,\frac{\partial e^*}{\partial w_{ij}} - v''_{e\hat{e}}\,\frac{\partial \hat{e}}{\partial w_{ij}} = 0 \tag{23b}$$

From equation (23b) we can then get the following comparative static result:

$$\frac{\partial e_i^*}{\partial w_{ij}} = \frac{v''_{e\hat{e}}\left(\frac{\partial \hat{e}}{\partial w}\right)}{-c''_{ee} - v''_{ee}} > 0 \tag{24}$$

Since both $c(\cdot)$ and $v(\cdot)$ are convex functions, the denominator is unambiguously negative, and $\frac{\partial \hat{e}}{\partial w}$ is positive by assumption. Therefore, this implies that the necessary condition for the existence of a positive wage effort reciprocity is that $v''_{e\hat{e}} < 0$. This condition is true by assumption, but recall that the interpretation of this condition is that a marginal increase in the moral obligation (resulting from increased wage by employer) raises the marginal gain to increased work effort on the part of the worker in term of a marginal reduction in moral disutility.

Numeric Example

To give a numeric application of this result, assume that the moral motivation is captured by the quadratic function $v_i(e_i - \hat{e_\iota}) = (e_i - \hat{e_\iota})^2$ and cost of effort is $(e_i) = e_i^2$. First order conditions for the worker are: $-2e_i - (2e_i - 2\hat{e_\iota}) = 0$, which leads to optimal worker effort

$$e_i^* = \frac{\hat{e_\iota}}{2} \tag{25}$$

Let the specific form of the moral obligation function be a linear weighted function of her Kantian effort imperative, K_i^e, and the fairness the worker interprets from the employer's wage offer, $F_i(w)$. Thus, we have: $\hat{e}_i = \theta K_i^e + (1-\theta)F_i(w)$. Assume an equal weight on each component of the moral obligation, $\theta = 0.5$, and assume the Kantian effort imperative $K_i^e \in [0.1,1]$ is $K_i^e = 0.5$ we get $\hat{e}_i = 0.25 + 0.5F_i(w)$.

Using equation (4a), we have the fairness component of the worker's moral disutility as:

$$F_i(w) = \frac{(w-20)}{120-20}(1-0.1)+0.1 \text{ with } w \in [20,120] \tag{26}$$

We can then rewrite the moral effort obligation as $\hat{e}_i = 0.25 + 0.5\big((\mathrm{w} - 20)0.009 + 0.1\big) = 0.201 + 0.0045w$. This implies optimal effort as a function of the wage is given by:

$$e_i^*(w) = \frac{0.201+0.0045w}{2} \tag{27}$$

In stage 2 of the game, firms choose wages by maximizing utility that depend on output per effort unit, *Q*, wages paid and effort levels (recognizing these are a function of wages). As with worker effort, we introduce moral considerations in terms of the wage offered compared to a moral ideal for the employer. The employer maximizes the following:

$$\text{Maximize}_{\,w} \quad \pi(w,e) = (Q-w)(e(w)) - v(w-\widehat{w}) \tag{28}$$

Assume that the moral motivation is captured by quadratic function such that $v_i(w-\widehat{w}) = (w-\widehat{w})^2$. The first-order condition can be written as (using *Q*=120 from Gächter and Falk, 2002):

$$\begin{aligned} \frac{\partial U}{\partial w}: \quad & (Q-w)e'(w) - e(w) - v'(w-\widehat{w}) = 0 \\ -\left(\frac{0.201+0.0045w}{2}\right) & + (Q-w)e'(w) - v'(\widehat{w}-w) = 0 \\ -\left(\tfrac{0.201+0.0045w}{2}\right) & + (120-w)e'(w) - 2w + 2\widehat{w} = 0 \end{aligned} \tag{29}$$

Consider that the principal has also moral motivation such that $\widehat{w} \in [20{,}120]$ is set at $\widehat{w} = 70$. From this first-order condition we can obtain optimal wage that depends on the differential moral standards such that $w^* = w^*(\widehat{w}) = 69.92$. Note that, as in the Dictator game predictions, the optional action is a bit less than one's moral target due to the monetary cost of behaving morally. By replacing *w* by its optimal value in $e_i^*(w)$ we get: $e_i^*(w) = 0.26$, which is above the minimal effort the worker could put forth. In other words, the predicted outcome is the gift exchange effect.

Altogether our model indicates that both intrinsic moral motivation coupled with reciprocity may explain why workers outperform under a flat wage scheme and why employers are willing to pay high wages.

3.2.3. The Voluntary Contribution Mechanism

A large and active literature in experimental economics has investigated the behavior of individuals who face social dilemmas such as the prisoner's dilemma, the common pool or the public good, and the factors that increase the extent of group-oriented behavior in such situations. One commonly used context, in which the conflict between individual and group incentives is studied, is the voluntary contributions mechanism. In this game, each individual member of a group receives an initial endowment of money. Each individual then has an opportunity to contribute any fraction of his endowment to a "group account". The allocation decisions are simultaneous in that others' choices are unknown at the time an individual makes his own decision. The total amount of money that all agents contribute to the group account is multiplied by a factor greater than 1 and then divided equally among all of the members of the group. Each individual has a dominant strategy to allocate zero to the group account, whereas the highest total group payoff is reached if all members contribute their entire endowment to the group account. The level of contribution can be interpreted as a measure of the extent that decisions are socially oriented. The value of the measure can be compared between treatments to identify factors that influence the level of cooperation.

Several experimental studies have documented strong empirical regularities in public goods experiments including (1) the fact that individuals do not act purely out of self-interest, nor do they act exclusively in the group interest and contribute more than predicted by the standard theoretical model; and (2) that average contribution declines steadily over time when the game is repeated under a finite horizon (See Ledyard, 1995; Croson 1996, Keser and van Winden 2000, Fehr and Gächter 2000, Masclet et al. 2003, Carpenter 2007, Sefton et al. 2007). A number of factors, properties of both the environment and the rules of interaction, which encourage cooperation, have been identified (see Ledyard, 1995, for a survey). In this current section we attempt to show that our model is compatible existing empirical regularities.

Consider a simplified version of the Voluntary Contribution Mechanism (VCM) with two players, i and j who can contribute voluntarily to fund a public good. Each player receives an endowment w and has to decide how much to contribute, x, to a group account.

The one shot VCM

If the game is one-shot, then player i's utility function is:

$$U_i(e_i) = w - x_i + \beta(x_i + x_j) - v(x_i - \hat{x}_i) \quad (30)$$

Where $\beta \in [0,1]$ is the marginal utility from consuming the public good, *w* is a player's endowment level, and x_i is player *i*'s contribution level. Here, β<1 yields the typical free-riding prediction, $x_i = x_j = 0$ that is at odds with the efficient outcome at $x_i = x_j = w$. The FOC from (30) is

$$-1 + \beta - v'(x_i - \hat{x}_i) = 0 \quad (31)$$

Solving this leads to the Nash equilibrium contribution of player *i:*

$$x^* = x^*(\hat{x}_i(K_i)) \quad (32)$$

Here we assume the social influence function $F(x_j)$ does not play a role given the one-shot nature of the game, and so one's moral obligation stems only from her Kantian obligation, *K*. Evaluating the FOC at the optimal effort level yields the identity:

$$-1 + \beta - v'(x^*(\hat{x}) - \hat{x}) \equiv 0 \quad (33)$$

The total differential of (33) can then be written as: $v''_{x,\hat{x}} d\hat{x} + v''_{x,x} dx = 0$, which produces the implicit derivative

$$\frac{\partial x^*}{\partial \hat{x}} = \frac{-v''_{x,\hat{x}}}{v''_{x,x}} \geq 0 \quad (34)$$

The sign $\frac{\partial x^*}{\partial \hat{x}} \geq 0$ is true by the earlier assumptions that $v''_{x,\hat{x}} < 0$ and $v''_{x,x} \geq 0$. Thus, one's optimal contribution is an increasing function of one's moral obligation, not surprisingly.

The two period VCM

Let us now consider a dynamic game consisting of two periods. To answer this question, we now turn to the dynamics of contributions over time. Assume now that the public good game is played for two periods. We assume that in each period players rely on their current updated moral motivation, which is determined by the observed contribution of player *j* of the previous period.

To solve this game, we use backward induction. In period 2, player *i*'s utility function with conditional moral motivation is (using numeric notation to identify the period and simplifying own-contribution notation to $x_{i,t} = x_t$):

$$max_{x_2} \quad U_2 = w - x_2 + \beta(x_2 + x_{j,2}) - v_i(x_2 - \hat{x}_2) \quad (35)$$

Recalling that *i's* moral obligation in period 2 is a function of the Kantian ideal *K* and player *j's* round 1 contribution through the fairness function, $\hat{x}_2 = \hat{x}_2(K, x_{j,1})$, we can write the FOC:

$$-1 + \beta - v'_{x,2}\big(x_2 - \hat{x}_2(K, x_{j,1})\big) = 0 \tag{36}$$

This FOC is solved to obtain the following period 2 contribution of player *i* as a function of player *j*'s period 1 contribution*:*

$$x_2^* = x_2^*\big(x_{j,1}\big) \tag{37}$$

By replacing (37) in equation (36) we get the following identity:

$$-1 + \beta - v'_{x,2}\big(x_2^* - \hat{x}_2(K, x_{j,1})\big) \equiv 0 \tag{38}$$

We can then establish the comparative static result regarding the effect of player *j*'s period 1 contribution, $x_{j,1}$, on player *i*'s optimal period 2 contribution, x_2^*:

$$-v''_{x_2x_2}(\cdot)\frac{\partial x_2^*(\cdot)}{\partial x_{j,1}} - v''_{x_2\hat{x}_2}(\cdot)\frac{\partial \hat{x}_2}{\partial x_{j,1}} = 0 \tag{39a}$$

$$\frac{\partial x_2^*(\cdot)}{\partial x_{j,1}} = \frac{v''_{x_2\hat{x}_2}(\cdot)\frac{\partial \hat{x}_2}{\partial x_{j,1}}}{-v''_{x_2x_2}(\cdot)} > 0 \tag{39b}$$

Proof:

By assumption $v''_{x_2\hat{x}_2} < 0$ and $v''_{x_2x_2}(\cdot) > 0$, and so the sign depends on $\frac{\partial \hat{x}_2}{\partial x_{j,1}}$. Assuming one's moral obligation increases in others' previous contribution is necessary to generate conditional cooperation. From equation (4a) we know that one's moral obligation increases in the other individual's previous contribution, and therefore the sign of equation (39b) is always positive.

Once the optimal stage 2 contribution is found, we then note that first period contributions are chosen to maximize

$$max_{x_1} \quad U_1 = w - x_1 + \beta\big(x_1 + x_{j,1}\big) - v_i(x_1 - \hat{x}_1) \tag{40}$$

The FOC for this are:

$$-1 + \beta - v'_x(x_1 - \hat{x}_1) = 0 \tag{41a}$$

From this we will again have:

$$x_1^* = x_1^*(\hat{x}_1(K)) \tag{42b}$$

Numeric Example

Let β=0.8 and *w*=20, and also assume one's Kantian moral imperative is full contribution, *K*=20. Suppose player *i's* moral motivation is captured by the quadratic function $v_i(x_i - \widehat{x_\iota}) = (x_i - \widehat{x_\iota})^2$. In a one shot VCM game the FOC can be written as:

$$-1 + \beta - 2x_i + 2\widehat{x_\iota} = 0 \tag{43}$$

This leads to optimal effort (recalling that $\beta = 0.8$):

$$x_i^* = \hat{x}_i - .1 \tag{44}$$

With *K*=20 and no influence from others, $\hat{x} = 20$, and the optimal contribution level of player *i* in the one-shot game is: $x_i^* = 19.9$. It is clear in this case how contributions in the one-shot game depend on the moral obligation of the player.

Now consider the repeated game where player *i* has Kantian imperative *K*=20, but player *j* has no moral motivations and will therefore freeride (i.e., $x_j^* = 0$). The FOC for player *i* in period 2 can now be written as:

$$x_2^* = \hat{x}_2\big(K, x_{j,1}\big) - .1 \tag{45}$$

Let us specify the player *i*'s moral obligation function in period two of the game, $\hat{x}_2$, be a linear weighted function of her Kantian effort imperative, K_i^e, and the contribution of player *j* from period 1, $x_{j,1}$ such that $\hat{x}_2 = \theta K + (1-\theta)x_{j,1}$. Assume an equal weight on each component of the moral obligation, $\theta = 0.5$, and since *K*=20 for player *i,* the Kantian effort imperative $K_i^e \in [0.20]$ is $K_i^e = 20$ and so we have the period two moral obligation as:

$$\hat{x}_2 = 10 + 0.5x_{j,1} \tag{46}$$

In period 1, player *i*'s moral motivation corresponds to his Kantian effort imperative without social (fairness) influence, and so player *i* will contribute the optimal one-shot level, $x_1^*(\widehat{x_\iota}) = 19.9$ we found above. Player *j*'s contribution level in period one is zero due to her absence of moral motivation. Consequently player *i*'s moral motivation in period 2 is updated via the function $F\big(a_j\big)$ from equations (4b) to yield: $\hat{x}_2 = 10$. By substituting $\hat{x}_2$ by its value in equation (40) we find player *i*'s contribution level in period 2: $x_2^* = 9.9$. Interestingly, our model explains both why people may contribute above zero (lack of moral motivation) but also

why cooperation declines overt time due to the conditional cooperation of those who have moral motivations but are also influenced by others' actions.

4. Conclusion

Ethics and moral standards are clearly different across individuals. Some care little about moral standards, while others place great importance on them and have disutility if deviating from their moral target behaviors. Still others care about morals, but are willing to let their moral standards be influenced by others. This may be the case due to overt peer pressure, for example, or by one's perceptions of moral consensus derived from more indirect observation of others' behavior. It is likely that other-regarding preferences and reciprocity loom large in interactive or strategic decision environments, while moral consensus, peer influences, and absolute ethical standards likely play a large role in the ethical domain of individual choice.

This paper describes a theoretical framework that incorporates moral considerations into one's preferences in a way that is intuitive and can capture behavior in individual or group decision domains. Our approach models utility as a function of one's own outcome as well as a function of one's moral target or standard of behavior. Importantly, this moral standard may be a Kantian imperative, or may be more derived from fairness considerations derived from others' behavior. *Homo Economicus* is explained in this model as one who places no weight on morals, while *Homo Kantis* cares only about one's moral obligation. Heterogeneity across individuals is explained by heterogeneous weights individuals may place on these two components of the general utility function. Yet further heterogeneity is predicted when one's moral obligation is allow to be either categorical, or malleable and subject to the influence of others' behavior. Given the model's ability to explain several stylized empirical and experimental results, we hope to stimulate an increased focus on how morals and ethical standards can help shed light on important behavioral tendencies.

References

Abbink K, Herrmann B. 2011. The moral costs of nastiness. *Economic Inquiry,* 49(2): 631-633.

Abbink K, Sadrieh A. 2009. The pleasure of being nasty. *Economics Letters,* 105: 306-308.

Abeler J, Becker A, & Falk A. 2014. Representative evidence on lying costs. *Journal of Public Economics*, 113: 96-104.

Abeler J, Nosenzo D, & Raymond C. 2019. Preferences for truth-telling. *Econometrica*, 87(4): 1115-1153.

Akerlof GA. 1982. Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4): 543-569.

Alger I, & Renault R. 2006. Screening ethics When honest agents care about fairness, *International Economic Review*, 47: 59–85.

Alger I, & Weibull J. 2013. Homo moralis preference evolution under complete information and asssortative matching. *Econometrica*, 81(6): 2269–2302

Allingham MG, & Sandmo A. 1972. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics,* 1(3-4): 323-338.

Andreoni J, Erard B, & Feinstein J. 1998. Tax compliance. *Journal of Economic Literature*, 36*:* 818–860.

Armentier O, & Boly A. 2011. A controlled field experiment on corruption. *European Economic Review, 55*: 1072-1082.

Arrow KJ. 1981. Optimal and voluntary income redistribution. In: Rosenfield, S. (Ed.), Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson. Cambridge University Press, Cambridge.

Arrow KJ. 1973. Some ordinalist utilitarian notes on Rawl's theory of justice. *Journal of Philosophy*, 70: 245-263.

Baron J. 1988. The employment relation as a social relation. *Journal of the Japanese and International Economy*, 2(4): 492- 525.

Battigalli P, Charness G, & Dufwenberg M. 2013. Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93(C): 227-232.

Becker GS. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76: 169-169.

Benabou R. & Tirole J. 2011. Identity, morals and taboos: beliefs as assets. *Quarterly Journal of Economics,* 126: 805–855.

Bolton GE, & Ockenfels A 2000. A theory of equity, reciprocity and competition. *American Economic Review*, 100: 166–193.

Brekke KS, Kverndokk S, & Nyborg K. 2003. An economic model of moral motivation. *Journal of Public Economics,* 87: 1967–1983.

Brüggen A, & Strobel M. 2007. Real effort versus chosen effort in experiments. *Economics Letters*, 96(2): 232-236.

Carpenter J. 2007. Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, 60: 31–51.

Case AC, & Katz LF. 1991. The company you keep: The effects of family and neighborhood on disadvantaged youths, NBER Working Paper No. 3708, Cambridge, MA.

Charness G, Masclet D, & Villeval MC. 2014. The dark side of competition for status. *Management Science,* 60(1): 38-55.

Coricelli G, Joffily M, Montmarquette C, & Villeval MC. 2010. Cheating, Emotions, and Rationality: An Experiment on Tax Evasion. *Experimental Economics* 13(2): 226-247.

Croson R. 1996. Partners and strangers revisited. *Economics Letters* 53: 25–32.

Deci EL. 1975. *Intrinsic Motivation.* New York: Plenum Publishing Corp.

Diekmann A, Przepiorka W, & Rauhut H. 2015. Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. *Rationality and Society*, *27*(3): 309–333.

Dickinson DL. *Labor Negotiations, Conflict, and Arbitration*. In eds. KF Zimmermann (Editor-in-Chief) and MC Villeval (Section Editor, Behavioral Economics) *Handbook of Labor, Human Resources and Population Economics*. Springer-Verlag, forthcoming.

Dickinson DL., 2000. Ultimatum decision-making: A test of reciprocal kindness. *Theory and Decision*, 48(2): 151-177.

Dohmen T. & Falk A. 2010. Performance pay and multi-dimensional sorting productivity, Preferences and Gender. *American Economic Review,* 101(2): 556-590.

Edgeworth FY. 1881 *Mathematical Psychics: An Essay on the Applications of Mathematics to the Moral Sciences*. L.S.E. Series of Reprints of Scarce Tracts in Economics and Political Sciences, No. 10, 1932.

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

Elster, J. 1989. Social Norms and Economic Theory. *Journal of Economic Perspectives*, 3(4): 99-117.

Engel C., 2011. Dictator games: a meta study. *Experimental Economics*, 14(4): 583-610.

Evans III JH, Hannan RL, Krishnan R, & Moser DV. 2001. Honesty in managerial reporting. *The Accounting Review*, 76(4): 537-559.

Falk A. & Ichino A. 2006. Clean evidence on peer pressure. *Journal of Labor Economics,* 24(1): 39-57.

Falk A. & Fischbacher U. 2002. Crime" in the lab-detecting social interaction. *European Economic Review.* 46(4-5): 859-869.

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

Fehr E, Kirchsteiger G, & Riedl A. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics,* 108: 437-459.

Fehr E, Kirchler E, Weichbold A, & Gächter S. 1998. When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor Economics*, 16(2): 324-351.

Fehr E. & Schmidt KM. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics,* 114(3): 817-68.

Fehr E. & Schmidt KM. 2006. The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. In *Handbook on the Economics of Giving, Reciprocity and Altruism*, ed. Serge-Christophe Kolm and Jean Mercier Ythier, 615-91. Amsterdam: Elsevier.

Fehr E, Gächter S, & Kirchsteiger G. 1997. Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica*, 65(4): 833-860.

Fehr E, Kirchler E, Weichbold A, & Gächter S. 1998. When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets. *Journal of Labor Economics*, 16(2): 324-351.

Figuieres C, Masclet D, & Willinger M. 2013. Weak moral motivation leads to the decline of voluntary contributions. *Journal of Public Economic Theory*, 15(5): 745-772.

Fischbacher U, & Föllmi-Heusi F. 2013. Lies in Disguise. An experimental study on cheating. *Journal of the European Economic Association*, 11(3): 525-547.

Forsythe R, Horowitz JL, Savin NE, & Sefton M. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3): 347-369.

Fortin B, Lacroix G. & Villeval MC. 2007. Tax evasion and social interactions. *Journal of Public Economics* 91: 2089-2112.

Gächter S. & Falk A. 2002. Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics*, 104: 1-27.
Gächter S, & Schulz JF. 2016. Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595): 496-499.
Gino F, Gu J, & Zhong CB. 2009. Contagion or restitution? When bad apples can motivate ethical behavior. *Journal of Experimental Social Psychology,* 45: 1299-1302.
Glaeser E, Sacerdote B, & Scheinkman J. 1996. Crime and social interactions. *Quarterly Journal of Economics,* 111: 507-548.
Gneezy U. 2005. Deception: The role of consequences. *American Economic Review*, 95: 384–394.
Gneezy U, & List J. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 74: 1365-1384.
Greiner B, Ockenfels A, & Werner P. 2011. Wage transparency and performance: A real effort experiment. *Economic Letters,* 111: 236-238.
Grover SL, & Hui C. 2005. How job pressures and extrinsic rewards affect lying behavior. *International Journal of Conflict Management*, 16: 287–300.
Güth W, Schmittberger R. & Schwarze B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4): 367-388.
Harsanyi J. 1980. Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and Decision*, 12: 115-133.
James H. 2005. Why did you do that? An economic examination of the effect of extrinsic compensation on intrinsic motivation and performance. *Journal of Economic Psychology*, 26(4): 549-566.
Kandel E, & Lazear E. 1992. Peer pressure and partnership. *Journal of Political Economy,* 100: 801-817.
Kant I. 1785. Groundwork of the Metaphysics of Morals.
Keizer K, Lindenberg S, & Steg L. 2008. The spreading of disorder. *Science,* 322(5908): 1681-1685.
Keser C, and Van Winden F. 2000. Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102: 23–39.
Kessler, D. and Levitt, S.D., 1999. Using sentence enhancements to distinguish between deterrence and incapacitation. *The Journal of Law and Economics*, *42*(S1), pp.343-364.
Kuhnen C. & Tymula A. 2012. Feedback, self-esteem and performance in organizations. *Management Science*, 58: 94-113.
Kreps D. 1997. Intrinsic motivation and extrinsic incentives. *American Economic Review*, 87(2): 359-364.
Kroher M, & Wolbring T. 2015. Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. *Social Science Research*, 53: 311-324.
Laffont JJ. 1975. Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, 42(168): 430-437.
Ledyard, J.O., 1995. Public goods: A survey of experimental research. In J. Kagel and A. Roth (eds), *The Handbook of Experimental Economics*. Princeton : Princeton University Press
Levitt, S.D., 1996. The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The quarterly journal of economics*, *111*(2), pp.319-351.
Ludwig J, Duncan GJ, &Hirschfield P. 2001. Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. *Quarterly Journal of Economics,* 116: 655-680.
Mas A, & Moretti E. 2009. Peers at work. *American Economic Review*, 99(1): 112-45.

Masclet D, Noussair C, Tucker S, &. Villeval MC. 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93: 366–380.

Mazar M, Amir O, & Ariely D. 2008a. More Ways to Cheat -Expanding the Scope of Dishonesty. *Journal of Marketing Research*, 45(6): 651-653.

Mazar M, Amir O, & Ariely D. 2008b. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, 45(6): 633–644.

Nagin DS, Rebitzer JB, Sanders S, & Taylor LJ. 2002. Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review,* 92: 850–873.

Nyborg K. 2000. Homo economicus and homo politicus: Interpretation and aggregation of environmental values. *Journal of Economic Behavior and Organization*, 42: 305–322.

Pruckner G, & Sausgruber R. 2013. Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association,* 11: 661–679.

Rabin M. 1993. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5): 1281–1302.

Roemer JE. 2010. Kantian Equilibrium. *Scandinavian Journal of Economics*, 112: 1–24.

Samuelson PA. 1993. Altruism as a problem involving group versus individual selection in economics and biology. *American Economic Review,* 83: 143–148.

Sefton M, Shupp Rm & Walker JM. 2007. The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45: 671–690.

Sen A. 1995. Moral codes and economic success. In: Britten, C.S., Hamlin, A. (Eds.), *Market Capitalism and Moral Values*. Edward Eldar, Aldershot.

Shalvi S, Dana J, Handgraaf MJ. & De Dreu CK. 2011. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2): 181-190.

Smith A. 1759. The Theory of Moral Sentiments. *Penguin*

Torgler B. 2002. Speaking to theorists and searching for facts: tax morale and tax compliance in experiments. *Journal of Economic Surveys*,16: 657–683.

Yitzhaki S. 1974. A Note on 'Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics* 3(2): 201-202.

Zizzo D. 2004. Inequality and procedural fairness in a money burning and stealing experiment. In: *Inequality, welfare and income distribution: Experimental approaches.* Emerald Group Publishing Limited, 215-247.

Zizzo D, & Oswald AJ. 2001. Are People Willing to Pay to Reduce Others' Incomes? *Annales d'Economie et de Statistique*, 63-64: 39-62