Narratives, Imperatives and Moral Reasoning[1]

Armin Falk[2]
and
Jean Tirole[3]

May 18, 2016

Preliminary. Comments welcome.

Abstract: This paper provides a theoretical framework to explain recent empirical findings on the malleability of morality. Building on a basic model with image concerns, we introduce the concept of narratives that allow individuals to maintain a positive image when in fact acting in a morally questionable way. We show how narratives, however feeble, inhibit moral behavior in downplaying externalities, magnifying the cost of moral behavior, or in suggesting not being pivotal. We further investigate why narratives are shared and get disseminated. We then turn to imperatives, i.e., moral rules or precepts, as a mode of communication to persuade agents to behave morally, and identify the conditions under which Kantian behavior will emerge in an otherwise fully utilitarian environment. Finally, we study how collective decision making and organizational design produces a sub-additivity of responsibility.

Keywords: Moral behavior, narratives, imperatives, Kantian reasoning, utilitarianism, organizations

JEL Codes: D62, D64, D78.

[1] This work owes much to discussions and joint work with Roland Bénabou. Exchanges with Johannes Hermle, Paul Seabright and Nora Szech and at presentations at the European University Institute, the second IAST-IEC workshop in Toulouse and the Personal Identity Conference (Chicago) as well as research assistance by Juliette Fournier, Pierre-Luc Vautrey and Thorben Woelk are also gratefully acknowledged. The European Research Council (European Community's Seventh Framework Programme (FP7/2007-2013) Grant Agreement no. 249429) provided research support.

[2] Institute of Behavior and Inequality and Department of Economics, University of Bonn.

[3] Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST), University of Toulouse Capitole.

1 Introduction

Moral behavior is malleable. This insight derives from human history as well as the recent and growing empirical literature in economics and psychology. The latter suggests that context, individual characteristics, and a williningness to engage in cognitive maneuvers justifying why behavior is not blameworthy, are important drivers of morality. This paper provides a coherent theoretical framework to offer a unified explanation of important empirical findings, and to allow analyzing individual and social circumstances that inhibit or promote immoral behavior. To do so, our basic framework models the trade-off between cost of acting morally and benefits from maintaining a positive self- or social image in agents with potential self-control problems. We introduce the role of narratives and show how they serve as an excuse for moral transgression without damaging image concerns. We then introduce a utilitarian perspective on imperatives to shed light on the long-standing dispute about two main conceptions of moral reasoning, utilitarianism and deontological ethics. Finally, we address the issue of collective decision making in its role to create a sub-additivity of responsibility, and characterize organizational design features that are crucial for shaping moral outcomes. The assumptions and predictions of our model are informed and supported by historical and empirical evidence, but they also raise new empirical questions, e.g., about the emergence of imperatives or the design of organizations.

Building on the premise that people strive to maintain a positive self-concept, and that moral behavior is central to individuals' image concerns (Aquino and Reed II, 2002; Mazar et al., 2008; Monin and Jordan, 2009), we model the trade-off between *image* and cost arising from acting morally. In the philosophical discourse, morality is typically described in terms of avoiding and preventing harm to others (Bentham, 1789; Mill, 1861; Gert and Gert, 2016). Accordingly, we define an action as moral if it produces a positive externality. Individuals are assumed to have an intrinsic valuation for the externality, but they are facing self-control problems when trading off image concerns with the cost of acting morally. In equilibrium an individual is more likely to act morally, the higher her image concerns and the perceived social benefits, and the lower the cost, her initial reputation and the level of self-control problems. These assumptions and predictions are supported by empirical evidence. Several papers have demonstrated the importance of increased social and self-image for the likelihood of pro-social behaviors. Likewise, recent work on moral cleansing and licensing suggests an important role of the initial reputation, i.e., an individual's image prior to taking a decision. We review evidence in support of an inverse relation between self-control problems and moral outcomes, and discuss experimental evidence showing that the likelihood of moral behavior increases in the level of the externality and decreases in the cost of enforcing it.

Our basic framework sets the stage for introducing *narratives*, a particularly relevant phenomenon for understanding variation in moral behavior. Narratives can be thought of as stories and descriptions we tell ourselves and others to organize experience and interpret circumstances; they are "instrument[s] of mind in the construction of reality" (Bruner, 1991, p. 6). Narratives are central to human existence and have received much attention outside economics, in particular in psychology and the humanities. Here we introduce a simple, economic notion of narratives (a signal about the externality), and study their relevance for moral outcomes. In the model – as in reality – narratives need not be true; it is sufficient that they are minimally plausible to allow reframing the decision context and to provide an excuse. This in turn permits the individual to engage in immoral action while maintaining a positive self-image. Narratives are extremely powerful in shaping behavior. We discuss a range of historical examples and experimental evidence focusing on narratives of "neutralization" (Sykes and Matza, 1957), including denial of responsibility, injury and victim. We show how narratives legitimize wrongful actions in downplaying externalities, reducing perceptions of being pivotal, magnifying costs and in encouraging omission bias as well as action and information avoidance. Clearly, if a decision maker perceives a victim as undeserving, i.e., the perceived externality is low, hurting him is not terribly damaging to his self-image. But narratives also have a social dimension. Suppose a decision maker who cares for his social image learns such a narrative and, while acting immorally, is observed by his peers: what would he do? He would want to *share* the narrative with his peers to have them evaluate him in a more positive way. We show that with image concerns, individuals prefer disclosing a narrative, even if this involves a cost. As a consequence, peers will find it easier to act immorally as well, generating the motive to spread the narrative even further. This creates an ever increasing demand and supply of narratives and provides an explanation for the endogenous dissemination of narratives and ideologies that goes along with an increased propensity to act immorally. By contrast, positive narratives will not go viral, unless individuals want to develop reputations for humbleness.

A major controversy in occidental moral thinking concerns the foundation of moral behavior in terms of either utilitatrian/consequentialist or deontological/rule-based reasoning. Classic utilitarians justify normative principles and actions in terms of their consequences, aiming at maximizing the Good, i.e., pleasure, happiness, desire satisfaction, or welfare (Alexander and Moore, 2015; Bentham, 1789; Johnson, 2014; Mill, 2002/1861; Sinnott-Armstrong, 2015). A major criticism of consequentialism is that it allows or even encourages harmful acts (such as killing or beating innocents), as long as these acts generate net positive consequences, e.g., saving lifes. In contrast, deontological approaches categorically prohibit harmful actions: Wrongful acts cannot be justified by their consequences, regardless of how beneficial they are. What constitutes an act as moral is acting in accordance with moral imperatives or rules, or put differently, deontological reasoning attributes a priority of the Right over the Good (Alexander and Moore, 2015). The tension between these two lines of moral reasoning has been demonstrated in various moral dilemma thought experiments, e.g., the well-known Trolley problem (Foot, 1967). Survey studies on these dilemmas typically document a co-existence of consequentialist and deontological reasoning.

In this paper we offer a formal account for such a co-existence. We show how *imperatives* can emerge in an otherwise fully utilitarian framework, and why acting in a Kantian way can be popular. The framework is a communication game between a principal (e.g., parents, or a religious leader) and an agent. The principal cares for the welfare of society and/or the agent, and can issue either a positive narrative or an imperative. The narrative concerns parameters such as highlighting the level of the externality, or the cost or visibility of behavior. In contrast, an imperative is a moral precept to act in a certain way, i.e., it does not focus on motives but dictates a particular action. Both instruments are potential means of persuasion to encourage the agent to act in a moral way. The analysis reveals that imperatives can emerge in a world populated with utilitarians. We identify the conditions under which Kantian behavior is likely, and discuss the costs and benefits of using imperatives: Conditions facilitating the emergence of imperatives are high levels of congruence of interests between principal and agent, large externalities and principals with moral authority, i.e., sound moral judgment. What matters also is an agent's self-control problems, but the effect is not monotonous.

These results reveal various costs and benefits of imperatives, relative to narratives. On the cost side, imperatives are effective only if issued by principals with moral authority while narratives can be issued by anybody. This suggests why imperatives are rarely used in the political arena, but are more common in parent-child interactions or in religious writings, such as the Ten Commandments. Another restrictive feature of imperatives is that they imply some rigidity in decision-making. They dictate a particular action and leave little room for adapting to contingencies, a feature reminiscent of the rule vs. discretion idea. In the philosophical debate this rigidity has often been identified as an important weakness of deontological reasoning. The imperative not to lie, put forward and defended in Kant (1785)[4] is easily at odds with moral intuition, as in Kant's famous example on lying to a murderer at the door in order to save the life of a friend. On the benefit side, imperatives are less vulnerable to interpretation uncertainty and to the threat of being debunked. Moreover, imperatives, when effective, expand the range of desired behaviors, i.e., encourage the provision of greater externalities. This offers a possible explanation why imperatives are typically referring to rather general prescriptions and duties with a large scope, such as not to lie, to steal or to kill. Finally, we show that under certain conditions, issuing imperatives rather than narratives is more likely if agents suffer from self-control problems. This offers a psychological explanation for imperatives, akin to the use of heuristics (Sunstein, 2005). Rules or heuristics help individuals who are vulnerable to taking impulsive decisions, in reaching their long-term goals. While, due to their rigidity, rules can misfire under certain circumstances, they generally do well in guiding behavior in the presence of lack of self-control.

We then argue that Kantian postures generate favorable social images, and therefore suggest why they turn out to be much more common than standard economics would predict. First, and following up on Bénabou-Tirole (2009), we show that the very act of questioning an imperative may have a deleterious image impact even when one ends up abiding by the imperative. Calculating individuals are perceived as uncommitted, as they are evaluated not only through their acts but also through their blind abidance by the imperative. Second, we formalize the domain of the incommensurable dear to Durkheim (1915) and Kant (1795) and the general reluctance of individuals to consider trade-offs in moral choices. To account for empirical evidence on the elicitation of preferences in the moral domain, we look at the strategy method, whereby the individual is asked to provide a response to a range of possible offers, among which one is then drawn. We provide conditions under which the individual refuses to take any offer, however large. The link with the first observation is that the attitude toward an offer impacts the individual's image beyond the occurrence of this particular offer; the individual must therefore trade off the benefit from taking very large offers with the collateral damage incurred overall on her image. Taken together, these observations fit well with the widespread propensity to resolve moral dilemma through indignation and rigid postures rather than through analysis.

[4] "To be *truthful* (honest) in all declarations is therefore a sacred command of reason prescribing unconditionally, one not to be restricted by any conveniences." (Kant 1797, 8: 427)

Many morally relevant decisions are taken in groups, rather than in isolation. In the final part of our analysis we therefore study moral implications of *shared control*, i.e., contexts where implementation of morally wanted outcomes depend on several agents' actions, as is characteristic for firms, bureaucracies and organizations in general. We show that shared control bears the potential to encourage immoral behavior but that the precise circumstances matter. In terms of production technology we differentiate between individual and collective veto power, i.e., contexts where either each agent or the collective is pivotal in guaranteeing the moral outcome. Two further distinctions concern whether the cost of acting pro-socially is associated to individual action (individual incentives) or collective outcome (team incentives), and whether reputation can be attributed to each agent's action separately (individual accountability) or happens at the level of the group (collective accountability). Perhaps surprisingly, we find that shared control with team incentives and individual accountability enhances scope for moral outcomes. The intuition is what we call a "cheap signalling effect": if the cost of acting morally may not materialize anyway, why not choose to act pro-socially and reap the image benefits? This result offers an explanation for why decisions of committees which publicly disclose individual voting outcomes may be soft on third parties. With individual incentives, however, taking decisions in groups may evoke a sub-additivity of responsibility. In comparison to decisions taken in isolation, the existence of moral equilibria is less likely in groups, regardless of whether production involves individual or collective veto power. Likewise, collective accountability may reduce incentives to act pro-socially. Intuitively, in cases where the audience cannot attribute outcomes to individual actions, reputation is blurred and agents may "hide" behind the group outcome. The analysis suggests several implications for organizations seeking to enhance corporate social responsibility.

The remainder of the paper is organized as follows. In the next section we introduce the basic framework and discuss implications and assumptions of the model in the light of empiricial findings. Section 3 introduces the notion of narratives, and conditions for sharing and debunking narratives. The emergence of imperatives and Kantian behavior is modeled in Section 4. Section 5 investigates the popularity of Kantians. In Section 6 we turn to organizational design and its potential impact on sub-additivity of responsibility. Section 7 concludes.

2 Basic moral trade-off

2.1 Modeling moral behavior

The theoretical model builds on Bénabou and Tirole (2006, 2011, 2012). There are three periods, $t = 0, 1, 2$. Date 0 will later refer to the "ex-ante self". At date 1, the individual will choose whether to engage in moral (or pro-social) behavior ($a = 1$) or not ($a = 0$). A moral choice yields a self- or social image benefit at date 2. The individual has deep value or altruism degree v (high, moral type) or 0 (low, immoral type), with probabilities ρ and $1 - \rho$. Let $\bar{v} = \rho v$ denote the expected value.

Suppose that the low-value type never engages in moral behavior (as will be the case under Assumption 1 below).[5] The issue then is whether the v-type chooses $a = 1$.

Let c denote the private cost of pro-social behavior, $\beta < 1$ the self-control or hyperbolicity parameter and μ the strength of the image concern. We will assume that the true value of the externality is $\omega = 0$ or $\omega = 1$, and we let $e = E[\omega]$ denote the expected externality, that is, e is the ex-ante probability that the externality is 1. The high type therefore has an intrinsic motivation for the moral action equal to ve. We assume that $v < 1$: the individual cannot value the externality more than the beneficiaries of the externality. Note that the preferences are explicitly consequentialist/utilitarian: The individual has low intrinsic motivation if he perceives the externality to be minor. This modeling choice will make the emergence of Kantian behaviors all the more interesting (see section 4).

In the specific context of self-signaling, the individual has some feel for his values at the decision date, but forgets that later on, to only remember the decision $a = 0$ or 1. Alternatively, image concerns might refer to social signaling: only the individual knows his true type, but the self's intended audience (e.g., his peers) do not. Either way, the moral type's objective function is

$$\left(ve - \frac{c}{\beta}\right) a + \mu\widehat{v}(a),$$

where $\widehat{v}(a)$ is the expected type conditional on action $a \in \{0, 1\}$.

Similarly the immoral type's objective function is

$$\left(-\frac{c}{\beta}\right) a + \mu\widehat{v}(a).$$

When the immoral individual has a dominant strategy not to contribute, as will be the case under Assumption 1 below, there is a multiple-equilibrium range (this is not the case when *a contrario* the high-value type has a dominant strategy to contribute). Intuitively, if the high type does not contribute, there is less stigma from not contributing, and so the high type in turn is less eager to contribute. In case of multiple equilibria, we choose the equilibrium that is best for both types, i.e., the no-contribution equilibrium (this selection is rather irrelevant for the conclusions as we will shortly note). In an equilibrium in which the high type does not contribute, the reputational gain from $a = 1$ is $\mu(v - \bar{v})$.[6]

We make an assumption that will guarantee that the high type always contributes when certain of the existence of an externality, while the low type never contributes:

Assumption 1.

$$v - \left(\frac{c}{\beta}\right) + \mu\left(v - \bar{v}\right) > 0 > -\left(\frac{c}{\beta}\right) + \mu v.$$

The right inequality in Assumption 1 simply says that the immoral type does not want to contribute even when the reputational benefit of contributing is maximal at

[5] We could have made the reverse assumption (type v always engages in pro-social behavior) without affecting the key results.

[6] We rule out strictly dominated strategies when updating for out-of-equilibrium behaviors.

$\mu(v-0) = \mu v$. Thus, it is a dominant strategy for the low type not to contribute. Consequently, in a pooling equilibrium, in which both types choose $a = 0$, we will assume that $\hat{v}(1) = v$, as only type v can gain from contributing.[7]

The left inequality in Assumption 1 says that the complete absence of contribution (pooling at $a = 0$) is not an equilibrium whenever the externality is certain, as the high type would then want to contribute (and reap image gain $\mu(v - \bar{v})$).

If $ve - (c/\beta) + \mu(v - \bar{v}) \leq 0 \leq ve - (c/\beta) + \mu v$, there exist both a pooling equilibrium at $a = 0$ and a "separating equilibrium" in which the high type contributes. But the separating equilibrium yields a lower payoff to both types[8] and is therefore ruled out by our selection criterion. The separating equilibrium selection yields the same comparative statics, except that now there is no influence of the initial reputation.[9]

Assumption 1 ensures that there exists e^* in $(0, 1)$ satisfying

$$ve^* - \left(\frac{c}{\beta}\right) + \mu(v - \bar{v}) = 0. \tag{1}$$

Then in equilibrium, the high type contributes if and only if:

$$e > e^*.$$

When $e \leq e^*$, the "unethical or immoral equilibrium", in which $a = 0$ is always selected, prevails. When $e > e^*$, the "ethical or moral equilibrium", in which type v chooses $a = 1$, obtains (and is actually the unique equilibrium).

Proposition 1 *The moral type contributes if and only if*

$$ve - \left(\frac{c}{\beta}\right) + \mu(v - \bar{v}) > 0.$$

Immoral behavior is encouraged by a low image concern (low μ), a good initial reputation (high $\bar{v}$), a low self-control (low β), a high cost of moral behavior (high c), and a low perceived social benefit from moral behavior (low e).

In the following we discuss the plausibility of our modeling assumptions in light of empirical evidence.

[7] This is for instance implied by refinement D1, but this refinement is stronger than needed here.

[8] For the low type, $\mu \cdot 0 < \mu\bar{v}$, and for the high type $ve - (c/\beta) + \mu v < \mu\bar{v}$.

[9] By contrast, one obtains a role for a certain type of excuse, the inability to pick $a = 1$, which does not arise in the selected equilibrium. Suppose that with probability x, only $\{a = 0\}$ is doable, and that the audience will not know whether the choice set was $\{a = 0\}$ or $\{a \in \{0, 1\}\}$. The condition for moral behavior is still given by (1) under the selected equilibrium, but becomes

$$ve - \left(\frac{c}{\beta}\right) + \mu(v - \tilde{v}) > 0$$

under the separating-equilibrium choice, where $\tilde{v} \equiv x\rho v/[1 - (1 - x)\rho] \in (0, \bar{v})$ if $x \in (0, 1)$: the better the excuse (the higher x is), the less moral the behavior.

2.2 Supportive evidence

Social and self-image concerns (μ) play a central role for our model, as well as for a whole class of related signaling models (e.g., Bénabou-Tirole, 2006, 2011, 2012). Social image refers to reputational concerns vis-à-vis others, e.g., in terms of a desire to be liked and well regarded by peers. Higher levels of social image concerns are predicted to positively affect moral behaviors. An example for supporting evidence is provided by Ariely et al. (2009). In their "Click for Charity" experiment subjects can donate to a charitable organization by repeatedly clicking two keys on a computer keyboard. They find that participants provide significantly greater effort in the presence of an audience than in private.

Self-image concerns, in contrast, are purely self-directed and refer to the self-awareness of congruency between internal standards or values and the self (e.g., in light of current behavior). Thus, higher levels of self-awareness (μ) should favor behaviors that are compatible with salient (moral) standards compared to states where attention is directed away from the self.[10] Consistent with this notion, experimental evidence in psychology suggests that self-awareness fosters fairness and honesty if moral standards are salient (Batson et al., 1999), decreases transgression rates for children (Beaman et al., 1979) and inhibits cheating in a performance test if moral standards (in contrast with competence standards) are salient (Diener and Wallbom, 1976; Vallacher and Solodky, 1979). To provide a more direct test for the relevance of self-image (μ) in the context of our model, and signaling models in general, Falk and Tirole (2016) have run a simple binary choice experiment. In the experiment participants can choose to exert a negative externality in return for a monetary reward ($a = 0$), or not to exert the negative externality ($a = 1$). In the main condition μ is exogenously increased by exposing participants to their "self-image" through videotaping them in real time. Relative to control treatments where participants do not see their face, the likelihood of exerting the negative externality is significantly lower.

The model predicts a higher likelihood of unethical behavior in the presence of a *high initial reputation* (high $\bar{v}$). In other words, present behavior is affected by previously aspired identity, potentially giving rise to "moral cleansing" as well as "moral licensing" effects (Merritt et al., 2010). The former describes enhanced moral activity in situations where moral self-worth has been threatened (Zhong and Liljenquist, 2006). In contrast, the latter suggests that salience of previous good behavior discourages moral action. There is ample evidence on moral licensing in the domains of political incorrectness (Bradley-Geist et al., 2010; Effron et al. 2009; Merritt et al., 2012; Monin and Miller, 2001), selfishness (Jordan et al., 2011; Khan and Dhar, 2006; Mazar and Zhong, 2010; Sachdeva et al., 2009), dieting (Effron et al., 2012) and consumption choices (Khan and Dhar, 2006). Monin and Miller (2001), e.g., study a hypothetical hiring decision: Subjects who had been able to demonstrate non-prejudiced attitudes were subsequently more likely to express the belief that a (police) job is better suited for a White than a Black person. Similarly, Effron et al. (2009) show that after being given the opportunity to endorse Obama, participants were more likely to favor Whites than Blacks. Sachdeva et al. (2009) show licensing effects in the realm of altruism. They find that after being asked to write

[10] For more details, see literature on Objective Self-Awareness Theory (Duval and Wicklund, 1972).

a self-relevant story including positive (negative) traits, participants donate less (more) to a charity.

Our model predicts a positive correlation between moral behavior and *self-control* (β). Intuitively, morally demanding decisions often imply the trade-off between immediate gratification (e.g., money), and costs accruing in the future, e.g., in terms of lower self-image or reputation. When facing such trade-offs, individuals lacking self-control (low β) will be more tempted to engage in immoral behaviors. This implication of the model is supported by numerous empirical studies. Achtziger et al. (2015), e.g., show that participants who are ego depleted share significantly less money in a dictator game compared to non-depleted participants. Ego depletion is a manipulation, which consumes self-control resources and favors impulsive behavior (Baumeister et al., 2007).[11] Related evidence shows that lack of self-control fosters socially inappropriate actions such as dishonesty (Gino et al., 2011; Mead et al., 2009), criminal behavior (Gottfredson and Hirschi, 1990), and undermines cooperation (Osgood and Muraven, 2015). Neuroscientific evidence further suggests that an inhibition of self-control (dorsolateral prefrontal cortex) through transcranial magnetic stimulation induces more selfish behavior (Knoch et al., 2006). Burks et al. (2009) measure β and the standard discount factor δ in price list experiments in a sample of truck drivers. They report positive correlations with both β and δ, and the level of individual cooperation in simple spell out games. Similarly, in a preschool children sample (Bartling et al., 2012), β is significantly positively correlated with prosocial behavior[12] in every day life. Using self-reported measures of self-control (Rosenbaum Self-Control Schedule), Martinsson et al. (2012) find that lack of self-control is associated with lower offers in dictator games.[13]

Many papers show that moral or prosocial behavior is sensitive to the *cost* of moral behavior (high c). In public goods games, e.g., the cost of providing a positive externality is inversely related to the level of cooperation (Goeree et al., 2002; Gächter and Herrmann, 2009). Likewise, the willingness to exert altruistic punishment in public goods games with a subsequent sanctioning stage decreases in the cost of punishment (Egas and Riedl, 2008, Nikiforakis and Normann, 2008). A recent meta-analysis of dictator game behavior also shows that stake size matters. Using data on 158 treatments that explicitly varied stake size, Engel (2011) finds that higher stakes reduce the willingness to give, i.e., if more is at stake dictators keep more for themselves, both in absolute and relative terms. A final example that uses a binary moral decision, as modeled here, is provided in Falk and Szech (2013, 2014, 2015). In their experiment subjects face the morally relevant decision to either kill a mouse in return for money, or to save the life of a mouse and to receive no money. In several treatments subjects were facing a price list with increasing monetary amounts offered for the killing option, and no money for saving the mice. In other words, for higher prices offered the cost c of acting morally is increasing. In line with the model

[11] In this sense, Baumeister and Exline (1999) have actually coined self-control as the “moral muscle”.

[12] This is measured using the prosocial sub-facet of the Strength and Difficulty Questionnaire (own calculations. We thank Fabian Kosse).

[13] Note that the relation between self-control and moral behavior is ultimately rooted in what constitutes prepotent responses or fundamental impulses: selfishness or pro-sociality. The view expressed here is that selfish impulses need to be controlled, akin to the “Freudian” notion that the (super-)ego needs to override the fundamentally selfish human nature. The scientific dispute about fundamental impulses, however, is not resolved. Rand et al. (2012), e.g., argue that people are predisposed to acting pro-socially, and become self-interested only after reflection and reasoning.

assumption, the likelihood of acting morally strongly decreases in c.

The fact that the level of the *externality* or social benefit (e) affects prosocial decision making is well documented, in particular in the literature on cooperation and voluntary contribution to public goods (Kagel and Roth, 1995, Chapter 2). A particularly clean study that disentangles higher external return (gain for each other person) and internal costs (cost for subject) is Goeree et al. (2002). They find that a higher external return is associated with higher contributions. A nice field application of the relevance of perceived social benefits is Gneezy et al. (2014). They show that donations to charity decrease when overhead increases. Moreover, informing potential donors that overhead costs are covered, significantly increases donation rates.

2.3 Demand for moral opportunities

Comparing the objectives of the ex-ante self and the ex-post one, we can point at two wedges:

a) The *self-control problem* acts against moral behavior in a way that may make it desirable to tilt the balance toward moral behavior. For instance, one would like to refrain from retaliating against or yelling at friends and relatives, but be concerned about impulsive behavior in this respect. Image investments can therefore be viewed as just any other form of investment, justifying standard strategies to thwart impulsive behavior.[14]

b) *Image concerns* behind the veil of ignorance create a "zero-sum game" or generate investment in a "positional good". They promote pro-social behavior, but perhaps too much from the point of view of the ex-ante self, who would then want to avoid situations in which he will feel compelled to act prosocially.

Let us ask ourselves: would the individual, behind the veil of ignorance (not knowing her type), want to face a restricted choice $\{a = 0\}$?

The ex-ante self has utility $E[(ve - c)a + \mu\hat{v}(a)]$. Oversignaling for instance occurs if condition (1) holds: $ve - c/\beta + \mu(v - \bar{v}) \geq 0$, but $\rho(ve - c + \mu v) + (1 - \rho) \cdot 0 < \mu\bar{v}$, or equivalently $ve - c < 0$.

Proposition 2 *The ex-ante self feels that his ex-post incarnation will oversignal if* $e > e^*$ *and* $ve - c < 0$. *Oversignaling may therefore occur when:*

$$\left(\frac{c}{\beta}\right) - c < \mu(v - \bar{v}). \tag{2}$$

There is undersignaling if $e \leq e^*$ *and* $ve - c > 0$, *which may therefore occur when (2) is satisfied with the reverse inequality.*

Remark: We have assumed that the warm glow term is not subject to hyperbolic discounting (or at least lingers longer than the perceived cost). We could have made the opposite assumption, in which case moral behavior would require $(ve - c)/\beta + \mu(v - \bar{v}) > 0$. There would then always be oversignaling.

[14] For example, the use of personal rules (see Bénabou-Tirole 2004).

3 Narratives

This section introduces the role of narratives, and their implications for moral behavior. Narratives are stories we tell ourselves or others to make sense of reality (Bruner, 1991). As is widely accepted, narratives are central to human existence, and have consequently been studied in various fields including psychology, history, sociology, criminology and the humanities. In modern personality psychology narratives are considered as integral part of human personality. In McAdams' three-tiered personality framework, e.g., personality is modeled as an interplay between dispositional traits, characteristic adaptations and life stories, where the latter provide life with a sense of meaning, unity and purpose (McAdams 1985, 2006).

Here, we seek to introduce the notion of narratives into economics with a particular focus on their specific role in legitimating harmful action.[15] In short, when facing morally demanding choices, narratives provide an interpretative framework allowing favorable representations of action and consequences as complying to moral standards, when in fact violating these standards. In particular, stories of "neutralization" (Sykes and Matza, 1957), i.e., denial of responsibility, injury or victim, condemnation of the condemned and appeal to higher loyalties allow actors to convince themselves that their actions are not violating norms or standards they are otherwise committed to. In this sense, narratives can be understood as a form of "moral lubricant" (Presser and Sandberg, 2015 p.288). Narratives originate both from the individual and cultural level. They can be understood as "psychosocial constructions, coauthored by the person himself or herself and the cultural context within which the person's life is embedded and given meaning" (McAdams, 2001 p.101). Importantly, these stories need not be true in an objective sense to nevertheless powerfully influence a person's behavior and judgment (see, e.g., Haidt et al., 2009).

Our account of narratives is deliberately simple and not meant to offer a comprehensive account of the phenomenon as such. Instead we provide a version parsimonious enough to be included in a model of self-image concerns and moral behavior. It captures the essential feature that narratives may not be true, but contain some grain of truth or subjective plausibility to credibly reframe the decision allowing decision makers to maintain the self-image of being a decent person despite acting immorally. We assume that the narrative is given, supplied by a third party. It can be a "narrative entrepreneur", a partisan of $a = 0$; or some cultural or situational context.[16] Alternatively, the individual may have "authored" the narrative herself. The analysis is then similar, but slightly more complex as it depends on when the narrative is acquired. The simplest case is when it was acquired "behind the veil of ignorance", that is before knowing her type. Then the analysis is the same as below (the search for a negative narrative can then be justified by a concern for oversignaling, as in condition (2) and that for a positive narrative by a concern for undersignaling). The case in which the individual already knows her type when searching for a negative narrative is more complex because the very search or presence of an excuse signals something about the type.

[15] Our focus is on "negative" narratives. We will return to the positive role of narratives in the context of moral development in section 4.

[16] Glaeser (2005) formalizes the notion that politicians seek to expand political power in sowing hatred against a minority through means of creating and spreading stories.

To formalize narratives suppose the decision maker receives a negative narrative telling her that the externality may not be that important anyway, or that she is not causing it. The narrative "N" is received always when there is no externality and with probability z when there actually is one, implying that the narrative contains a grain of truth or has some minimal plausibility. Thus the received signal σ is either "N" (narrative) or "1" (there is an externality). The information structure is depicted in Figure 1. Thus with probability $e(1-z)$ (no narrative), the high type contributes, since $v - \left(\frac{c}{\beta}\right) + \mu(v - \bar{v}) > 0$, from Assumption 1.

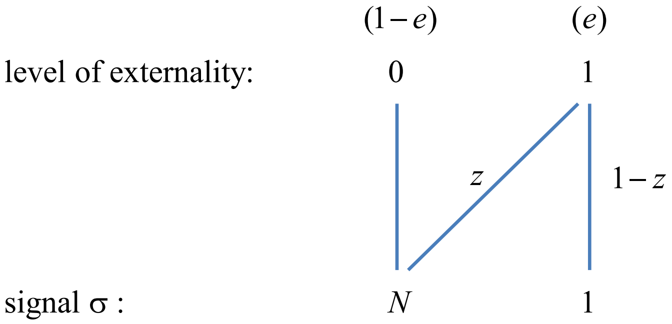

Figure 1: Information structure (negative narrative)

Let us assume that

$$e > e^*$$

so that in the absence of a signal (or equivalently, $z = 1$, which confers the lowest possible credibility on the "narrative"), the individual behaves morally.

If $e_1 \equiv \frac{ez}{[ez+1-e]}$ is such that

$$ve_1 - \left(\frac{c}{\beta}\right) + \mu(v - \bar{v}) \leq 0 \quad \text{or} \quad e_1 \leq e^* < e \tag{3}$$

then the narrative, although possibly very weak (z may be close to 1), is powerful as it eliminates any contribution.

Note that an excuse affects the individual's image and not only her incentive. The direct effect of an excuse is the decreased worthiness of the cause (a decrease in e), that reduces her motivation to contribute. The indirect effect operates through equilibrium behavior: The absence of contribution ($a = 0$) now carries less stigma, as the induced reputation is $\bar{v}$ instead of 0. Both effects go in the direction of reducing the contribution.

Proposition 3 *Suppose that the negative narrative leads to posterior belief e_1, where $e_1 \leq e^* < e$. The ex-ante probability of moral behavior falls from ρ to $e(1-z)\rho$. Thus even weak narratives (z close to 1) can have large effects on moral behavior.*

3.1 Applications

This simple result on effective narratives sheds light on a large number of empirical regularities, both from the field and the lab. We discuss a few of these in the following, focusing on narratives that downplay the externality, reduce the perception of being pivotal, magnify the cost of acting morally, or lead to omission bias as well as action and information avoidance (moral wiggle room).

(a) Countless narratives, sometimes backed up by pseudo-scientific insight[17], *downplay the externality* inflicted on third parties in suggesting that the victim is undeserving, that violence against the victim is an act of "self-defense"; or using allegations of conspiracy, power and control. To illustrate, consider the case of religious narratives lending interpretative support for the Indian removal policy and atrocities in nineteenth century America. Keeton (2015) convincingly argues that with recourse to the Exodus narrative from the Old Testament, Native Americans were casted "as characters in a biblical script" (p.126), a framing that endorsed the Indian Removal Act of 1830, ultimately legitimizing the harm of their removal. In his detailed analysis he identifies specific references drawn from the Bible: The Exodus story signifies America as the New Israel (doctrine of discovery), the story of creation motivates subduing tribal lands, and the story of Jacob and Esau serves to reduce the target and to deny harm. In sum, these narratives supported removal in portraying Native Americans as "deserving" relocation "as people who failed to uphold the divine mandate to till the soil" (p.144). A recent field study on the Rwandan genocide provides another illustration of the power of narratives: Yanagizawa-Drott (2014) studies the role of state-sponsored propaganda against the Tutsi minority. Using variation in radio coverage and violence, he estimates that about 10 percent of the overall violence can be attributed to "hate radio". An infamous historical example concerns the *Protocols of the Elders of Zion*, the most widely distributed antisemitic publication of modern times. The Protocols are fiction but have turned out to "credibly" blame Jews for conspiracy and other ills, rendering fighting them a "legitimate act of self-defense".[18]

Narratives of downplaying externalities are often encoded in wording and language. Using "dehumanizing language" Nazi ideology, e.g., degraded Jews and other victims to a "lesser" race and a "subhuman" level, facilitating their killing (Levi, 1988; Zimbardo 2007; Bandura 1999; Glover 2012; Petrinovich and O'Neill 1996). Likewise, euphemistic labeling by means of "sanitizing language" is used to camouflage morally problematic activities (Gambino, 1973). Examples comprise soldiers "wasting" people rather than killing them, referring to bombing missions as "servicing the target" or attacks as "clean, surgical strikes" leading to "collateral damage" in case civilians are killed; framing reactor accidents as a "normal aberration" or lies as "a different version of the facts" as in the Watergate hearings (see Gambino, 1973, and Bandura, 1999). Finally, note that reinterpretation by language is also key in understanding "moral framing" effects. Slight semantic shifts in an otherwise identical choice problem have been shown to systematically affect moral judgment. An example is the so-called Asian disease problem, where differences in outcomes stem from framing a decision either as "lives saved" or "lives lost" (Kahneman and Tversky, 1984).

(b) Narratives often suggest lack of responsibility alluding to *not being pivotal*: "If I do not do it, someone else will". This consequentialist reasoning reflects the idea that the

[17] An example is the "scientific" support of global warming being a natural phenomenon, rather than being being manmade.

[18] See also the literature on blaming and scapegoating, which are narrative based forms of reinterpretation of victims, in an attempt to reduce perceived externalities (Darley 1992, Glick 2002).

expected externality exerted by one's immoral behavior is reduced if the third party is likely to be hurt by someone else anyway. The classic example for the logic of diffusion of responsibility is the so-called bystander effect (Darley and Latané, 1968, and for a recent meta study Fischer et al., 2011). Typical bystander experiments study pro-social behaviors such as helping behavior in response to a staged emergency (e.g., the experimenter becomes injured). These studies often find that aggregate helping behavior is inversely related to group size. Subjectively plausible narratives in this context are that someone else will probably take responsibility, that other group members are more qualified to help, or that helping may result in negative legal consequences. A particularly striking historical example for the diffusion of responsibility is reported in Lifton (1986). He interviewed German doctors stationed in Auschwitz. They were operating in a nightmarish environment with one of their objectives being to "select" between prisoners who would be allowed to live and those who would be gassed right away. One of the frequently made justifications for the obvious evil was that the "horrible machinery would go on", regardless of whether or not a particular doctor would continue to participate.[19]

(c) For low externality and low pivotality, the excuse relates to the possibility of minor consequences. Alternatively, we could have formalized an excuse as a *magnification of the cost c*. "I only followed orders" or "I had a bad day" reminds the audience (self or external) that there was a non-negligible personal cost to behaving ethically.

(d) A further application concerns *acting by omission vs. by commission.* Many people share the intuition that, e.g., withholding the truth is not as bad as lying, or that stealing from a poor person is worse than not supporting him. We generally consider harmful omissions morally more acceptable than committing harmful acts, even if consequences are identical (Baron and Ritov, 2004). There is a vivid debate in philosophy, psychology, and law[20] about why we draw a distinction between omission and commission. As discussed in Spranca et al. (1991), omissions may result from ignorance while commissions usually do not, as the latter involve stronger effort, intention and attribution of causality. A well-known questionnaire study in this context asks subjects about rating a new vaccination showing that subjects favor inaction over statistically preferable action (Ritov and Baron 1990). Similarly, in Spranca et al. (1991) subjects read scenarios involving either an omission or a commission. Subjects are then asked to judge the morality of the actor in the given situation.[21] They often rated these scenarios, subjects often rated the harmful omission as less immoral than the harmful commission.

These findings mirror the moral intuition that failing to disclose privately-held information is less damaging to self- or social-esteem than a deliberate misrepresentation of the

[19] We will return to the issue of diffusion of being pivotal in section 6 on organizational design.

[20] In fact, the law treats the two differently (Feinberg, 1984). In American constitutional law, e.g., people have a right to withdraw life-sustaining equipment but no right to physician-assisted suicide (Sunstein, 2005).

[21] The context is a tennis tournament where the actor plays against another tennis player who is known to be the better player and having a particular allergy. The night before the match they meet for dinner. The actor, but not the other tennis player, knows that one of the possible two dishes contains an allergic substance. In one scenario the other player orders the allergic one and the actor does nothing to influence this decision. In the other scenario the other player orders the non-allergic one but the actor recommends the allergic dish to him. In both cases, the other player gets a bad stomach ache and the actor wins the tournament, i.e., outcomes are identical.

truth, even if the two behaviors have similar consequences. In our mind, the key difference between the two is the ex-post plausibility of an excuse. In the case of an omission ("I forgot", "I did not draw the connection", "I was in a rush"[22], and even "OK, but I did not lie") behavior can be interpreted as mere thoughtlessness (see also Spranca et al., 1991). In contrast, the act of fabricating/enouncing false evidence is much harder to rationalize other than as a deliberate act to deceive the other party. In the framework of the model, the reputational inference of an omission is weaker than for an act of commission.

(e) *Moral wiggle room.* In the presence of a morally demanding choice problem, people tend to avoid action or information. An intuitive example is changing sidewalks to avoid being confronted with a beggar. This intuition is supported by a series of papers on action and information avoidance. Dana, Weber and Kuang (2007), e.g., study a dictator game with hidden information about consequences. Although uncertainty can be resolved at no cost, many subjects choose to remain ignorant. In comparison to an otherwise identical dictator game without uncertainty about consequences, decisions in the hidden information treatments are significantly less prosocial. Relatedly, subjects avoid environments in which sharing is possible (Lazear, Malmendier and Weber, 2012 and Oberholzer-Gee and Eichenberger, 2008), or delegate decision rights to a principal-friendly decision-maker (Hamman, Loewenstein and Weber 2010, Bartling and Fischbacher, 2012). In both cases, prosocial behaviors are significantly less frequent in comparison to identical games that do not allow for action avoidance. In terms of our model, the unwillingness to learn the state of nature or to avoid/delegate choices can have one of two foundations:

a) The choice is made at a moment at which the agent does not yet know his type (v or 0). Then he wants to avoid having to self-signal if, as we saw in Proposition 2, $(c/\beta) - c < \mu(v - \bar{v})$, that is if he will feel compelled to act nicely if confronted with the opportunity. There are then two possible strategies:

   - avoiding actions, by eliminating the pro-social option (the action set is restricted to $a = 0$). This is the strategy of changing sidewalks so as not to be confronted with the beggar.
   - avoiding information, making $a = 0$ the equilibrium, as in the experiments on information avoidance.

b) The agent uses a feeble, but undebunked narrative ("I may not have hurt the other", "Someone else did it") to rationalize away the antisocial action. For instance, one may want to delegate the decision to an agent for "legitimate" reasons (eliciting a second opinion, delegated expertise). This explanation raises the issue of fragile excuses, and of the failure to debunk the narrative. This will bring us to a discussion of whether narratives are likely to be examined (see Section 3.3).

[22] A nice example of using the narrative excuse of being in a rush is the study of Dana et al. (2007) who show that subjects in a dictator game give less when facing a *non-binding* time constraint. If subjects do not decide in due time they automatically give nothing, an act of omission.

3.2 Sharing narratives

An intriguing consequence of narratives in the context of agents with image concerns is people's preference to share their narratives with others. Intuitively, if a decision maker misbehaves and learns an excusing narrative, his reputation vis-à-vis others will improve if these others know about the narrative as well. Our model thus provides a reputation-based mechanism for the fact that narratives, and moral judgments in general, will spread through the population (Haidt et al., 2009): Once a narrative exists it will "spread out over time and over multiple people", and it will "circulate and affect people, even if individuals rarely engage in private moral reasoning for themselves" (Haidt, 2001, p. 828-829).

The social-interaction context considered here involves a single decision-maker with an audience watching him. The reputational term reflects the desire to gain social prestige vis-à-vis peers. In a first step we assume that only one individual, the agent, has a decision on a pro-social action. Let us compare two situations, assuming, say, that the audience (the peers) either does or does not receive the narrative:

(i) *Shared narrative.* Peers know whether the individual has received the negative narrative or not (on top, of course, of observing a). Then the immoral equilibrium under the negative narrative prevails if and only if $e_1 < e^*$ (see above).

(ii) *Privately-known narrative.* Peers do not know whether the individual has received a narrative. Let $v_L \equiv \rho[1 - e(1 - z)]v/[(1 - \rho) + \rho[1 - e(1 - z)]]$ denote the expectation of the agent's type conditional on $a = 0$ and on being in an equilibrium in which the v type does not contribute when receiving the negative narrative. Note that $v_L < \bar{v}$, and so type v contributes when (secretly) receiving narrative N if

$$ve_1 - \left(\frac{c}{\beta}\right) + \mu(v - v_L) > 0. \tag{4}$$

This means that the absence of narrative sharing puts more pressure on the individual to contribute. Immoral behavior is facilitated by a shared narrative, which provides the individual with an "excuse" for not contributing.

Proposition 4 *When* (4) *is satisfied, the individual, regardless of her type, is strictly better off when the narrative is shared than when it is not (disclosure is irrelevant when* (4) *is violated).*

Self-serving disclosures. We have assumed that the disclosure of the narrative, if any, comes from a third party who is unknowledgeable about the individual's type. Suppose by contrast that the individual herself chooses whether to disclose the narrative, and that disclosure involves an arbitrarily small, but positive cost. Because the moral action $a = 1$ can only come from the high type, no disclosure will ever be associated with the moral action. Next consider the immoral action $a = 0$, for which the high and low types have exactly the same payoffs. In equilibrium, both types disclose the narrative (if there is one) to the peers and thereby obtain utility μv_L.

Viral spreading of negative narratives. This last observation can be used to show how narratives can spread endogenously to generate immoral behavior. Suppose that agents $i = 1, 2, \ldots$ pick $a_i \in \{0, 1\}$ sequentially and that agent i cares about his image vis-à-vis agent $i + 1$ (the same reasoning would still work if he cared about the opinion of all other agents). For simplicity, we will assume that an individual, when passing on the negative narrative, does not internalize the induced changes in behavior of other individuals: She only aims at preserving her reputation vis-à-vis the next peer. If agent 1 receives or conceives the negative narrative, then agent 1 behaves immorally and passes it along to agent 2 as long as the cost of doing so does is small enough. In turn, agent 2 behaves immorally and passes it along to agent 3 and so forth. Thus a negative narrative can spread virally and generate a cascade of morally questionable behaviors. We think that the proposed mechanism for the endogenous dissemination of narratives provides an important insight as to why "ideologies develop their own legs" (Glick, 2002 p. 137), potentially transforming belief systems of groups and societies.

3.3 Debunking and upholding a shared narrative

We now add the possibility of examining the narrative. Suppose at cost ψ (perceived as ψ/β due to present-biased preferences), the individual can find out whether the narrative really makes sense (in which case the externality is 0) or not (in which case the externality is 1, say).

Does the high type want to examine the validity of the narrative? We can entertain various assumptions concerning what the individual knows and/or recalls, with qualitatively similar results. Let us assume that the individual is aware of his type when choosing whether to question the narrative and that the debunking choice is not observed by the audience (social signaling, or self-signaling with imperfect memory about this choice). Let the debunking choice be made by the date-1 self. The no-investigation, no-prosocial behavior equilibrium exists if and only if the high type does not want to investigate in a pooling situation in which none of the types questions the narrative.

$$e_1 \left[v - \frac{c}{\beta} + \mu[v - \bar{v}] \right] \leq \frac{\psi}{\beta}. \tag{5}$$

Application: One interpretation of ψ is the nature of the choice context the decision maker is facing: The complexity of the decision problem, distraction and lack of attention, or simply time pressure may create effective constraints on the feasability of examination. Alternatively, ψ can be understood as an individual-specific skill parameter: Investigating whether a narrative actually makes sense requires reflection, which may be limited by available cognitive skills. This interpretation would imply a positive correlation between cognitive skills and moral behavior.[23] Suggestive evidence is provided in Deckers et al.

[23] There are two caveats with respect to expecting a general link between cognitive skills and moral behavior, however. First, in case of self-constructed narratives (see section 3), more intelligent people may be more creative in making up a convincing narrative to start with. Second, if narratives are promoting moral behavior, rather than serving as an excuse for doing bad (as we study in section 4), debunking narratives would imply a negative correlation between IQ and moral behavior.

(2016) who show that more intelligent subjects are less likely to engage in immoral decision making than subjects with lower IQ.[24]

4 Moral development and imperatives

4.1 Narratives vs. imperatives

This section models the formation of morality as a persuasion game between an agent and a principal. The latter can be thought of as an ex-ante incarnation of the individual, a parent, society, or as political or religious leaders. Depending on his objectives, the principal wants to generate a particular behavior from the agent. At the principal's disposal lie several routes of persuasion. A *narrative* is like earlier an argument that the agent may use to reach a decision. It can be defined as a signal regarding the parameters such as externality, cost or visibility of behavior. In contrast an *imperative* refers to a recommended behavior (e.g., to $a = 1$). It does not focus on the motive for the decision, but on the decision itself; it is a moral precept to behave in a certain way.

Both narratives and imperatives are commonly used to instill moral behavior. Numerous scholars have argued that oral, written, and cinematic narratives are essential components of "effective moral education" (Vitz 1990, p. 709).[25] Imperatives, in contrast, do not take the form of stories, they are commands of the form "Thou shalt not kill" or "Thou shalt not lie", as codified in the Ten Commandments. The notion of imperatives is reminiscent of deontological or rule-based moral reasoning, famously put forward by Immanuel Kant. An essential difference between utilitarianism/consequentialism on the one side and deontological normative theories is that in the former only consequences (pleasure, happiness, welfare) conceivably justify moral decision making, whereas the latter postulates a categorical demand or prohibition of actions, no matter how morally good their consequences: what makes a choice right is its conformity with a moral imperative.[26] The tension between these concepts has been illustrated in the Trolley or the Transplant

[24] They also find that the damaging effect of markets on morality reported in Falk and Szech (2013) is less pronounced, the higher subjects' IQ. Further evidence is reported by Corgnet et al. (2016) showing a positive correlation between the outcomes in a Cognitive Reflection Test (CRT) and altruism in simple non-strategic decisions. Based on a globally representative survey data set, Falk et al. (2015) find that cognitive skills are positively correlated with altruism, and that the likelihood of donating money, helping a stranger or volunteering time are all significantly increasing in the level of cognitive skills.

[25] The psychological rationale for the importance of stories (think, e.g., of children's stories and fairy tales or movies portraying moral heroes) in developing pro-social conceptions is that they allow identification and self-representation, increase empathy and awareness, and provide positive role models, see, e.g., Bennett 1993, McAdams and Koppensteiner 1992, Tappan and Brown 1989, Mar and Oatley 2008, Johnson 2012. See also the empirical literature on moral persuasion in experimental economics, e.g., Dal Bó and Dal Bó (2014), and references therein.

[26] The discussion about these two main lines of thought in occidental moral philosophy is, of course, much more involved, see Alexander and Moore (2015). For example, imperatives (or rules) as modeled here are not unconditionally justified, in contrast to deontological reasoning. Our notion is more akin to so-called rule-consequentialism which understands an act as morally wrong if it is forbidden by rules that are justified in terms of their consequences. Similarly, philosophers (starting with John Stuart Mill and more recently in Hare, 1993, and Cummiskey, 1996) have suggested a *teleological* reading of the categorial imperative, as a means to produce the best overall outcome.

problem, where the act of killing one implies saving many. While consequentialist reasoning requires killing (throwing a man in front of a trolley to stop it; killing the healthy patient to use his organs), deontological reasoning calls for obeying the imperative not to kill, regardless of the consequences and without reference to any ends. Empirically, both notions co-exist.

In the formal analysis below, the principal issues a narrative or an imperative (or the two). We show that imperatives can emerge in a population of purely utilitarian individuals, and unveil the factors that are conducive to their existence. These factors stem from both demand and supply, i.e., characteristics of agent and principal.[27] First, imperatives are effective only if issued by highly trusted principals, while everyone can use the narrative route to attempt to persuade. Second, imperatives, precisely because they focus directly on the decision, do not let the agent adapt her behavior to contingencies; they therefore imply some rigidity in decision-making. Third, relative to narratives, the cheap talk nature of imperatives makes them less fragile to interpretation uncertainty or to the threat of debunking. Fourth, imperatives allow to pool states in which the agent would be reluctant to behave in the recommended way with other states where he would be eager; imperatives thereby expand the scope for the desired behavior.[28] On the demand side, the agent, like in any communication game, will be more influenced by the principal if they are more congruent. Also, Proposition 2 suggests that, if the communication game occurs at date 0 (that is, involves the ex-ante self and an external principal), the agent will be more receptive to messages leading to $a = 1$ if she is more hyperbolic and has low image concerns (β and μ are low).

The theoretical predictions map into several empirical findings, but also suggest new avenues for empirical research. For example, in moral dilemmas such as the Trolley problem there is a puzzling coexistence of deontological and consequentialist reasoning. While from a consequentialist perspective it is clearly preferred to kill one in order save five, many people's intuition prohibits the act of killing. Likewise, no general consensus is reached when it comes to so-called repugnant goods (Roth, 2007). While advocates of markets for organs, sex or surrogate mothers highlight potential welfare gains, skeptics refer to the imperative not to treat humans as a means, but always as an end in itself. The inherent rigidity of imperatives is connected to the rules vs. discretion idea. In this sense, imperatives can be viewed as some moral heuristic that works well in many cases but may misfire or actually produce moral mistakes, at least from a utilitarian point of view (Sunstein, 2005). The relation between imperatives and low levels of self-control and self-image, draws a possible connection to imperatives as commitment devices: sticking to simple behavioral rules may help attenuating the undersupply of moral behavior resulting from an agent's present bias or lack of self-image. A minimum of congruence as a prerequisite for imperatives suggests why imperatives might work better in parent-child

[27] Kant has formulated his categorical imperative from both perspectives; agent ("act only in accordance with that maxim through which you can at the same time will that it become a universal law." (Kant, 1785, 4:421)) as well as principal ("the Idea of the will of every rational being as a will that legislates universal law" (Kant, 1785, 4:432)), i.e., both in terms of a universal law giver as well as universal law followers (see Johnson, 2014).

[28] Other desirable features of imperatives that are not captured by our model (but could be) include lower communication costs (it is often easier to define an action than to explain reasons why this action is desirable) and, due to their 0/1 nature, easier monitoring of compliance.

relations[29] rather than between loosely related interaction partners. Likewise, moral authority is a precondition for imperatives to work, which sheds light on why narratives are common in the political arena, while religious leaders can also rely on imperatives.

4.2 Modeling

Suppose that the principal learns a narrative leading to posterior belief e, while the agent does not. The prior distribution over $e \in [0,1]$ is $F(e)$ with density $f(e)$ and mean $E_F[e] = e_0$, the prior mean probability that there is an externality. A convincing positive[30] narrative is defined as one such that

$$e > e^* > e_0,$$

where, recall, $ve^* - (c/\beta) + \mu(v - \bar{v}) = 0$. So we assume for conciseness that good behavior is not innate so that the agent does not behave prosocially unless prompted by a narrative or an imperative: $e_0 < e^*$.

The principal's objective function, conditional on type v, is $U^P(e)a(v)$; similarly, the agent's utility when having type v is $U^A(e)a(v)$. We assume that $U^P(\cdot)$ and $U^A(\cdot)$ are increasing affine functions and that, as depicted in Figure 2, the principal wants to promote pro-social behavior. Let $e^P < e^*$ be defined by $U^P(e^P) = 0$ and $U^A(e^*) = 0$. We will assume that the agent knows $U^P(\cdot)$, i.e., how trustworthy the principal is. In what follows, we will identify e^P with the congruence between the principal and the individual: $e^P = e^*$ indicates perfect congruence; and as e^P decreases, so does the level of congruence.

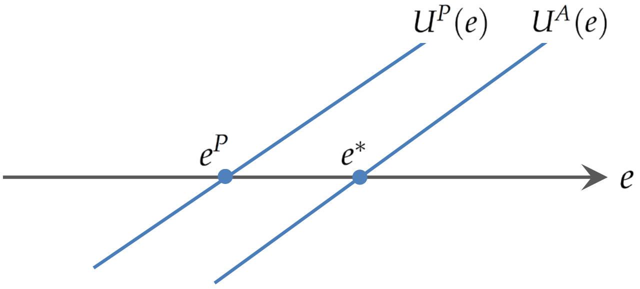

Figure 2: P and A's preferences

For example, the principal may maximize a weighted average of the third party's benefit from a pro-social action and the individual's ex-ante welfare (ex ante, the individual discounts the future less than his present-biased self; furthermore, reputation is positional and investments in a self- or social image is a zero-sum game):

$$E_{\tilde{v}}\left[\left[we + (\tilde{v}e - c)\right]a(\tilde{v})\right] \equiv \rho\left[we + (ve - c)\right]a(v).$$

[29] Empirically, parents not only place high value on the utility of their children (corresponding to a low weight w in the principal's objective function, see below), they are also similar in terms of their preferences (Dohmen et al., 2012). For models of intergenerational transmission of values see Bisin-Verdier (2001), Dessi (2008) and Tabellini (2008).

[30] Similarly, a convincing negative narrative would be defined as one satisfying $e < e^* < e_0$.

Weight $w = 0$ corresponds to parents maximizing their child's welfare, $w = \frac{1}{2}$ to a utilitarian social planner and $w = +\infty$ to a private entity or any party wanting to promote pro-social behavior but having no empathy for the individual; $w = -\infty$ would represent an immoral entrepreneur with no empathy for the agent. In the context of this example, we thus focus on a pro-social principal ($w \geq 0$) for conciseness.

It may be useful to examine the conflicts of objectives. First, we see that the principal desires more pro-social behavior than the agent's ex-ante self. Second, comparing the agent's ex-ante and ex-post selves, the latter's present bias creates an extra cost $c\frac{1-\beta}{\beta}$, leading to an undersupply of pro-social behavior from the point of view of the ex-ante self; by contrast, the ex-post self's image concern $\mu\,(v - \bar{v})$ leads to an oversupply of pro-social behavior from the point of view of the ex-ante self. Thus strong image concerns and low present bias create a demand against moral rules while weak image concerns and high present bias create a demand for moral behavior.

The timing goes as follows. First, the principal chooses between disclosing the narrative and issuing an imperative (say, "$a = 1$"), or both.[31] The agent then selects $a \in \{0, 1\}$.

4.3 Uncertainty about interpretation/communication

Suppose that when faced with the narrative, the agent understands it with probability $x < 1$ and fails to fathom the argument with probability $1-x$ ("babbling"); the probability of understanding the narrative can be arbitrarily close to 1. The agent is unable to distinguish between a narrative she does not understand and a meaningless narrative that the principal can always provide regardless of the value of e.

For the moral imperative to be effective in equilibrium (i.e., trigger action $a = 1$ as recommended), it must be the case that

(i) Anticipating obedience, the principal recommends $a = 1$ if and only if $U^P(e) \geq 0$, or $e \geq e^P$.

(ii) The (high type) agent obeys the imperative. That is, she picks $a = 1$ when told so. This requires that

$$M^+(e^P) \equiv E[e \mid e \geq e^P] \geq e^* \tag{6}$$

Suppose that condition (6) is satisfied. If the agent is provided with a babbling narrative, her posterior beliefs are, say, $M^-(e^P) = E[e \mid e < e^P] < e^*$ and so she picks $a = 0$.

The principal's behavior is then optimal as a principal with information $e \geq e^P$ prefers $U^P(e)$ to what is obtained through the narrative (or for that matter through the combination of a narrative and an imperative), namely $xU^P(e)$ if $e \geq e^*$, and 0 if $e < e^*$. By contrast, there is no equilibrium imperative if condition (6) is violated. Provided that $x < 1$, this is also the unique equilibrium under our Pareto selection criterion.

[31] See Dewatripont-Tirole (2005) for an analysis of rival modes of communication. In equilibrium, the principal will not use both routes. Intuitively, disclosing the narrative on top of the imperative exposes the principal to the risk of miscommunication described below. Conversely, the principal chooses the narrative route only if the imperative route is ineffective. More on this below.

We can also perform some comparative statics. Factors facilitating the emergence of an imperative are:

(i) *Congruence.* Suppose that e^P increases, then $M^+(e^P)$ increases and so (6) is more likely to be satisfied.

Interestingly, the principal should not be viewed as too much of an advocate for pro-social actions (that is, w should not be too high in the weighted-utility illustration of $U^P(\cdot)$). To be convinced, the agent must perceive the principal as standing for her interests. Principals who are too dogmatic about the course of action to be taken will not be listened to.

(ii) *Ability to make sound judgment (moral authority).* Let the distribution of e be indexed by a parameter ρ indexing the precision of the principal's information. More accurate narratives intuitively mean a spread in the induced posterior beliefs. If $\partial M^+(e,\rho)/\partial\rho > 0$, then condition (6) is more likely to be satisfied as ρ increases. This condition is satisfied in the case of the uniform, Pareto and exponential distributions,[32] or more generally when ρ denotes a rotation ($\int edF(e,\rho) = e_0$ and $e < e_\rho \Rightarrow F_\rho \geq 0$ and $e > e_\rho \Rightarrow F_\rho \leq 0$) and $F_\rho(e^P,\rho) \geq 0$.

If we interpret ρ as a parameter of capability of good judgment, principals with high moral standing (in the sense of safe judgment, not dogmatism as we just saw) are more likely to be able to issue imperatives; narratives by contrast can be spread by everyone.

(iii) *Large expected externalities.* Index the distribution by a shift parameter θ: $F(e-\theta)$ with corresponding mean $e_0+\theta$. Then, assuming that the hazard rate $f/[1-F]$ is increasing (as it is for familiar distributions), if $\theta_1 < \theta_2$

$$M^+(e^P,\theta_1) \geq e^* \Longrightarrow M^+(e^P,\theta_2) \geq e^*,$$

so large expected externalities facilitate the emergence of imperatives.[33]

Proposition 5 *Suppose that there is at least a slight probability of miscommunication of a narrative. Then the equilibrium is unique. It features an imperative if $M^+(e^P) \geq e^*$, and if $M^+(e^P) < e^*$, a narrative if $e \geq e^*$ and no communication otherwise.*

An imperative is more likely when congruence is high, when the externality is likely to be large, and when the principal is perceived to have sound judgment.

[32] Uniform: $e \sim U[e_0-\rho, e_0+\rho]$
Pareto: $1-F(e,\rho) = (1/\rho e)^{\alpha(\rho)}$ on $[1/\rho,\infty)$ where $e_0 = \frac{\alpha(\rho)}{\alpha(\rho)-1}\frac{1}{\rho}$.
Exponential: $1-F(e,\rho) = e^{-\lambda(\rho)(e-(1/\rho))}$ on $[1/\rho,+\infty)$ where $e_0 = \frac{1}{\rho}+\frac{1}{\lambda(\rho)}$.
In all three cases, ρ denotes a rotation (with rotation points e_0 for the first and the third, and above e_0 for the second). For the last two examples, e should be interpreted as the magnitude of the externality rather than a probability (so it can exceed 1).

[33] To show this, note that $M^+(e^P,\theta) = \theta + M^+(e^P-\theta)$. Furthermore, $(M^+)' \in (0,1)$ if the hazard rate condition is satisfied.

Combining narratives and imperatives

In Proposition 5 above, the principal uses either a narrative or an imperative, but not both. This is because the narrative is a sufficient statistic for what can be known, and so an imperative has no information value once the narrative has been disclosed. In practice, morals and religions as well as upbringing processes combine a variety of narratives and imperatives; for instance, a story about someone who has stolen or lied and had to regret it badly, followed by a generalization to "thou shalt not steal/lie".

The insights of this section still hold, except for a possible co-existence of narratives and imperatives, once one allows for multiple narratives. Suppose for instance that there are multiple positive narratives and that congruence is too small to generate a credible imperative on a stand-alone basis. The principal may then start with one or several positive narratives, that raise the perceived level of congruence enough so that an imperative becomes effective.

Providing a formal treatment of the narrative-imperative mix lies outside the scope of this paper. In general, one will have to make assumptions on how narratives combine to generate updated beliefs and how likely it is that they will be understood by the agent. Here we content ourselves with a highly stylized example, whose only purpose is to make the point that the principal's optimal strategy may combine a narrative and an imperative. Let $\hat{e}$ be defined by $M^{+}(\hat{e}) = e^{*}$, and suppose that $e^{P} < \hat{e}$. Let the principal have a signal: N, received when $e < \hat{e}$, and P, received when $e \geq \hat{e}$. The principal can disclose this signal (which will then be understood with probability equal to 1, say). Then disclosing P induces the moral agent to contribute. Note that we could add a more precise signal, say the true value e. But as in Proposition 5, the principal would not want to release this second signal and would content himself with the imperative to contribute.

4.4 Demand for flexibility

Let us now consider the case of a privately informed agent. His information could be some independent signal; alternatively, as will be formalized here, it could be some thought that is generated by combining the disclosure of the narrative and the agent's own experience or information (by contrast, an imperative does not trigger such thinking).

Let us thus assume that provided with narrative e, the agent formulates externality judgment σ distributed according to differentiable function $G(\sigma \mid e)$ with $E_G(\sigma \mid e) = e$ such that (i) an increase in e shifts the distribution of σ to the right in the sense of the monotone-likelihood-ratio property $\frac{G(\sigma|e_2)}{G(\sigma|e_1)}$ is increasing in σ if $e_1 < e_2$ and (ii) σ is a sufficient statistic for (σ, e).[34] From (ii), U^P and U^A depend on σ, not on e. Let

$$V^P(e) \equiv \int_{e^*}^{1} U^P(\sigma) dG(\sigma \mid e)$$

denote the principal's welfare under a narrative.

Let us find conditions under which an imperative exists. Let $I \equiv \{e \mid P$ picks the

[34] For instance, $G(\sigma \mid e) = \sigma^{\frac{e}{1-e}}$ (so $E(\sigma \mid e) = e$) on $[0, 1]$.

imperative "$a = 1$"}. Obedience by the agent requires that

$$E[e \mid e \in I] \geq e^*. \tag{7}$$

To determine the set I of principal's types selecting the imperative, note that picking the narrative switches behavior from $a = 1$ to $a = 0$ whenever $\sigma < e^*$, and so equilibrium behavior requires that

$$\Delta(e) \equiv \int_0^{e^*} U^P(\sigma) dG(\sigma \mid e) \geq 0 \quad \text{for all } e \in I, \tag{8}$$

while $\Delta(e) \leq 0$ for $e \notin I$ ($-\Delta$ is the "value of information"). Also (8) is never satisfied at $e = e^P$:

$$\int_0^1 U^P(\sigma) dG(\sigma \mid e^P) = U^P(e^P) = 0 = \int_0^{e^*} U^P(\sigma) dG(\sigma \mid e^P) + \int_{e^*}^1 U^P(\sigma) dG(\sigma \mid e^P)$$

where the first equality results from the linearity of U^P; and so $\Delta(e^P) < 0$. Under the monotone-likelihood-ratio property[35], there exists $e^\dagger > e^P$ such that $\Delta(e) \geq 0 \Leftrightarrow e \geq e^\dagger$ and so $I \subseteq \left(e^\dagger, 1\right]$. This result is reminiscent of that in Bénabou-Tirole (2002), who show that a more self-confident individual is more information averse. Here the more optimistic principal selects the imperative, a form of information aversion.

We thus conclude that an imperative exists if and only if

$$M^+(e^\dagger) \geq e^* \quad \text{where} \quad \Delta(e^\dagger) = 0. \tag{9}$$

Congruence. As congruence (again measured equivalently by e^P or $(-w)$) increases, the principal is more and more keen on delegating decision-making to the individual ($e^\dagger$ increases) and the probability that an imperative exists increases. However, conditional on the existence of an imperative, the probability that it is chosen decreases with congruence and tends to 0 as the two parties converge to perfect congruence.

Self-control. This analysis illustrates the ambiguous impact of self-control on the possibility of an imperative. On the one hand, as the self-control problem worsens, *the principal is more tempted to go for an imperative* (e^* increases and so $e^\dagger(e^*)$ given by $\int_0^{e^*} U^P(\sigma \mid e^\dagger) dG(\sigma \mid e^\dagger) = 0$ decreases: I expands). On the other hand, at some point *the obedience condition $M^+(e^\dagger(e^*)) \geq e^*$ may no longer be satisfied* as $M^+(e^\dagger)$ decreases and e^* increases. So a worsening self-control problem facilitates the emergence of an imperative, but only up to a point.

[35] Recall that U^P is affine; so let $U^P(\sigma) = \alpha\sigma - \gamma$. Then

$$\Delta(e) = \int_0^{e^*} (\alpha\sigma - \gamma) dG(\sigma \mid e) = \alpha G(e^* \mid e) \left[\int_0^{e^*} \frac{\sigma dG(\sigma \mid e)}{G(e^* \mid e)} - \frac{\gamma}{\alpha} \right] = \alpha G(e^* \mid e) \left[e^* - \frac{\gamma}{\alpha} - \int_0^{e^*} \frac{G(\sigma \mid e)}{G(e^* \mid e)} d\sigma \right].$$

Finally, the monotone-likelihood-ratio property implies that $\frac{G(\sigma|e)}{G(e^*|e)}$ is decreasing in e for $\sigma < e^*$, and so $\Delta(e_2) > 0$ if $\Delta(e_1) \geq 0$ and $e_2 > e_1$.

Proposition 6 *Suppose that the individual can use private information to refine the principal's narrative and so imperatives have a cost in terms of flexibility. An imperative exists if and only if $M^{+}(e^{\dagger}) \geq e^{*}$ where $\Delta(e^{\dagger}) \equiv 0$.*

The probability of an imperative is non-monotonic in congruence: An imperative requires a minimum level of congruence; but beyond this level, the probability of an imperative decreases with congruence (to converge to 0 for perfect congruence). The effect of self control on imperatives is ambiguous.

5 The popularity of Kantians

5.1 Refraining from questioning moral imperatives

Viewed from a deontological perspective, imperatives have to be obeyed irrespective of consequences or calculating cost and benefits. In the following we provide a rationale for such *rule-worship* showing that even the very act of *questioning* the imperative, and not only violating it, can be a dangerous strategy.[36] The corollary of our analysis here is that we do in fact more than is subjectively morally warranted, in pretending to be more Kantian than we really wish to be. To study this, return to the basic framework. Assume, for technical simplicity only, that U^P does not depend on the agent's type (e.g., $U^P(e) = e - \kappa$, where κ is a constant). Letting as earlier e^P be defined by $U^P(e^P) = 0$, the imperative corresponds to realizations $e \geq e^P$. Suppose that there are two varieties of the high type v_1 and v_2, in proportions $1 - \lambda$ and λ, so

$$v = \lambda v_2 + (1 - \lambda)\, v_1;$$

that the better type v_2 is highly moral and, regardless of reputational incentives, always chooses $a = 1$ when the principal so desires:

$$v_2 e^P - \frac{c}{\beta} \geq 0.$$

Accordingly, type v_1 will be called the "morally fragile type". Suppose further that, if the principal issues an imperative instead of disclosing e, the agent can still learn e at an infinitesimal cost and that this questioning of the imperative is observable.

We look for an equilibrium in which

(i) The principal issues an imperative if and only if $e \geq e^P$.

(ii) The high type (whether v_1 or v_2) does not attempt to learn e and conforms to the imperative ($a = 1$), while the low type also does not attempt to learn e and picks $a = 0$.

[36] The argument here builds on Bénabou-Tirole (2009), in which asking for the price fetched in a repugnant transaction backfires even if the deal is not concluded in the end. We here add the idea that the agent can challenge an endogenous imperative set by a principal.

(iii) Were the agent to learn e, an off-the-equilibrium-path event, society would form posterior beliefs $v = v_1$.[37]

For this to be an equilibrium, it must be the case that type v_1 obeys the imperative:

$$\int_{e^P}^{1} \left(\frac{c}{\beta} - v_1 e \right) \frac{dF(e)}{1 - F(e^P)} \leq \mu \left[v - \frac{\rho(1-\lambda)v_1}{1-\rho\lambda} \right]. \tag{10}$$

Second, type v_1, if he acquired the information, would reveal his type. So he would pick $a = 1$ if and only if $v_1 e \geq c/\beta$. A sufficient condition (a necessary one if the information cost is low enough) for v_1 not to want to acquire the information

$$\int_{e^P}^{\frac{c}{\beta v_1}} \left(\frac{c}{\beta} - v_1 e \right) \frac{dF(e)}{1 - F(e^P)} < \mu \left(v - v_1 \right) = \mu\lambda \left(v_2 - v_1 \right). \tag{11}$$

The left-hand side of this inequality stands for the flexibility benefit of being informed, in that the agent does not feel compelled to behave morally when he does not really want to. The right-hand side represents the opprobrium raised by a departure from Kantianism, a cost that is borne even if the agent ends up behaving morally: Only morally fragile agents would dare to even question the imperative.

Simple computations show that the RHS of (10) exceeds the RHS of (11). Because the LHS of (11) exceeds the LHS of (10), condition (10) is verified if (11) is.

Proposition 7 *Suppose that there are three types, as described above: The highly moral type (who always wants to behave pro-socially), the morally fragile type and the immoral type. Let the principal issue an imperative. Then the morally fragile type does not question the imperative (even if the cost of doing so is zero) provided that* (11) *is satisfied: The morally fragile type mimics the Kantian behavior of the highly moral type by fear of being perceived as a "calculating individual". This behavior is more likely, the more congruent the principal and the higher the ratio of highly moral to morally fragile types.*

Thus the "calculating individuals", who question the imperative, are perceived as persons of mediocre moral standing even when they behave prosocially. And if the reputation loss is sufficient, they do not question the imperative and behave like their high-moral-standing Kantian counterparts. Individuals in equilibrium demonstrate the high moral standing by being Kantian and engaging in information avoidance. A recent study provides evidence for the popularity of Kantians. In a series of experiments Everett, Pizarro and Crockett (in press) show that participants who make deontological judgments in dilemma problems are preferred as social partners; they are perceived as more moral and trustworthy.

[37] This belief selection for example results from using the D1 refinement. Intuitively, type v_1 gains most from the information.

5.2 The incommensurable and the refusal to consider trade-offs

We just saw that questioning a dogma can be deleterious even if one does not abandon it. Another facet of Kantian behavior is the refusal to consider trade-offs, as formulated in Kant's famous statement, "*in the kingdom of ends everything has either a price or a dignity. What has a price can be replaced by something else as its equivalent; what on the other hand is above all price and therefore admits of no equivalent has a dignity*" (Kant 1795). The often used approach to eliciting preferences in experiments, the strategy method, is particularly apt at capturing Kant's intuition. We assume that the subject is asked at what minimum level of reward c she is willing to take the immoral action $a = 0$, knowing that the actual c will be drawn in some distribution $f(c)$ with support included in $[0, \infty)$ (in many experiments this distribution is taken to be uniform on some finite interval, but it need not be so).

Let the support of f be $[0, \bar{c}]$ with $\bar{c} \leq +\infty$. We first assume that the expected extrinsic incentive exceeds the maximal reputation gain: $L(\bar{c}) \geq \mu v$, where

$$L(\gamma) \equiv \int_0^\gamma \frac{c}{\beta} f(c) dc$$

Let γ^S (where "S" stands for "least-cost separating") be defined by

$$L(\gamma^S) = \mu v$$

(the loss of monetary reward implied by threshold γ^S is equal to the reputational benefit of masquarading as a high type). Any threshold $\gamma \geq \gamma^S$ is weakly dominated by threshold $\gamma = 0$ for the low type. To cut down on uninteresting cases, we assume that

$$\gamma^S > \beta v e. \tag{12}$$

For, if $\gamma^S \leq \beta v e$, the high type can set $\gamma = \beta v e$, his first-best threshold and obtain reputation μv (as we rule out weakly dominated strategies).

To further reduce the number of cases under consideration, let us assume that the probability of a high type, ρ, is low enough, so that

$$L(\beta v e) \geq \mu \bar{v}. \tag{13}$$

Condition (12) implies that the high type does not want to set a threshold exceeding γ^S, which suffices to reveal a high type. Suppose that the high type in equilibrium sets threshold $\gamma < \gamma^S$. Necessarily, the low type pools with some probability x and accepts all offers with probability $1 - x$. As long as $L(\gamma) \geq \mu \bar{v}$, the beliefs following threshold γ are given by

$$L(\gamma) = \mu \hat{v} \equiv \mu \frac{\rho v}{\rho + (1 - \rho) x},$$

where $\hat{v}$ is the reputation following Kantian behavior. And so the high type's payoff,

$$\int_0^\gamma vef(c)dc + \int_\gamma^{\bar{c}} \frac{c}{\beta} f(c)dc + \mu\hat{v}$$
$$= \int_0^\gamma vef(c)dc + E\left[\frac{c}{\beta}\right],$$

is increasing in γ. So the optimal payoff for the high type corresponds to the separating threshold $\gamma = \gamma^S$.[38]

Next, suppose that $\beta ve \leq \bar{c} < \gamma^S$. The equilibrium can no longer be separating, as $L(\bar{c}) < \mu v$. It involves Kantian behavior by the high type, who refuses all offers ($\gamma = \bar{c}$) and mixing by the low type, who also adopts a Kantian behavior with probability x and accepts all offers ($\gamma = 0$) with probability $1 - x$, where

$$L(\bar{c}) = \mu\hat{v}.$$

Finally, suppose that

$$L(\bar{c}) < \mu v$$

Then we can set $\gamma^S \equiv +\infty$ and the analysis is the same as in the previous case, with Kantian behavior by the high type and perhaps the low type as well. In particular, note that if $\bar{c} = +\infty$, the individual turns down arbitrarily large offers.

Proposition 8 *Suppose that the strategy method is used to elicit the individual's willingness to act prosocially and that the probability distribution of the incentive c has density $f(c)$ on $[0, \bar{c}]$ where $\bar{c} \leq +\infty$. Suppose that conditions (12) and (13) are satisfied. Let γ^S be defined by $L(\gamma^S) = \mu v$ if there exists a solution to this equation and $\gamma^S = +\infty$ otherwise.*

(i) If $\bar{c} \leq \gamma^S$, the high type adopts a Kantian behavior of never behaving immorally, whatever the incentive to do so (which can go to infinity if $\bar{c} = \gamma^S = +\infty$). The low type randomizes between the Kantian behavior and the fully immoral behavior of always behaving immorally.

(ii) If $\bar{c} > \gamma^S$, then the high type sets threshold γ^S and the low type always behaves immorally.

Several empirical papers suggest the existence of such cost-invariant moral behaviors. Chen and Schonger (2013) identify deontological motivation by varying the probability of a decision being consequential. In Falk and Szech (2014), about 10 percent of subjects choose not to kill a mouse conditional on the belief that the chance of being pivotal is exactly zero. In Falk and Szech (2015) the average amount of money needed to accept the killing of mice does not depend on whether one, two or three mice are at stake, suggesting rule-based thinking in the sense that the externality is irrelevant. Likewise, they collected

[38] If $L(\gamma) \leq \mu\bar{v}$, then the equilibrium is a pooling one. The high type's payoff in that range, $\int_0^\gamma vef(c)dc + \int_\gamma^{\bar{c}} (c/\beta)f(c)dc + \mu\bar{v}$, is increasing in γ.

evidence on various price-list treatments and find "bunching" at the Kantian behavior in the sense that there is no c that would push individuals to act immorally.[39]

6 Shared control and the sub-additivity of responsibility

This section is a first stab at looking at moral consequences if outcomes result from decisions taken by groups, rather than by a single individual. Intuitively, if control about outcomes is shared, patterns of being pivotal, accountability and image may be shifted. This in turn bears the potential to create a *sub-additivity of responsibility* in the sense that agents act less morally when interacting in groups rather than in isolation. Relevant applications are corporate social responsibility and the design of organizations[40], market interaction and voting. Here we study the moral relevance of three generic organizational features, both in terms of individual actions and collective outcomes: patterns of being pivotal, accountability of action vs. outcome, and individual vs. team incentives.

To formally study the sub-additivity of responsibility, we will use a dyad (our results generalize to more than two individuals). We differentiate between two production technologies, or patterns of being pivotal:

(a) *Individual veto power.* In some environments, it takes all to misbehave for the bad outcome to happen or, equivalently, each agent by herself can guarantee the moral outcome. One can have in mind a potential bilateral trade that, if concluded, would exert a negative externality on the rest of society, or an unethical accounting, risk management or safety behavior that each member of a division's executive team can report to upper management. More formally, suppose that there are two individuals, $i = 1, 2$. Each individual selects an action $a_i \in \{0, 1\}$; she can by herself make sure that the moral outcome is selected, i.e., each has veto power: $1 - a = (1 - a_1)(1 - a_2)$.

(b) *Collective veto power.* By contrast, there are environments in which all agents must behave well in order to enforce a moral outcome: $a = a_1 a_2$. In this context, if only one other group member misbehaves, an agent is no longer pivotal. It hence refers to the standard "If I don't do it, someone else will", so prominent in workers, management and

[39] This holds even when the maximum amount is 100 euros. This evidence does of course not show that there is *no* price at which subjects would eventually be willing to exchange life for money. It suggests, however, that for the given range of stakes subjects obey the imperative not to kill.

[40] Numerous accounts of organizational failure to comply with ethical standards come from the car, oil, financial or pharmaceutical industry: Disasters in the oil industry comprise the Piper Alpha case, the 1982 Ocean ranger disaster off Newfoundland, BP's Baku-Tvilisi-Ceyhan pipeline and Deepwater Horizon, and the Exxon Valdez disaster; two examples from the car industry involve the recent "clean diesel" scandal of Volkswagen or the case of Ford Pinto, which was sold for years despite executives being aware that the gas tank was likely to rupture (in particular in crashes) and thus incinerate car drivers; examples from the financial industry in the context of the financial crisis such as among mortgage originators or LIBOR rigging abound; finally think of the production and marketing of pharmaceuticals with bad side-effects such as Contergan/Thalodomide. The critical role of organizations has also been pointed out in most extreme cases such as the organization of the Holocaust; see the literature on the "production of evil" that asks "how organizations produce killers" (Darley 1992, p. 204).

countries' defending the manufacturing and sales of weapons to dubious clients, or in defendants' response to accusations of war crimes.

We will assume that choices are simultaneous. Let us, for now, further assume that the cost c per individual is incurred whenever the moral action is taken ($a = 1$), i.e., irrespective of whether agent i chooses $a_i = 1$ or $a_i = 0$. Thus, agents face *team incentives*, with rewards determined by the outcome rather than by individual behavior. Individuals are "consequentialists". That is, provided that she has type $v_i \in \{0, v\}$, individual i's valuation of the moral action $a = 1$ is $v_i e$. Agent i's utility is then given by:

$$\left(v_i e - \frac{c}{\beta}\right) a + \mu \hat{v}(a_i).$$

An important observation is that, as it stands, the predicted behavior is the same as with a single individual (a "dictator"), as long as there is no image concern ($\mu = 0$); for, individual i should reason in the following manner: "if $a_j = 1$ with individual veto power, or $a_j = 0$ with collective veto power, my behavior is irrelevant; so I should reason as if I were pivotal". Hence, with team incentives and in the absence of image concerns, agents will display the same behavior regardless whether they act in groups or in isolation.

In the following we will study situations where shared control does affect individual behaviors: To this end, we first study the basic model with $\mu > 0$. We then add some richness to the environment. In particular, we show that shared control together with two important organizational features – individual incentives and team accountability – may lead to more immoral behavior at the individual level. We will focus on symmetric equilibria. Furthermore, because an agent may no longer be pivotal, Assumption 1 no longer guarantees that contributing is a strictly dominated strategy for the immoral type. Thus, the refinement based on the elimination of strictly dominated strategies no longer pins down out-of-equilibrium beliefs. We will therefore make another weak refinement: We require that beliefs be monotonic $\hat{v}(1) > \hat{v}(0)$.

6.1 The cheap signaling effect

This section uses the basic paradigm, but assumes that there are two agents subject to team incentives and individual accountability. We rule out narratives and so both agents have the same estimate e of the externality brought about by $a = 1$.

In the case of collective veto power ($a = a_1 a_2$), we will keep assuming that if no one ever contributes, then $\hat{v}(1) = v$. This assumption is not crucial for Proposition 9, but it shortens the analysis and is natural.[41]

Proposition 9 *Under team incentives and individual accountability, individual behavior is more pro-social than in the single decision-maker case; this holds for collective veto power* ($a = a_1 a_2$) *as well as for individual veto power* ($(1 - a) = (1 - a_1)(1 - a_2)$).

Proof. See Appendix. ■

[41] It is implied for example by refinement D1 and the existence of arbitrarily small trembles giving an (arbitrarily small) probability that $a_i = 1$.

The intuition can be grasped from considering agent i's decision of whether to contribute, given by

$$E[a_j]\left(v_i e - \frac{c}{\beta}\right) + \mu[\hat{v}(1) - \hat{v}(0)] \gtrless 0 \qquad \text{if } a = a_i a_j,$$
$$E[1-a_j]\left(v_i e - \frac{c}{\beta}\right) + \mu[\hat{v}(1) - \hat{v}(0)] \gtrless 0 \qquad \text{if } (1-a) = (1-a_i)(1-a_j).$$

There is a trade-off only if there is a cost of acquiring a good image: $v_i e - (c/\beta) < 0$. In either case, this cost is reduced by the uncertainty of being pivotal, and so the incentive to signal a pro-social inclination is higher. We call this the "cheap signaling effect": It is cheap to make sacrifices that may not materialize, i.e., reaping image gains at no or little cost.

Proposition 9 suggests an explanation for why *transparent committees*, i.e., committees for which individual voting records are publicly disclosed can be soft on third parties. This will be the case if the rules of committees are strict and decision makers would like to appear friendly to third parties. An example is the European council of finance ministers (EcoFin), which did not sanction any of the many violations of the Maastricht treaty: Why risk to attract the wrath of the infringing country when one may not even be pivotal to the decision?

Up to this point we have only discussed individual behavior, but it is also important to consider aggregate outcomes: In the case of collective veto power, the aggregate outcome is a fortiori more pro-social than in the case of a dictator. But even in contexts with two gatekeepers to pro-social behavior, i.e., individual veto power, it may be the case that the outcome is more pro-social in comparison to the dictator case (which would be impossible in the absence of image concerns).

6.2 Individual incentives

In many situations, the cost of moral behavior depends on one's individual action rather than on the collective action/outcome. In the following we therefore introduce the distinction between *team incentives* vs. *individual incentives.* As an example of the latter, think of jointly coming to the rescue of a victim, which may imply being shot at or knifed by the aggressor. Likewise, in organizations, an agent's cost c to enforce a pro-social outcome can be associated with a loss of material or other incentives, e.g., not receiving a bonus or experiencing pressure from a superior. Shared control with team incentives implies that all agents receive c if $a = 0$. In contrast, individual incentives mean that agent i receives c if $a_i = 0$, irrespective of the other agents' behavior. Agent i's utility under individual incentives is therefore

$$(v_i e)a - \left(\frac{c}{\beta}\right) a_i + \mu \hat{v}(a_i).$$

Proposition 10 *Under individual incentives, individual behavior is less pro-social than in the single decision-maker case.*

Proof. Assumption 1 implies that the immoral type never contributes regardless of what the other agent decides. So let us focus on the moral type. From Proposition 1, the moral type contributes if and only if

$$v\tilde{e} - \frac{c}{\beta} + \mu(v - \bar{v}) > 0,$$

where

$$\tilde{e} = \begin{cases} E[a_j]e & \text{if} \quad a = a_i a_j \\ [1 - E[a_j]]\, e & \text{if} \quad a = 1 - (1 - a_i)(1 - a_j). \end{cases}$$

In an equilibrium in which the moral type contributes therefore,

$$\tilde{e} = \begin{cases} \rho e & \text{if} \quad a = a_i a_j \\ (1 - \rho)e & \text{if} \quad a = 1 - (1 - a_i)(1 - a_j). \end{cases}$$

Either way, the condition for the existence of an equilibrium in which the moral type contributes is stricter than for a dictator. Suppose indeed that $e > e^*$, so a dictator contributes when moral. If $a = a_i a_j$, then if $\rho e < e^*$, the equilibrium involves no contribution. If $a = 1 - (1 - a_i)(1 - a_j)$, no contribution cannot be an equilibrium if $e > e^*$. The equilibrium is then in mixed strategies: The moral type contributes with probability x, where

$$(1 - \rho x)e = e^*,$$

so again individual behavior is less moral compared to the case of a single decision maker.

Note that even when one agent suffices to impose the moral outcome, collective behavior may be less pro-social than with a dictator (this is the case provided that e is not much higher than e^*). ■

Application:

Falk and Szech (2013) show that unethical individual behavior is more frequent with individual veto power (two gatekeepers) than with one. In their "bilateral trade" treatment, two parties can each, by refusing to trade, ensure that a negative externality associated with trading is avoided. Unethical behavior is more common than when a single party can pick an action with the same material payoff and the same externality as under trade in the bilateral trade treatment. Dana, Weber and Kuang (2007) show that comparing a standard dictator game with a simultaneous move game with two dictators, who can each implement a fair outcome vis-à-vis a receiver. They find that fair divisions are more likely in the former than in the latter. The case of collective veto power is studied in Falk and Szech (2014). They compare an individual decision making treatment with decision making in groups. In both conditions subjects (simultaneously) choose whether to kill a mouse in exchange for c, or not to kill. If only one subject decides to kill in the group treatment eight[42] mice are killed, irrespective of the other group members's decisions. Moral behavior is less frequent in groups than in the individual decision condition. They also report an effect of $E[a_j]e$ on the killing decision: the more likely an agent believes she is pivotal, the less likely she is to kill.

[42] The externality is eight mice in the group of eight treatment to keep the externality per individual identical to the individual condition (one subject, one mouse).

6.3 Collective accountability

Social image concerns promote pro-social behavior. With shared control, however, individual attribution is not guaranteed, and reputational effects can be easily blurred. In many contexts with collective decision making, the relevant audience does not observe agents' individual choices a_i *(individual accountability)*, but only the aggregate outcome a *(collective accountability)*. In the case of team incentives and collective accountability, agent i's utility is

$$\left(v_i e - \frac{c}{\beta}\right) a + \mu \hat{v}(a).$$

From a social image perspective it clearly makes a difference whether the *outcome* a or the individual *behavior* (a_i) is observed, i.e., whether reputation $\widehat{v}$ depends on a or a_i. Intuitively, collective accountability (only a is observed) allows hiding in a group, hoping that the audience will blame others for a bad outcome. Shared control creates a sub-additivity of responsibility and hence reduces incentives for moral behavior. This holds for individual veto power. With collective veto power the immoral outcome is in fact always an equilibrium, the other's immoral behavior supplying a perfect excuse!

Proposition 11 *Collective accountability makes the conditions for the existence of a moral equilibrium, in which the moral type contributes, more stringent, and the condition for the existence of an immoral equilibrium equally or less stringent. The Pareto-selection criterion implies that – relative to a single decision-maker – moral behavior is unchanged if* $a = a_1 a_2$ *and less frequent if* $1 - a = (1 - a_i)(1 - a_j)$.

In sum, shared control bears a potential to dilute responsibility and to encourage immoral behaviors. The precise circumstances matter, however. As we have seen we would not expect any negative effect for group members without image concerns, quite to the contrary. Moreover, organizational patterns of individual vs. collective veto power matter, and so do individual incentives and individual accountability. Organizations seeking to promote corporate social responsibility should therefore enforce a maximum of individual responsibility: This calls for carefully selecting employees on the basis of their image concerns, to make sure that individual actions are transparent, and to properly align individual incentives.

7 Concluding remarks

In this paper we have modeled the malleability of moral behavior in a coherent framework. Key issues that explain variation in morality are image concerns, self-control problems, narratives, imperatives and organizational design. We believe that the exercise is helpful in organizing important empirical findings, and to provide new insights into moral reasoning in general. At the same time, the analysis is clearly incomplete. One issue is the definition of what constitutes a moral act. In our analysis we have adopted the commonly accepted notion of avoiding and preventing harm to others, modeled in terms of an externality. We think that any model of morality must consider externalities, but acknowledge that other notions of morality may be important as well. Haidt (2007) for instance criticizes the reduction to the fairness-harm (externality) conception and suggests to include other phenomena such as loyalty, authority, and spiritual purity. A possible way to connect these notions to our framework is to interpret violations of loyalty or purity as creating a negative external effect among those who care for loyalty or purity; but this would be an incomplete account of the latter.

Another issue concerns agreement about the morally desired. The model starts from the premise that, for a given context, actors agree on what is right and wrong. This does not capture situations of multiple moral norms, characteristic of many important societal conflicts. Consider engagement in anti-abortion or anti-gay movements for example. For some people such engagement is considered a moral duty, for others it is the exact opposite. If agreement on what constitutes a positive externality is not commonly shared, however, it becomes unclear, e.g., how to achieve social image. In these cases our model implicitly assumes that we care for social image only among those who share our values, while disregarding other people's opinions and approval. A similar problem arises if a particular action simultaneously creates positive and negative externalities. Here, the model can be interpreted as considering the net effect, a utilitarian notion of treating as moral whatever creates the biggest net benefit. Finally, we have not explicitly addressed the important role of ex-post justifications of immoral action, e.g., in terms of wishful thinking or self-serving beliefs. This could be incorporated, however, as an extension to the section on narratives. The model would then specify the active search and/or self-construction of excusing narratives, after the decision has been taken. As before, the rationale would be the reinterpretation of circumstances and actions, in an attempt to preserve a good image.

References

Achtziger, A., Alós-Ferrer, C., and A. K. Wagner (2015) "Money, Depletion, and Prosociality in the Dictator Game," *Journal of Neuroscience, Psychology, and Economics*, 8(1): 1–14.

Alexander, L., and M. Moore (2015) "Deontological Ethics," *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/`

Aquino, K. and Reed II, A. (2002) "The self-importance of moral identity," *Journal of personality and social psychology*, 83(6): 1423–1440

Ariely, D., Bracha, A., and S. Meier (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1): 544–555.

Bandura, A. (1999) "Moral Disengagement in the Perpetration of Inhumanities," *Personality and Social Psychology Review*, 3(3): 193–209.

Baron, J., and I. Ritov (2004) "Omission Bias, Individual Differences, and Normality," *Organizational Behavior and Human Decision Processes*, 94(2): 74–85.

Bartling, B., Fehr, E., and D. Schunk (2012) "Health Effects on Children's Willingness to Compete," *Experimental Economics*, 15(1): 58–70.

Bartling, B., and U. Fischbacher (2012) "Shifting the Blame: On Delegation and Responsibility," *Review of Economic Studies*, 79(1): 67–87.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., and J.A. Strongman (1999) "Moral Hypocrisy: Appearing Moral to Oneself Without Being So," *Journal of Personality and Social Psychology*, 77(3): 525–537.

Baumeister, R. F., Vohs, K. D., and D. M. Tice (2007) "The Strength Model of Self-Control," *Current Directions in Psychological Science*, 16(6): 351–355.

Beaman, A. L., Klentz, B., Diener, E., and S. Svanum (1979) "Self-Awareness and Transgression in Children: Two Field Studies,"*Journal of Personality and Social Psychology*, 37(10): 1835–1846.

Bennett, W. (1993) *The Book of Virtues*, New York: Simon & Schuster.

Bentham, J. (1789) "An Introduction to the Principles of Morals," London: Athlone.

Bénabou, R., and J. Tirole (2002) "Self Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117: 871–915.

_______ (2004) "Willpower and Personal Rules," *Journal of Political Economy*, 112(4): 848–886.

_______ (2006) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5): 1652–1678.

——— (2009) “Over My Dead Body: Bargaining and the Price of Dignity,” *American Economic Review, Papers and Proceedings*, 99(2): 459–465.

——— (2011) “Identity, Morals and Taboos: Beliefs as Assets,” *Quarterly Journal of Economics*, 126(2): 805–855.

——— (2012) “Laws and Norms,” IZA Discussion Papers 6290, Institute for the Study of Labor (IZA).

Bisin, A., and T. Verdier (2001) “The Economics of Cultural Transmission and the Dynamics of Preferences,” *Journal of Economic Theory*, 97: 298–319.

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., and C. McKenna (2010) “Moral Credentialing by Association: The Importance of Choice and Relationship Closeness,” *Personality and Social Psychology Bulletin*, 36(11): 1564–1575.

Bruner, J. (1991) “Relational Contracts and the Nature of Market Interactions,” *Econometrica*, 72(3): 747–780.

Burks, S. V., Carpenter, J. P., Goette, L., and A. Rustichini (2009) “Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment,” *Proceedings of the National Academy of Sciences of the United States of America*, 106(19): 7745–7750.

Chen, D. and M. Schonger (2013) “Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations,” Discussion Paper, ETH Zurich.

Corgnet, B., Espin, A. M. and R. Hernán-González (2015) “The Cognitive Basis of Social Behavior: Cognitive Reflection Overrides Antisocial but Not Always Prosocial Motives,” *Frontiers in Behavioral Neuroscience*, 9: 289.

Cummiskey, D. (1996) *Kantian Consequentialism*. New York: Oxford University Press.

Dal Bó, E., and P. Dal Bó (2014) “‘Do the right thing’: The effects of moral suasion on cooperation,” *Journal of Public Economics*, 117: 28–38.

Dana, J., Weber, R., and J. Kuang (2007) “Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preferences for Fairness,” *Economic Theory*, 33: 67–80.

Darley, J. M. (1992) “Social Organization for the Production of Evil,” *Psychological Inquiry*, 3(2): 199–218.

Darley, J. M., and B. Latané (1968) “Bystander Intervention in Emergencies: Diffusion of Responsibility,” *Journal of Personality and Social Psychology*, 8 (4, Pt.1): 377–383.

Deckers, T., Falk, A., Kosse, F., and N. Szech. (2016) “Homo Moralis: Personal Characteristics, Institutions and Moral Decision-Making,” WZB Discussion Paper SPII 2016–302.

Dessi, R., and B. Monin (2012) “Noblesse Oblige: Moral Identity and Prosocial Behavior in the Face of Selfishness,” TSE Working Paper 12–347, Toulouse School of Economics.

Dewatripoint, M. and J. Tirole (2005) “Modes of Communication,” *Journal of Political Economy*, 113: 1217–1238.

Diener, E., and M. Wallbom (1976) “Effects of Self-Awareness on Antinormative Behavior,” *Journal of Research in Personality*, 10(1): 107–111.

Dohmen, T., Falk, A., Huffman, D. and U. Sunde (2012) “The Intergenerational Transmission of Risk and Trust Attitudes,” *Review of Economic Studies*, 79(2): 645-677.

Durkheim, E. (1915) *The Elementary Forms of the Religious Life*, London George Allen & Unwin ltd.

Duval, S. and R. A. Wicklund (1972) *A Theory of Objective Self Awareness*, Oxford: Academic Press.

Effron, D. A., Cameron, J. S., and B. Monin (2009) “Endorsing Obama Licenses Favoring Whites,” *Journal of Experimental Social Psychology*, 45(4): 590–593.

Effron, D. A., Monin, B., and D. T. Miller (2012) “The Unhealthy Road Not Taken: Licensing Indulgence by Exaggerating Counterfactual Sins,” *Journal of Experimental Social Psychology*, 49(3): 573–578.

Egas, M. and A.Riedl (2008) “The Economics of Altruistic Punishment and the Maintenance of Cooperation” *Proc. R. Soc. B*, 275: 871–878.

Engel, C. (2011) “Dictator Games: A Meta Study,” *Experimental Economics*, 14(4), 583–610.

Everett, J. A. C., Pizarro, D., and M. J. Crockett (in press) “Inference of Trustworthiness from Intuitive Moral Judgments,” *Journal of Experimental Psychology: General.*

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and U. Sunde (2015) “The Nature and Predictive Power of Preferences: Global Evidence,” IZA Discussion Paper No. 9504.

Falk, A. and N. Szech (2013) “Morals and Markets,” *Science*, 340: 707–711.

_______ (2014) “Diffusion of Being Pivotal and Immoral Outcomes,” Discussion Paper, University of Bonn.

_______ (2015) “Irrelevance of Numbers and Concave Moral Costs ” Discussion Paper, University of Bonn.

Falk, A. and J. Tirole (2016) “In Face of Yourself - A Note on Self-Image,” mimeo.

Feinberg, J. (1984) *Harm to others.* Oxford, Oxford University Press.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., and D. Frey (2011) “The Bystander-Effect: A Meta-Analytic Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies”, *Psychological Bulletin*, 137(4): 517–537.

Foot, P. (1967) “The Problem of Abortion and the Doctrine of the Double Effect,” *Oxford Review*, 5: 5–15.

Gambino, R. (1973) “Watergate lingo: A language of non-responsibility,” *Freedom at Issue*, 22(7–9): 15–17.

Gächter, S. and B. Herrmann (2009) “Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment,” *Phil. Trans. R. Soc*, 364: 791–806.

Gert, B. and J. Gert (2016) “The Definition of Morality,” *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/spr2016/entries/morality-definition/`

Gino, F., Schweitzer, M. E., Mead, N. L., and D. Ariely (2011) “Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior,” *Organizational Behavior and Human Decision Processes*, 115(2): 191–203.

Glaeser, E. L. (2005) “The Political Economy of Hatred,” *Quarterly Journal of Economics*, 120: 45–86.

Glick, P. (2002) “Sacrificial Lambs Dressed in Wolves’ Clothing: Envious Prejudice, Ideology, and the Scapegoating of Jews,” In Newman, L. S. and R. Erber *Understanding Genocide: The Social Psychology of the Holocaust.* New York, Oxford University Press.

Glover, J. (2012) *Humanity: A Moral History of the Twentieth Century* (2nd ed.). New Haven, Yale University Press.

Gneezy, U., Keenan, E. A., and A. Gneezy (2014) “Avoiding Overhead Aversion in Charity,” *Science*, 346(6209): 632–635.

Goeree, J. K., Holt, C. A., and S. K. Laury (2002) “Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior,” *Journal of Public Economics*, 83(2): 255–276.

Gottfredson, M. R. and T. Hirschi (1990) *A General Theory of Crime.* Stanford, Stanford University Press.

Haidt, J. (2001) “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment,” *Psychological Review*, 108(4): 814–834.

Haidt, J. (2007) “The new synthesis in moral psychology, ” *Science*, 316, 998–1002.

Haidt, J., Graham, J., and C. Joseph (2009) “Above and Below Left-Right: Ideological Narratives and Moral Foundations,” *Psychological Inquiry*, 20(2–3): 110–119.

Hamman, J., Loewenstein, G., and R. Weber (2010) “Self-interest Through Delegation: An Additional Rationale for the Principal-Agent Relationship,” *American Economic Review*, 100(4): 1826–1846.

Hare, R. M. (1993) “Could Kant Have Been a Utilitarian?” In Dancy, R. M. *Kant and Critique: New Essays in Honor of W.H. Werkmeister.* Dordrecht, Springer Science & Business Media.

Johnson, D. R. (2012) “Transportation into a story increases empathy, prosocial behavior, and perceptual bias toward fearful expressions,” *Personality and Individual Differences*, 52: 150–155.

Johnson, R. (2014) “Kant’s Moral Philosophy”, *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/sum2014/entries/kant-moral/`

Jordan, J., Mullen, E., and J. K. Murnighan (2011) “Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior,” *Personality and Social Psychology Bulletin*, 37(5): 701–713.

Kagel, J. H. and A. E. Roth (1995) *The Handbook of Experimental Economics.* Princeton, Princeton University Press.

Kahneman, D. and A. Tversky (1984) “Choices, Values, and Frames,” *American Psychologist*, 39(3): 341–350.

_______ (2000) *Choices, Values, and Frames.* Cambridge, Cambridge University Press.

Kant, I. (1785) *Grundlegung zur Metaphysik der Sitten.*

Kant, I. (1797) “On a Supposed Right to Lie from Philanthropy,” In Kant, I. and M. J. Gregor (1996) *Practical Philosophy*, Cambridge: Cambridge University Press.

Keeton, R. M. (2015) “‘The Race of Pale Men Should Increase and Multiply’: Religious Narratives and Indian Removal”In Presser, L. and S. Sandberg (2015) *Narrative Criminology: Understanding Stories of Crime*, New York and London: New York University Press.

Khan, U. and R. Dhar (2006) “Licensing Effect in Consumer Choice,” *Journal of Marketing Research*, 43(2): 259–266.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., and P. Brugger (2006) “Disruption of Right Prefrontal Cortex by Low-Frequency Repetitive Transcranial Magnetic Stimulation Induces Risk-Taking Behavior,” *Journal of Neuroscience*, 26(24): 6469–6472.

Lazear, E. P., Malmendier, U., and R. A. Weber (2012) “Sorting in Experiments with Application to Social Preferences.” *American Economic Journal: Applied Economics*, 4(1): 136–163.

Levi, P. (1988) *The Drowned and the Saved.* New York: Summit.

Lifton, R. J. (1986) *The Nazi Doctors: A Study of the Psychology of Evil.* London, Macmillan.

Mar, R. A. and K. Oatley (2008) "The Function of Fiction is the Abstraction and Simulation of Social Experience," *Perspectives on Psychological Science*, 3(3): 173–192.

Martinsson, P., Myrseth, K. O. R., and C. Wollbrant (2012) "Reconciling Pro-Social vs. Selfish Behavior: On the Role of Self-Control," *Judgment and Decision Making*, 7(3): 304–315.

Mazar, N., Amir, O and D. Ariely (2008) "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, XLV: 633–644.

Mazar, N. and C.-B. Zhong (2010) "Do Green Products Make Us Better People?," *Psychological Science*, 21(4): 494–498.

McAdams, D. P (1985) *Power, Intimacy, and the Life Story*, Homewood, IL: Dorsey.

_______ (2001) "The Psychology of Life Stories," *Review of General Psychology*, 5: 100–122.

_______ (2006) "The Role of Narrative in Personality Psychology Today," *Narrative Inquiry*, 16(1): 11–18.

McAdams, D. P and R. Koppensteiner (1992) "The Manager Seeking Virtue: Lessons from Literature," *Journal of Business Ethics*, 11(8): 627–634.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., and D. Ariely (2009) "Too Tired to Tell the Truth: Self-Control Resource Depletion and Dishonesty," *Journal of Experimental Social Psychology*, 45(3): 594–597.

Merritt, A. C., Effron, D. A., Fein, S., Savitsky, K. K., Tuller, D. M., and B. Monin (2012) "The Strategic Pursuit of Moral Credentials," *Journal of Experimental Social Psychology*, 48(3): 774–777.

Merritt, A. C., Effron, D. A., and B. Monin (2010) "Moral Self-Licensing: When Being Good Frees Us to Be Bad," *Social and Personality Psychology Compass*, 4(5): 344–357.

Mill, J. S. (2002) *Utilitarianism: edited with an introduction by Roger Crisp*, New York: Oxford University Press, Originally published in 1861.

Monin, B. and A. H. Jordan (2009) "The dynamic moral self: A social psychological perspective" In Narvaez, D. and D. Lapsley (eds.) (2009) *Personality, Identity, and Character: Explorations in Moral Psychology.* New York: Cambridge University Press.

Monin, B. and D. T. Miller (2001) "Moral Credentials and the Expression of Prejudice," *Journal of Personality and Social Psychology*, 81(1): 33–43.

Nikiforakis, N. and H.-T. Normann (2008) "A Comparative Statics Analysis of Punishment in Public-Good Experiments," *Exp Econ*, 11: 358–369.

Oberholzer-Gee, F. and R. Eichenberger (2008) "Fairness in Extended Dictator Game Experiments," *The BE Journal of Economic Analysis & Policy*, 8(1): 16.

Osgood, J. M. and M. Muraven (2015) "Self-Control Depletion Does Not Diminish Attitudes About Being Prosocial but Does Diminish Prosocial Behaviors," *Basic and Applied Social Psychology*, 37(1): 68–80.

Petrinovich, L. and P. O'Neill (1996) "Influence of Wording and Framing Effects on Moral Intuitions," *Ethology and Sociobiology*, 17(3): 145–171.

Presser, L. and S. Sandberg (2015) *Narrative Criminology: Understanding Stories of Crime*. New York and London: New York University Press.

Rand, D., Greene, J. D., and M. A. Nowak (2012) "Spontaneous Giving and Calculated Greed," *Nature*, 489(7416): 427–430.

Ritov, I. and J. Baron (1990) *Reluctance to Vaccinate: Omission Bias and Ambiguity.* Wharton School, Risk and Decision Processes Center.

Roth, A. E. (2007) "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, 21(3): 37–58.

Sachdeva, S., Iliev, R., and D. L. Medin (2009) "Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation," *Psychological Science*, 20(4): 523–528.

Sinnott-Armstrong, W. (2015) "Consequentialism", *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/win2015/entries/consequentialism/`

Spranca, M., Minsk, E., and J. Baron (1991) "Omission and Commission in Judgment and Choice," *Journal of Experimental Social Psychology*, 27(1): 76–105.

Sunstein, C. R. (2005) "Moral Heuristics," *Behavioral and Brain Sciences*, 28: 531–573.

Sykes, G. M. and D. Matza "Techniques of Neutralization: A Theory of Delinquency," *American Sociological Review*, 22(6): 664–670.

Tabellini, G. (2008) "The Scope of Cooperation: Values and Incentives," *Quarterly Journal of Economics*, 123(3): 905–950.

Tappan, M. and L. Brown (1989) "Stories Told and Lessons Learned: Toward a Narrative Approach to Moral Development and Moral Education," *Harvard Educational Review*, 59 (2): 182–205.

Vallacher, R. R. and M. Solodky (1979) "Objective Self-Awareness, Standards of Evaluation, and Moral Behavior," *Journal of Experimental Social Psychology*, 15(3): 254–262.

Vitz, P. C. (1990) "The Use of Stories in Moral Development," *American Psychologist*, 45:709–720.

Yanagizawa-Drott, D. (2014) "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics*, 129(4): 1947–1994.

Zhong, C.-B. and K. Liljenquist (2006) "Washing Away Your Sins: Threatened Morality and Physical Cleansing," *Science*, 313(5792): 1451–1452.

Zimbardo, P. (2007) *The Lucifer Effect: Understanding How Good People Turn Evil.* New York, Random House.

Appendix

Proof of Proposition 9 (cheap signaling effect)

We focus on symmetric equilibria, and assume that behavior is moral under a single decision-maker (otherwise the Proposition is trivially satisfied):

$$ve - \frac{c}{\beta} + \mu(v - \bar{v}) > 0$$

Collective veto power ($a = a_i a_j$). Assume first that

$$\rho\left(-\frac{c}{\beta}\right) + \mu v \leq 0,$$

which is a necessary condition for a separating equilibrium to exist. This separating equilibrium (with moral behavior by the high type) yields payoff for the high type equal to

$$\rho\left(ve - \frac{c}{\beta}\right) + \mu v.$$

Under the assumption that $\hat{v}(1) = v$, there is no pooling equilibrium at $a_i \equiv 0$ for all i. So the equilibrium is necessarily either a separating equilibrium or an equilibrium in which the high type randomizes between $a_i = 1$ and $a_i = 0$ (the condition $\rho(-c/\beta) + \mu v \leq 0$ imposes $a_i = 0$ for the low type). Hence $\hat{v}(0) \leq \bar{v}$: The payoff of the high type is bounded above by $\mu\bar{v}$ in any equilibrium in which the high type picks $a_i = 0$ with positive probability. Hence, the separating equilibrium is necessarily Pareto dominant if

$$e > e^{**},$$

where $e^{**} < e^*$ is given by

$$\rho\left(\frac{c}{\beta} - ve^{**}\right) = \mu(v - \bar{v}).$$

So the range of parameters leading to moral behavior is larger. Furthermore, there is some moral behavior even if $e < e^{**}$.

When $\rho(-c/\beta) + \mu v > 0$ by contrast, an equilibrium in which there is some moral behavior has the low type play $a_i = 1$ with probability x and $a_i \equiv 0$ with probability $1 - x$. Letting $\hat{\rho} \equiv \rho + (1-\rho)x$ and $\hat{v} \equiv \rho v/\hat{\rho}$, the low type's indifference writes:

$$\hat{\rho}\left(-\frac{c}{\beta}\right) + \mu\hat{v} = 0.$$

This defines a unique $\hat{\rho}$ and therefore a unique mixing probability x. Behavior is therefore more moral as the high type, and sometimes the low type behave morally.

Individual veto power $(1-a=(1-a_i)(1-a_j))$. Under individual veto power, immoral behavior by both types is an equilibrium provided that

$$ve-\frac{c}{\beta}+\mu v \leq \mu \bar{v}.$$

So immoral behavior is not an equilibrium under the assumption made earlier.

As earlier, we can look for an equilibrium in which the high type behaves morally and the low type may or may not play morally.

A separating equilibrium requires that

$$(1-\rho)\left(ve-\frac{c}{\beta}\right)+\mu v>0 \text{ and } (1-\rho)\left(-\frac{c}{\beta}\right)+\mu v<0.$$

The first condition is always satisfied. A pooling equilibrium (both types behave morally) always exists (if a_j is always equal to 1, agent i is never pivotal and might as well invest in her image). Finally, randomization by the low type requires that

$$(1-\hat{\rho})\left(-\frac{c}{\beta}\right)+\mu\hat{v}=0$$

with $\hat{\rho}\equiv\rho+(1-\rho)x$ and $\hat{v}=\rho v/\hat{\rho}$, and x is the low type's probability of behaving morally.

When $(1-\rho)(-c/\beta)+\mu v<0$, the Pareto-dominant equilibrium is the separating one:

$$\rho\left(-\frac{c}{\beta}\right)>\left(-\frac{c}{\beta}\right)+\mu\bar{v} \qquad \text{for the low type}$$

and

$$ve-\frac{c}{\beta}+\mu v>ve-\frac{c}{\beta}+\mu\bar{v} \quad \text{for the high type}$$

When $(1-\rho)\left(-\frac{c}{\beta}\right)+\mu v>0$, the Pareto-dominant equilibrium is the semi-separating equilibrium:

$$\hat{\rho}\left(-\frac{c}{\beta}\right)>\left(-\frac{c}{\beta}\right)+\mu\bar{v} \qquad \text{for the low type}$$

and

$$ve-\frac{c}{\beta}+\mu\hat{v}>ve-\frac{c}{\beta}+\mu\bar{v} \quad \text{for the high type}$$