NBER WORKING PAPER SERIES

SELF-ESTEEM, MORAL CAPITAL, AND WRONGDOING

Ernesto Dal Bó
Marko Terviö

Working Paper 14508
http://www.nber.org/papers/w14508

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2008

We thank Roland Bénabou, Jeremy Bulow, Pedro Dal Bó, Erik Eyster, Botond Köszegi, Keith Krehbiel, John Morgan, Santiago Oliveros, Matt Rabin, Tim Williamson, and seminar participants at Berkeley, Birming- ham, Essex, Helsinki School of Economics, Princeton, Stanford, UCSD, and Universidad de San Andrés for useful conversations and comments. Juan Escobar provided excellent research assistance. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Self-Esteem, Moral Capital, and Wrongdoing
Ernesto Dal Bó and Marko Terviö
NBER Working Paper No. 14508
November 2008, Revised September 2009
JEL No. D83,K4,Z1

ABSTRACT

We present an infinite-horizon model of moral standards where self-esteem and unconscious drives play key roles. In the model, an individual receives random temptations (such as bribe offers) and must decide which to resist. Individual actions depend both on conscious intent and a type reflecting unconscious drives. Temptations yield consumption value, but keeping a good self-image (a high belief of being the type of person that resists) yields self-esteem. We identify conditions for individuals to build an introspective reputation for goodness ("moral capital") and for good actions to lead to a stronger disposition to do good. Bad actions destroy moral capital and lock-in further wrongdoing. Economic shocks that result in higher temptations have persistent effects on wrongdoing that fade only as new generations replace the shocked cohorts. Small parametric differences across societies may lead to large wrongdoing differentials, and societies with the same moral fundamentals may display different wrongdoing rates depending on how much past luck has polarized the distribution of individual beliefs. The model illustrates how optimal deterrence may change under endogenous moral costs and how wrongdoing may be compounded as high temptation activities attract individuals with low moral capital.

Ernesto Dal Bó
University of California, Berkeley
Haas School of Business
545 Student Services Building #1900
Berkeley, CA 94720-1900
and NBER
dalbo@haas.berkeley.edu

Marko Terviö
Helsinki School of Economics
PL 1210
00101 Helsinki
Finland
marko.tervio@hse.fi

1 Introduction

We investigate the dynamics of wrongdoing in a model where individual moral standards emerge endogenously. We develop our framework in two parts. In the first part we investigate individual-level incentives to adhere to a moral objective, and in the second part we aggregate behavior to the level of a society in a demographic steady state. In our model, an infinitely lived individual receives a sequence of stochastic consumption opportunities ("temptations") and must decide which to resist, if any. The model is rooted on two fundamental assumptions. The first is that individual actions when facing temptation depend not just on conscious intentions but also on unconscious impulses. The second is that the individual, although aware that temptations are enjoyable, also derives utility from thinking she has "a good heart." In other words, she would like to be the type of person whose unconscious nature is geared towards rejecting temptations.

The idea that people may have an incentive to behave morally because they want to maintain a high opinion of themselves is old – it goes back at least to Adam Smith's Theory of Moral Sentiments. But it poses immediate challenges. How are opinions on the self anchored, and when will individuals prefer to forgo enjoyable consumption for the sake of self-image? After all, a life of mischief may be rewarding, too. Second, will behavior that is driven by an introspective reputation be self-reinforcing? Third, what are the implications for the social dynamics of wrongdoing when individuals balance self-esteem with consumption-based utility?

We derive the individual's optimal policy and isolate conditions under which (i) individuals resist actions that are deemed immoral but yield consumption value, (ii) individuals improve their self-image by resisting immoral actions, and (iii) an improvement in self-image strengthens the inclination to resist immoral actions, while events that damage self-image weaken the inclination to act morally. A self-reinforcing path of wrongdoing results. An example is that of a person who, perhaps because the country is going through hard times, faces a surge in temptations. Under hardship, a person may do things that erode her self-image, such as taking a bribe. A damaged self-image reduces the incentive to behave morally, even after economic conditions have returned to normality. These results require specific conditions, both in terms of the preferences over self-image as well as in terms of what actions are more likely to reveal information about the self.

In the second part of the paper we derive the aggregate wrongdoing rate in a society in demographic steady state, and then perform comparative statics and impulse-response type

exercises. For instance, a higher distribution of temptations yields more wrongdoing not just because temptations are in average higher, but because it triggers an endogenous decrease in individual moral standards. This result highlights how small differences in fundamentals across societies may create relatively large differences in wrongdoing rates. We also show how shocks that trigger wrongdoing during a "crisis" period will continue to raise wrongdoing rates well after the economy has got back to normal. In addition, wrongdoing across societies may not respond solely to "moral fundamentals" such as the share of "good" vs. "bad" people, but to events in the past that have polarized the beliefs that individuals hold about themselves. Our model yields conditions for the emergence of phenomena such as taboos and the use of harsher punishments for repeat offenders. The latter result illustrates the important point that optimal deterrence schemes may shift when moral standards are endogenous. We also explain why "high temptation" activities (such as politics) could attract individuals with low moral standards, making such activities conducive to high wrongdoing not just through their stronger temptations but also because they attract the individuals least equipped to resist them.

It is important to note that we do not attempt to explain the content of morality. We take it as given that individuals believe that utilizing certain consumption opportunities is wrong, and that goodness is the feature of types who do not do wrong. The content of morality may follow from evolutionary forces, and be transmitted by culture and parental authority. The question of why people derive utility from thinking that they are good and what counts as wrongdoing is beyond the scope of our enquiry. We study the determination of moral standards, seen as the degree of adherence to established moral principles. This is important because there is indication that moral standards are both endogenous and important for behavior.

First, there is evidence that intrinsic motivations, and in particular notions of what is right and what constitutes a duty, can be important determinants of behavior. For example, experimental evidence indicates that people are willing to give up consumption in exchange for avoiding telling lies (Gneezy 2005), and for imparting justice in the form of punishment against those who "misbehave" (Ostrom, Walker and Gardner 1992, Fehr and Gächter 2002).[1] Second, there is a revealed preference argument for the idea that moral costs are

[1] Fisman and Miguel (2007) show that traffic violations correlate with national origin even when individuals share the same environment. One interpretation is that intrinsic aspects of motivation that are culturally defined affect dispositions towards wrongdoing. Also, considerations of fairness appear to vary across cultures, and affect how individuals distribute resources (Heinrich and Smith 2004).

both important and predictably sensitive to intervention. Nontrivial amounts of resources are spent with the objective of shaping moral costs. Parental discourse toward children, and expenditures in education (from the elementary level to MBA Ethics courses) are arguably serving the purpose of having individuals internalize moral standards. However, many models in economics and politics studying wrongdoing (crime, tax evasion, corruption) tend to consider "moral costs" a given.[2] As will be illustrated by our model, the optimal design of deterrence mechanisms may change once we incorporate the fact that moral standards are endogenous.

The structure of the paper is as follows. The next section offers an overview of our model and of the related literature. Section 3 presents our basic model featuring the problem of an individual. Section 4 aggregates the problem of individuals and studies determinants of wrongdoing rates at the social level. Section 5 provides applications and Section 6 concludes.

2 Overview of the model and related literature

In our model, infinitely-lived individuals receive an independently and randomly drawn consumption opportunity, or temptation, in each period. Individuals may resist the temptation or succumb to it. Individuals have time-consistent preferences and an unconscious type that may be good or bad. Good types always adhere to the moral principle of resisting temptations. Individuals do not know their type, but hold beliefs about the probability that they are good. These beliefs constitute individuals' self-image, an introspective reputation that gets tarnished when deviating from what good types would do. Similarly to Kőszegi (2006), we assume that individuals derive utility not just from consumption but also from self-image and that they may be risk averse with respect to that self-image.

In a standard model where intent to resist translates always into actual resistance, every individual could behave as a good type, but then resistance could not be taken as evidence that one has the good type. In our model, however, bad individuals cannot be certain to behave well even when they try, so good behavior does convey information about a person's type. This is due to an important aspect of our model, namely that individuals have limited control over their actions. In our model individuals choose whether they intend to actively resist the temptation, or to give up. Giving up leads to one's type determining one's action, while a bad type could fall for the temptation even when trying to resist. As a consequence,

[2] A classic reference in the economic theory of crime is Becker (1968). His model (and much subsequent work) posits an exogenous parameter for an individual's inherent disposition to commit crimes.

the individual can infer something about her type after observing her own actions: after all, her actions may have been affected not by her intent, of which she is conscious, but by her type, of which she is not.

This framework helps think about the Weberian account of the Calvinist Ethic, according to which individuals are born saved or damned, but do not know their predestination status. Given that uncertainty, the account goes, individuals resist mundane temptations in order to reduce their fear that they were born damned. Thus individuals resist temptations in order to maintain and even improve their self-image. An immediate question is: how can the Weberian Calvinist improve her own confidence of having been born saved when her good actions were deliberately chosen to convince herself that she is saved?[3] In our model, the reason why confidence on having a good type can improve after good behavior is that the individual does not select her actions exclusively through conscious deliberation. Rather, she can only select her intent, and in each period she may suffer an independent random disturbance allowing her type to "defeat" her intent and determine the action. This lack of perfect correlation between intent and actions is dubbed "imperfect free will" throughout the paper.

Is it reasonable to assume that individuals are not in full control of their own actions? A large literature in psychology has documented the role of visceral impulses and unconscious bias in decision-making. For example, Berridge (2003) discusses how the mesolimbic dopamine system causes ultimate decisions to reflect unconscious drives, thereby introducing a wedge between what we 'like' (or what we would like to want) and what we actually 'want.' (For a previous model where decisions are affected by unconscious 'gut' feelings, see Prelec and Bodner 2003. On visceral impulses see also, i.a., Loewenstein 1996, and Bernheim and Rangel 2004 for a model of addiction rooted in the neuroscience of impulse control). The permanent nature of unconscious drives is what is captured in one's type, while period-specific, random factors alter the strength of those drives. The view in the paper is that people may still select an intent that could override, and generate good news about, the nature of those drives. People presumably care about being the type who "misbehaves" precisely because society condemns such misbehavior. This is consistent with the Calvinist view that what matters is one's predestination status, and that human actions only count to the extent that they convey information about that status.[4]

[3]This question has been studied by Prelec and Bodner (2003) and Bénabou and Tirole (2004), whose work we discuss below.

[4]This does not imply that one could not endorse a more benign ethical view where what counts is not

An example of actions being affected by forces outside of conscious designs is when the ability to control a visceral impulse is diminished by a shock to external circumstances or even to an internal organic disposition. This mechanism, though substantively different, works in a way that is similar to forgetfulness in Bénabou and Tirole (2004). In their model, individuals forget their past actions with some probability and therefore learn from outcomes, despite understanding that they had acted under an incentive to manipulate their own beliefs. In our model, the imperfect free will associated with unconscious drives is the condition for good behavior to improve self-image.

The model also characterizes conditions under which a stronger self-image (i.e., a higher posterior on one having the good type) leads to a stronger disposition to select an intent to resist, and hence to actually be more likely to resist temptation. The resulting pattern is one where self-image is costly to improve (it requires forgoing temptations) but improvements have lasting benefits, so self-image works as a form of capital, which we call moral capital. Under the individual's optimal policy, morality emerges as a cumulative process of habituation through action, which parallels Aristotle's account of the attainment of virtue.

A number of recent papers offer insights that help understand the shaping of moral standards. Kaplow and Shavell (2007) focus on the relative convenience of investing in instilling guilt and virtue versus using incentives to induce good behavior. Tabellini (2007) studies investments in the transmission of cooperative values in an overlapping generations framework. In a related model, Baron (2008) investigates different social arrangements for ensuring high levels of cooperation and compares the attractiveness of generalized vs. restricted morality. These studies address important aspects of moral behavior, but abstract from the internal process that makes individuals want to adhere to received moral objectives. In all of these models adherence to values responds directly to a given investment in their inculcation. Our model illustrates that although moral objectives might be internalized, inculcation should also target the determinants of the degree of adherence to those objectives, such as beliefs about one's goodness and one's ability to transform intent into actions.

Work on cognitive dissonance has emphasized the link between self-image and belief manipulation. In this connection, Rabin (1994) relies explicitly on a link between self-image and moral behavior, as do Brekke, Kverndokk and Nyborg (2003) in their model of voluntary contributions, as well as Cervellati, Esteban and Kranich (2006) in their model

one's type, which is after all a given, but one's attempts at dealing with it, which are the result of a choice. From a positive point of view, the model seeks to capture the regularity that people appear to value having a "good nature."

of moral sentiments and redistribution. The operationalization of self-image in those papers is very different from ours, which follows Kőszegi's (2006) formulation of ego-utility in his study of overconfidence and task choice. In his model, agents who are risk-averse about their beliefs about their own ability will choose tasks that are less informative. The demand for information about self also plays a crucial role in our model: when good behavior is relatively uninformative about one's type, risk-averse individuals will be willing to forgo temptations in order to preserve their introspective reputation, causing self-restraint to emerge. Risk aversion is necessary for individuals to be willing to sacrifice consumption for the sake of self-esteem because, beliefs being a martingale, no individual would sacrifice consumption for no expected improvement in her self-image.

Besides having a different focus–Kőszegi is concerned with overconfidence and not moral standards–a difference between our model and his is that in his setup individuals have uncertainty about their ability, a trait which increases extrinsic payoffs, while in our setup uncertainty relates to the moral type, a trait that decreases extrinsic payoffs. This difference leads to divergences in results. For instance, in Kőszegi's proposition 4, more confidence on having the high type makes taking a risky action more likely, while in our setup more confidence of having the good type leads to more resistance, which is akin to taking the less risky action. That is, the correlation between confidence and the risk taken in our paper is the opposite than in Kőszegi's model.

One implication of the martingale property of beliefs is that when individuals' priors match objective probabilities they cannot affect their moral capital on average. Then wrongdoing rates at the social level will depend on the dispersion of individual moral capital. This resonates with the early findings by Carrillo and Mariotti (2000) who study a model where an individual manipulates her beliefs in order to prevent dynamically inconsistent behavior. They note that beliefs cannot be manipulated in expectation, but higher moments can be.

In our paper, the individual wishes to manipulate her beliefs for purely intrinsic reasons. The models by Prelec and Bodner (2003), Bénabou and Tirole (2006, 2007), and Kőszegi (2006) can also be understood this way. Other work considers instrumental rationales for manipulating beliefs. In Carrillo and Mariotti (2000) and in Bénabou and Tirole (2004) the individual manipulates her beliefs in order to help herself overcome time-inconsistency.[5]

[5]Beliefs are manipulated for instrumental reasons also in the model of Compte and Postlewaite (2004), in which an individual wants to stay optimistic because such psychological state will improve her performance at a given task. Hermalin and Isen (2008) offer a model the where mood affects the choice of actions and vice versa, leading to potential multiple equilibria in individual behavior.

Bénabou and Tirole introduce several aspects that we revisit, such as self-reputation playing a role, and past actions of the individual having the power to affect that reputation, thereby opening the door to self-reinforcing patterns of behavior.[6] But the setups have important differences, notably that the individual in their model has time-inconsistent preferences, and is modeled as a sequence of selves who play a noncooperative game (our setup is decision-theoretic).[7] In addition, unconscious forces play a central role in our model.

In this regard our model lies closer to Prelec and Bodner's (2003) where the 'gut' makes decisions under the constraint that the conscious mind may disapprove of the gut's tendencies. They study a self-signaling game in which the gut makes a decision with an eye to concealing its own nature as evidenced by the decisions made. In Prelec and Bodner's model the 'gut' can be seen as a fully strategic player, while in our model unconscious forces are just 'firing away' (like behavioral types in reputation models), and the individual, not fully aware of their nature, may do well or badly at overriding them. We believe this is an attractive modeling choice to capture unconscious impulses.

To summarize, most previous work contains one or more of the following traits: individuals have time-inconsistent preferences; the individual is conceptualized as a sequence of different selves who play a non-cooperative game amongst themselves; models are static or have finite horizons; unconscious bias, when modelled, acts as a strategic player. Our model features an individual that contains a single self, uses Bayes' rule to update beliefs, and has time-consistent plans. The individual has an unconscious bias and a preference for feeling confident that such a bias is compatible with received morality. We characterize the full dynamics of individual behavior over an infinite horizon. This is convenient for our analysis of the accumulation of moral capital, as finite horizon settings confound the effects of a state variable that evolves over time (beliefs about self) with those of a terminal date.

3 The Model

The individual lives in an infinite horizon discrete time world and discounts the future by a factor $\lambda \in (0,1)$. The individual is characterized by a type, good or bad, that is unknown to her, and she is born with an initial belief that she is good with probability μ_0. She has two

[6] See also Tirole (1996) for a theory of corruption persistence based on the impact of stereotyping on extrinsic incentives.

[7] Prelec and Bodner (2003), Brocas and Carrillo (2005) and Fudenberg and Levine (2006) also adopt a non-cooperative approach to modeling intra-personal conflict. The latter model can be expressed in decision-theoretic terms, although it abstracts from self-image considerations.

additively separable sources of utility: "self-esteem," which depends on her belief that she is good, and consumption. What matters for our purposes is the additional consumption that the individual could gain by dishonest means. We call this additional consumption utility a "temptation."

In each period t the individual faces a temptation x_t, drawn randomly from nonnegative numbers according to a distribution function F, with associated density f. We assume that F is continuous, $f(0) > 0$, and $Ex < \infty$. For concreteness, think of a bureaucrat facing an opportunity for taking a bribe each period. The temptation is the additional consumption utility obtained by consuming the bribe.

Given the lack of restrictions on the shape of F, we can assume without loss of generality that utility is linear in x. To see what our reduced-form temptation x means, denote the consumption utility function $v(\cdot)$, the consumption available by honest means by c_h, and the additional consumption available by dishonest means by c_w. Then $x \equiv v(c_h + c_w) - v(c_h)$ measures the additional utility from the bribe that is tempting the individual. For example, a period when c_h is lower—say because an inflationary shock lowers real wages in the public sector—results in a higher x due to concave v. A shift in the distribution F towards higher values of x reflects an environment where wrongdoing opportunities are relatively more attractive.

An individual can take one of two actions in a given period: yield to the temptation or resist. However, the individual cannot select her action directly, but rather can select her intent. We will talk of "positive intent" when the individual is actively attempting to resist temptation, and of "no intent" or "giving up" when the individual is not trying. When selecting a positive intent, a bad individual will in fact resist the temptation only if her free will works in that period, which it does with probability $\phi \in (0, 1)$ (drawn independently each period). When free will works then intent determines the action, and when free will fails then the underlying type determines the action. This formulation separates an agent's intentions from her actions. One interpretation of imperfect free will is that an external shock alters the ability of the individual to transform her intent into her action. Another possibility is that of an internal shock, as humans have biological and subconscious impulses that may thwart the designs of conscious thought. The role of imperfect free will in our model is that actions may reflect not just the agent's intention, but also her type. As a result, the agent may learn about her type by observing her own actions. Note that in a world without free will there would be no choice. And in a world with perfect free will ($\phi = 1$) it would be impossible to learn anything about one's own type by looking at one's

own actions. When there are limitations on free will then self-discovery will have a role.

The individual can consciously perceive utility from temptations, and utility from self-esteem. The individual with belief μ_t in period t enjoys a self-esteem $u(\mu_t)$ during that period. We assume that

$$u(\mu) = \mu^{1-\rho}, \tag{1}$$

where $\rho \in [0, 1)$ is the coefficient of relative risk aversion. Preferences over beliefs are not standard in economics, but can be rationalized on the basis of psychological evidence that people care about their own attributes for non-instrumental reasons – that is, for reasons that are not connected to outcomes, but to the experience of living with a certain degree of self-worth.

Conditional on t, individual beliefs can only take one of three values, $\mu_t = \{0, \hat{\mu}_t, 1\}$. We call individuals with a belief $\hat{\mu}_t \in (0, 1)$ *unaware*, while those who know their type for sure, $\hat{\mu}_t \in \{0, 1\}$, are called *aware*. An unaware person who enters period t with beliefs $\hat{\mu}_{t-1}$ and who successfully resists a temptation in period t will, applying Bayes' rule, update her belief to $\hat{\mu}_t = \hat{\mu}_{t-1} / \left(\hat{\mu}_{t-1} + \left(1 - \hat{\mu}_{t-1}\right)\phi\right)$. Thus, having been born with the initial belief μ_0, an individual who has successfully resisted t times remains unaware and has the belief

$$\hat{\mu}_t = \frac{\mu_0}{\mu_0 + (1 - \mu_0)\,\phi^t}. \tag{2}$$

The beliefs about one's goodness improve when seeing oneself do good, even when knowing that one has selected a positive intent. Note that, in any given period, the individual obtains utility $U_t = x_t + u(0) = x_t$ if taking the temptation, or $U_t = u(\mu_t)$ if never having taken one. Figure 1 shows the timeline in any given period t.

Our formulation of types and free will can be rationalized in an expanded setting following Bernheim and Rangel (2004). They model individual actions as being automatically triggered whenever a level of sensitivity to a cue surpasses some threshold level. Building on their premise, now assume that the realized level of sensitivity at a point in time depends additively on a baseline, permanent level, and a temporary sensitivity disturbance.[8] Individuals differ in their baseline sensitivity. "Good types" have a very low baseline sensitivity, while "bad types" have a very high baseline sensitivity. If the baseline sensitivity of good (bad) types is low (high) enough relative to the extent of the support of temporary factors, we will obtain good (bad) types that always (never) resist. The workings of a positive intent can then be rationalized as raising the threshold for falling into temptation, so that bad types

[8] The additive formulation parallels the approach in Prelec and Bodner (2003).

with favorable temporary shocks will get an overall realized sensitivity below the threshold and resist. Then ϕ can be seen to capture the measure of temporary disturbances that, under positive intent, would bring the realized sensitivity of bad types below the threshold. This representation is clearly a simplification of the unconscious biological basis of behavior. However, it is related to views in neuroendocrinology of how hormones may affect behavior through an organizing and a situational impact (e.g., in connection with aggression and sexual differentiation, see i.a. Hays 1981 and Sussman et al. 1987). The organizing effect occurs before birth and in the first few years of life, shaping the central nervous system and fixing the baseline sensitivity. The situational impact is related to hormonal changes due to circumstantial shifts, providing the changing disturbance that completes the determination of realized sensitivity.

3.1 The individual's objective

The problem of the agent is to select a policy $\hat{x}_1, \hat{x}_2, \hat{x}_3, \ldots$ to maximize expected lifetime utility. The policy specifies cutoff values such that temptations above them will be met with a positive intent to avoid them. For now we assume that the optimal policy will take such a cutoff form. We check later that this assumption is verified.

To set up the expected lifetime utility as a function of the cutoffs, it is useful to consider first the contribution of just one generic future period t, as perceived before the realization of x_1. (Later we combine these contributions into the present value of expected utility.) At the end of period t the agent could be in four different states in terms of the expected utility contributed by period t: (i) she could remain unaware about her type, (ii) she could have found out she has a good type, (iii) she could have found out in period t that she has the bad type, (iv) she could have found in a period previous to t that she has the bad type. To calculate the probability for each of the states we introduce the following

Definition 1 *The term*

$$\begin{aligned} H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right) &\equiv \prod_{s=1}^{t} F\left(\hat{x}_s\right), \\ H_0 &\equiv 1, \end{aligned} \tag{3}$$

denotes the probability that the agent has received shocks that she meets with positive intent in all periods up to, and including, t.

An agent who is aware of being good will enjoy the self-esteem rewards of her certainty, with value $u(1)$. Someone who ends period t unaware of her type is someone who has

not yet fallen for a temptation and who has beliefs $\hat{\mu}_t \in (0,1)$ that she is good. Her utility will be $u(\hat{\mu}_t)$. Conditional on being good (which has prior probability μ_0), the two relevant states have respective probabilities $\Pr(\text{unaware}|\hat{x}_1,\ldots,\hat{x}_t) = H_t(\hat{x}_1,\ldots,\hat{x}_t)$ and $\Pr(\text{aware}|\hat{x}_1,\ldots,\hat{x}_t) = 1 - H_t(\hat{x}_1,\ldots,\hat{x}_t)$. Combining these probabilities with the respective conditional utilities, the contribution to the expected utility of a good type from future period t is,

$$EU_t|\text{good} = H_t(\hat{x}_1,\ldots,\hat{x}_t)\,u(\hat{\mu}_t) + [1 - H_t(\hat{x}_1,\ldots,\hat{x}_t)]\,u(1).$$

Someone who had already learned that she has the bad type before period t will enter the period with no self-esteem, $u(0) = 0$, and will take any temptation x_t. Her expected utility is just Ex. In the event that she finds out in period t that she is bad she faces different expected utilities depending on more specific circumstances. One possibility is that she faces a temptation above her cutoff $\hat{x}_t$, does not attempt to resist, and sees herself seize the temptation. This provides full evidence that she is bad, so $u(0) = 0$, and the expected consumption utility conditional on this event is $E[x|x \geq \hat{x}_t]$. But it could also be that the agent faces a temptation below $\hat{x}_t$, selects a positive intent, but lacks free will. Her bad type chooses the action for her, providing full evidence of being bad. Conditional on this instance the expected utility is $E[x|x < \hat{x}_t]$. Conditional on being bad, at the end of period t the agent may remain unaware or be in one of the three awareness states just described. The four possible awareness states have probabilities given by

$$\begin{aligned}
&\Pr(\text{unaware}|\hat{x}_1,\ldots,\hat{x}_t) = \phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t),\\
&\Pr(\text{aware before}|\hat{x}_1,\ldots,\hat{x}_t) = \left[1 - \phi^{t-1} H_{t-1}(\hat{x}_1,\ldots,\hat{x}_{t-1})\right],\\
&\Pr(\text{newly aware, high } x|\hat{x}_1,\ldots,\hat{x}_t) = \phi^{t-1} H_{t-1}(\hat{x}_1,\ldots,\hat{x}_t)\,[1 - F(\hat{x}_t)],\\
&\Pr(\text{newly aware, low } x|\hat{x}_1,\ldots,\hat{x}_t) = \phi^{t-1} H_{t-1}(\hat{x}_1,\ldots,\hat{x}_t)\,[F(\hat{x}_t)(1-\phi)].
\end{aligned}$$

Combining these probabilities with the respective expected utilities (suppressing the arguments of H_t for brevity) yields an expression for the expected utility accruing to a bad type from some future period t:

$$EU_t|\text{bad} = \begin{pmatrix} \phi^t H_t u(\hat{\mu}_t) + \left[1 - \phi^{t-1} H_{t-1}\right] Ex + \\ \phi^{t-1} H_{t-1}(1 - F(\hat{x}_t))\,E[x|x \geq \hat{x}_t] + (1-\phi)\,\phi^{t-1} H_{t-1} F(\hat{x}_t)\,E[x|x < \hat{x}_t] \end{pmatrix}.$$

Because at the beginning of period 1 the agent attaches probability μ_0 to being good, her expected utility from period t is

$$EU_t = \mu_0\left[H_t u(\mu_t) + (1 - H_t)\,u(1)\right] + (1-\mu_0)\,EU_t|\text{bad}. \tag{4}$$

The sequence of utilities conditional on remaining unaware, $u(\hat{\mu}_1), u(\hat{\mu}_2), \ldots$, is just a known increasing sequence of numbers that converges to $u(1)$, hence we denote these numbers as u_t. Summing up and discounting the expected utilities (4) from all periods $t = 1, 2, \ldots$ gives (after rearrangement) the individual objective function

$$\begin{aligned} V_0(\hat{x}_1, \hat{x}_2, \ldots) &= \sum_{t=1}^{\infty} \lambda^{t-1} EU_t = \frac{\mu_0 u(1) + (1-\mu_0) Ex}{1-\lambda} + \\ &+ \sum_{t=1}^{\infty} \lambda^{t-1} \left\{ \begin{array}{c} \left[\mu_0 + (1-\mu_0)\phi^t\right] H_t u_t \\ -\mu_0 H_t u(1) - (1-\mu_0)\phi^t H_{t-1} \int_0^{\hat{x}_t} x f(x) dx \end{array} \right\}. \end{aligned} \tag{5}$$

3.2 Optimal policy

The problem of the individual is to select a sequence of cutoffs $\hat{x}_1, \hat{x}_2, \ldots$ to maximize the objective function (5). The cutoff $\hat{x}_t$ gives the highest temptation that she will intend to resist in period t conditional on remaining unaware as of the beginning of period t. (If she is aware of her type in period t there is nothing to choose; good types are unable to do bad, and bad types get zero utility from self-esteem so they take every temptation). The first order condition with respect to the cutoff in an arbitrary period s is,

$$\begin{aligned} &\frac{\partial V_0}{\partial \hat{x}_s} = \lambda^{s-1} H_{s-1} f(\hat{x}_s) \left\{ u_s \left[\mu_0 + (1-\mu_0)\phi^s\right] - \mu_0 u(1) - (1-\mu_0)\phi^s \hat{x}_s \right\} \\ &+ \frac{f(\hat{x}_s)}{F(\hat{x}_s)} \sum_{t=s}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F(\hat{x}_{t+1}) \left[\mu_0 + (1-\mu_0)\phi^{t+1}\right] u_{t+1} - \\ -\mu_0 F(\hat{x}_{t+1}) u(1) - (1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} = 0. \end{aligned} \tag{6}$$

Substantially rearranging this condition yields the extremum

$$\hat{x}_s^* = \frac{g_s}{(1-\mu_0)\phi^s} + \sum_{t=1}^{\infty} \lambda^{t+s-1} \frac{H_{t+s-1}}{H_s} \left\{ \frac{F(\hat{x}_{t+s}) g_{t+s}}{(1-\mu_0)\phi^s} - \phi^t \int_0^{\hat{x}_{t+s}} x f(x) dx \right\}, \tag{7}$$

where $g_s \equiv \left[\mu_0 + (1-\mu_0)\phi^s\right] u_s - \mu_0 u(1)$.

This last expression (7) characterizes a sequence $\hat{x}_1^*, \hat{x}_2^*, \ldots$ of solutions to the problem where each threshold is a function of future (but not past) policies. (The optimal policy is thus time-consistent). Note that $H_{t+s-1}/H_s = F(\hat{x}_{s+1}) \times \cdots \times F(\hat{x}_{t+s-1})$. Using the generic expression for $\hat{x}_s^*$, we then obtain the particular case of $\hat{x}_1^*$:

$$\hat{x}_1^* = \frac{g_1}{(1-\mu_0)\phi} + \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F(\hat{x}_s) \right) \left\{ F(\hat{x}_{t+1}) \frac{g_{t+1}}{(1-\mu_0)\phi} - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \right\}. \tag{8}$$

Remark 1 *The structure of expected lifetime utility at any period t, conditional on being unaware, is identical to the problem of a newborn individual, with the only difference that*

a newborn individual has prior belief μ_0 whereas an unaware individual has the belief $\hat{\mu}_{t-1}$. Therefore $\hat{x}_1^*$ is identical to that of $\hat{x}_s^*$ up to the time indices.

The problem of selecting the optimal policy from period 1 onwards is entirely analogous to that of selecting a policy, while unaware of type, from some period $t > 1$ onwards. So the problem of a person who is born with initial belief μ' is identical to the problem facing a person who has, after t periods of successful resistance to temptations, obtained the updated belief equal to μ'.

The following proposition characterizes optimal individual behavior.

Proposition 1 *There exists a unique solution to the agent's maximization problem. If the agent is risk averse in the utility over beliefs about herself ($\rho > 0$) then the solution is a strictly positive and convergent sequence of cutoffs $\hat{x}_1^*, \hat{x}_2^*, \ldots$ such that, while she remains unaware of her type, she selects a positive intent to pass on every temptation x_t such that $x_t \leq \hat{x}_t^*$, and give up otherwise.*

This result indicates that risk aversion is a necessary condition for self-restraint - the intuition for this is explained in the next subsection. From now on we assume $\rho > 0$. We assumed that the optimal policy in each period would adopt the cutoff structure. The fact that the FOCs have a unique solution $\hat{x}_s^*$ in each period and that the objective function is concave in each cutoff imply that the optimal policy has to adopt the cutoff form. Because the problem at hand is time-consistent, the cutoffs that the agent "plans" for future periods will still characterize her behavior if she were to reach those periods in a state of unawareness.

Note that we did not assume that larger temptations are harder to resist: The probability of intended resistance turning into actual resistance is independent of the size of the temptation. The fact that individuals are more likely to resist small temptations is entirely due to their optimization behavior.

3.3 Characteristics of individual behavior

3.3.1 The role of risk aversion

As just shown, a necessary and sufficient condition for wanting to resist temptations is to be risk averse over self-image, or beliefs about one's type. Let us go back to the behavior that Weber associated with the Calvinist ethic. We mentioned earlier a problem with the Weberian view. Why would one want to incur a cost in terms of forgone consumption

in order to maintain any conviction one may have, when this conviction cannot change in expectation? An attractive alternative could be to just find out the truth about one's type and then live accordingly. According to our model, individuals who fit the Weberian account must dislike risk over their own beliefs about their salvation. The reason is that beliefs are a martingale, which means that the agent cannot alter her beliefs in expectation. Why would then she attempt to pass on a positive temptation? The intuition is that by intending to resist individuals reduce the risk over their beliefs, which is valuable to a risk averse individual. A similar logic arises in Kőszegi's (2006) model of task choice.

A necessary and sufficient condition for the agent to be interested in attempting to resist temptations is for her to have risk averse preferences over beliefs about her type. Let us go back to the example of the behavior that Weber associated with the Calvinist ethic. According to our model, individuals that behave in that way must dislike risk over their own beliefs about their salvation. Why is risk aversion a requirement for such behavior? The reason is related to beliefs being a martingale, which means that the agent cannot alter her beliefs in expectation. Why would then she attempt to pass on a positive temptation? The intuition is that by resisting individuals reduce the risk over their beliefs, which is valuable to a risk averse individual. The appendix includes a formal explanation in the simple case of an individual who lives for a single period. It is worth noting that the role of risk aversion in creating endogenous self-restraint can be made in a static model. The dynamic model is necessary to study the evolution of beliefs, and as discussed later, the infinite horizon is important in order to abstract from terminal date effects.

Are people really risk averse regarding their beliefs? While we do not know of systematic evidence, the behavior of individuals facing a probable worrying medical diagnosis is suggestive that risk aversion over beliefs may play a role in human behavior. Most individuals who have a parent with Huntington's disease, and therefore a 50% probability of having the disease themselves, prefer not to take the genetic test.[9] If these individuals were typical expected utility maximizers that only care about outcomes, they would want to find out whether they have the disease, in order to make adjustments prior to the onset of this incurable disease that sets in during middle age and is ultimately lethal. The fact that so many refuse the test is suggestive of risk aversion over beliefs.

[9] "Facing Life With a Lethal Gene." New York Times, March 17, 2007.

3.3.2 The role of imperfect free will

There is a second puzzling aspect to the Weberian account of the Calvinist ethic. It is not obvious how one should interpret favorably any good acts that one has undertaken with the known objective of producing favorable evidence of one's own salvation. One possibility is that individuals may forget why they took an action in the past, as highlighted by Bénabou and Tirole (2004). But if an individual remembers her motivation to produce just that evidence of salvation, she could attribute the good acts to these deliberate attempts, and not to any underlying unknown type. Indeed, if intent always turns into action ($\phi = 1$) individuals cannot learn about their type. As long as they always choose a positive intent, they remain unaware and keep the prior belief μ_0.

However, choosing no intent induces a gamble that involves learning one's type. A high enough temptation can lure the agent to accept the gamble. She now faces an optimal stopping problem in a stationary environment. As there is no growth in self-image, $\hat{x}^*$ is the same in every period as long as the individual remains unaware. Therefore it is defined by a stationary version of (8), where $\hat{x}^*_s = \hat{x}^*$ for all s and $g_s = g = u(\mu_0) - \mu_0 u(1)$ which is positive for risk averse individuals. With the constant cutoff it simplifies to

$$\hat{x}^* = \frac{g}{1-\mu_0} + \left(\frac{1}{1-\lambda F(\hat{x}^*)}\right)\left(\frac{g}{1-\mu_0} - E[x|x<\hat{x}^*]\right) \tag{9}$$

This fixed point equation defines the optimal stationary cutoff. Note that the LHS is a 45-degree line, while RHS begins at a positive value and grows towards a finite limit. Therefore there has to be at least one solution. This shows that, while risk aversion (which implies $g > 0$) is necessary to have individuals pass on temptations, imperfect free will is not. Imperfect free will is necessary for people to learn from past actions of resistance.

3.3.3 Endogenous moral standards, moral capital, and Aristotelian virtue

An important question is whether a person who begins by selecting a positive intent has more or less of a reason to keep doing that as time goes by and she sees herself resist. In his treatment of moral virtues in Nichomachean Ethics (see esp. Book II) Aristotle held that a moral disposition is developed by the performance of moral acts. In his view, learning plays a role in moral development, and the more a person behaves virtuously, the easier it gets to continue to behave that way. Is this true of the individual in our model?

In our model, and individual who, having selected a positive intent at time t, resists, will update her prior $\hat{\mu}_{t-1}$ to a higher level $\hat{\mu}_t$. This makes the utility to be had in terms of

self-esteem even higher, suggesting that higher beliefs over time should push the individual to attempt to resist higher temptations. However, selecting a positive intent is counter-productive in the event that one is truly good (a state that is now deemed more likely), because the self-esteem return will be only $u(\hat{\mu}_t)$ instead of $u(1)$. To put it simply, the cost of positive intent is a possible forgone temptation, and the benefit is a lower variance of beliefs, but this variance goes to zero when self-image gets very high. As a result, it is not obvious that individuals with higher self-image should have higher cutoffs. But we are able to show,

Proposition 2 *Individuals who are successful in resisting temptations become more predisposed to resist further temptations. Formally, the sequence* $\hat{x}_1^*, \hat{x}_2^*, \ldots$ *is increasing.*

Proof. See Appendix. ■

Good actions bring stronger confidence of having the good type. This higher confidence, which we call *individual moral capital*, in turn predisposes one to resist even larger temptations. The key to the proof is that although gains from reducing the variance of beliefs get smaller as beliefs get close to certainty, the expected cost in terms of forgone consumption goes to zero faster. To see that the latter costs must decrease, note that the agent's intent will get in the way of her enjoying a temptation in period t only if she is bad and has free will in t. This event has a joint probability $\left(1-\hat{\mu}_{t-1}\right)\phi$. Therefore, as beliefs $\hat{\mu}_t$ get close to one the cost in terms of forgone temptations gets close to zero.

An important aspect of the last proposition is that the effective propensity of (bad) individuals to submit to temptations is endogenous. In other words, we can interpret the sequence of cutoffs $\hat{x}_t^*$ as the individual's moral standards, and we see that these standards evolve over time, depending on the history of temptations, intent decisions, and actions. Bad individuals who have always received temptations below their thresholds, and who have always had free will, will become morally robust over time. However, their high standards owe nothing to any underlying superiority in terms of fixed individual traits, and owe much to having had a quiet life in terms of temptations and luck at having been in control of their actions. Any bad type may suddenly lose her moral capital for two reasons: (i) having selected a positive intent, she may lack free will and see herself take the temptation; (ii) alternatively, she may receive a temptation above her current cutoff, and select no positive intent, which will also trigger her taking the temptation. This will immediately take her posterior to zero. After that, she will take every temptation coming to her because her standards, as measured by cutoffs in the space of temptations, have dropped to zero.

3.3.4 Discussion on modelling features

Now that the basic characterization of individual behavior is complete, we make a few remarks regarding our modelling approach.

Infinite horizon

The point that risk averse individuals will resist some temptations can be made in simpler finite horizon settings. But investigating whether past good behavior has the effect of strengthening moral dispositions requires our using an infinite horizon model. The reason is as follows. An individual's decision to resist a temptation takes into account the value of the current temptation against the stream of self-esteem returns net of future expected temptations. A shorter future diminishes that net value of future self-esteem returns. Thus, the stream of payoffs associated with good behavior depends both on the state variable capturing moral capital as well as on the remaining lifetime. Because individuals accumulate moral capital over time, isolating the effects of moral capital in a finite horizon model would be difficult, as these effects would be confounded with those of a shortening horizon. Indeed, with a finite horizon cutoffs eventually decrease even as beliefs continue to increase – the policy function becomes non-monotonic due to a shortening of the horizon and not to any other change in the connection between beliefs and incentives to resist. An infinite horizon model offers a setting that is stationary up to the value of the state variable, and hence allows us to isolate the effect of interest.

Dichotomous types

We assumed that good types always behave, while bad types may not, so types are very different. In a more general version of the model, one could imagine that both types may misbehave, with good types having a lower chance of wrongdoing when deciding to resist. In fact, the model we use is a limit case of a richer one where, in the absence of an active intent to resist, good types behave with a probability α_g while bad types resist with a lower probability α_b. When attempting to resist, both types will behave for sure if their free will works, and only with their type-related chance if their free will fails. That is, good types will behave with probability $\alpha_g (1-\phi) + \phi$ while bad types behave with the lower probability $\alpha_b (1-\phi) + \phi$. This model would again imply that good behavior leads to a higher self-image, while bad behavior leads to a lower self-image, although beliefs do not go down to zero in the event of wrongdoing. Working out the full dynamics in this richer model is very difficult because the number of states explodes, while dynamic programming methods are unable to deal with our model. This is due to the fact that the conditions usually invoked in order to characterize policy functions when using dynamic programming are stronger than

necessary and not met in our model.[10] However, the basic facts of the static version of the model with a single period can still be proved: a decision to resist yields a lower variance gamble in terms of future beliefs and therefore risk averse individuals will choose to resist temptations.

We have, however, simulated this richer model. According to our numerical results, even if $\alpha_b > 0$ the policy function is monotonic and the results in the paper remain. If $\alpha_g < 1$ good types may at times err, so the policy function becomes eventually decreasing for high enough beliefs. The reason is that for very high beliefs that one has the good type, a fall is interpreted as a tremble from one's type, rather than as evidence of having the bad type. Therefore the dynamic path of the unaware contains a part where eventually the individual becomes sanctimonious while lowering his own standards. In this version of the model the results in the paper can be established as possibility results for a subset of initial priors.

Deciding to be bad

We assumed that free will only gets in the way when attempting to resist. In other words, there is no symmetric decision to actively seek to commit a crime, decision which could be thwarted by a lack of free will. We believe the version we have used better captures the essence of wrongdoing: most of morality is defined around trying to control impulses towards self-serving goals. But a symmetric version of the model is possible, where imperfect free will may cause an attempt to misbehave to fail. Our results go through in this formulation provided one condition on parameter values is met. That condition ensures that selecting a positive intent leads to a lower-variance gamble in terms of future beliefs about self.

3.3.5 Comparative statics

We now examine the role of the initial prior μ_0 and of the effectiveness of free will ϕ (note ϕ could just be a belief). We also analyze the role of a brighter future in the form of an alternative distribution of temptations G that is first order stochastically dominated by F (i.e., G tends to generate lower temptations than F). For example, G could capture a better environment where the individual does not need bribes to live well. We then have

Proposition 3 *The sequence $\hat{x}_1^*, \hat{x}_2^*, \ldots$ is higher when*

(a) temptations x are drawn from G rather than F, where $G(x) > F(x)$ for all x.

[10] To prove monotonicity of the policy function through a dynamic programming approach we would have to rely on results hinging on two sufficient conditions: that the per period expected payoff function and the transition function describing the probabilities over future beliefs be supermodular in x_t and $\hat{\mu}_t$. The first condition can be met with a minimal change in the utility function we use. The second condition is violated.

(b) the initial belief μ_0 is higher.

(c) the effectiveness of free will ϕ (or the belief in it) is higher (shown numerically under exponential distribution of temptations).

(d) the discount factor λ is higher.

Proof. See Appendix. ■

Part (a) means that when the individual expects lower temptations in the future she will choose more stringent moral standards today.

Part (b) means that an individual with higher initial beliefs will also choose more stringent standards. This suggests that if parents desire that their offspring resist temptations they would want to inculcate in their offspring a high belief in their own goodness.

Part (c) means that when individuals believe that they have more control over their actions they will choose more stringent standards. This result could only be shown numerically for exponential distributions over temptations, and is far from obvious as there are forces pushing in opposite ways. With stronger free will a positive intent is more likely to secure a self-esteem gain but it is also more likely that it will preclude enjoying the temptation. Moreover, a higher ϕ reduces the positive updating that takes place in case the temptation is resisted.

These results imply that a better environment (in terms of higher μ and ϕ, if we take the beliefs to be rooted in the true values, and in terms of the distribution of temptations) reduce the probability that the individual has done wrong by a given date due to two effects. Taking the case of the distribution of temptations, the direct effect is that, given the individual's standards, a better environment makes it less likely that a high enough temptation will materialize so as to induce the individual to give up. The indirect effect is that the expectation of a better environment leads the individual to resist even larger shocks, complementing the direct effect. This positive feedback suggests that small differences in the environment could generate relatively large departures in the propensity to do wrong.

Finally, part (d) states that when the individual cares more about the future she will attempt to resist more temptations.

The results (c) and (d) match the emphasis by criminologists holding that the inability to control impulses and to take the future into account play a role in the disposition toward crime (see Gottfredson and Hirschi 1990, and Nagin and Paternoster 1993).

4 Moral capital and wrongdoing in a society

We now consider a society of individuals each facing the problem introduced in the previous section. We assume that shocks are independent across individuals and that the society is large in the sense that the law of large numbers can be used to derive the wrongdoing rates in the society. We first analyze the evolution of the wrongdoing rate within a cohort of individuals. Then we introduce an exogenous death rate in order to analyze wrongdoing rates in a society that is in a demographic steady state.

Our analysis of individual behavior proceeded without specifying the actual probability that an individual has a good type, because individual decisions depend only on subjective probabilities. In what follows, the individual choice variables $\hat{x}_t$ should be interpreted as having been optimized given beliefs μ_0 and ϕ. While the individual intent to resist temptations depends on $\hat{x}_t$, the ability to actually resist temptations conditional on intent depends on whether one really is a good type and has free will. We denote the actual share of good types by μ and assume that ϕ is a correct belief.

4.1 Wrongdoing rate within a cohort

Consider a cohort of individuals born into age $t = 1$ with initial belief $\mu_0 \in (0, 1)$ that may or may not be equal to μ. The share of aware individuals—those with the belief $\hat{\mu}_t \in \{0, 1\}$—increases over time, and a fraction $1 - \mu$ of the aware individuals will do wrong. We know from Proposition 2 that as a cohort ages the resistance cutoff $\hat{x}_t$ increases. The only ones to resist temptations at age t are those who either have the good type, or those who, despite being bad, end up the period continuing to be unaware of their type. Those who end age t aware of being bad did wrong at age t. (This includes individuals who only became aware during age t, i.e., after doing wrong for the first time). Thus the population wrongdoing rate at age t is the probability that an individual has become aware of being bad by the end of age t:

$$\begin{aligned} w_t &= (1-\mu)\left(1 - \Pr\left(\text{unaware}|\hat{x}_1, \ldots, \hat{x}_t, \text{bad}\right)\right) \\ &= (1-\mu)\left(1 - \phi^t H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\right). \end{aligned} \tag{10}$$

As the cohort ages, the term $\phi^t H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)$ approaches zero and the wrongdoing rate w_t increases monotonically converging to the share of bad types $1 - \mu$. (All convergence in this model is only asymptotic, in this case because $\phi^t H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)$ is strictly positive for any finite t.) Resisting individuals must become less numerous because bad types eventually

become aware – either because a very high temptation materializes, or because their free will fails in some period.

The evolution of wrongdoing rates is linked to the evolution of the distribution of beliefs. Notice first that, at age t, there are only three possible beliefs. The aware either know for sure that they are bad or that they are good; the unaware have used the Bayesian updating formula t times and so hold the same belief.

Type	Belief μ_t	Population share
Aware good	1	$\mu\left[1-H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right)\right]$
Aware bad	0	$(1-\mu)\left[1-\phi^t H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right)\right]$
Unaware	$\hat{\mu}_t=\frac{\mu_0}{\mu_0+\phi^t(1-\mu_0)}$	$\mu H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right)+(1-\mu)\,\phi^t H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right)$

The average belief at age t is therefore

$$\bar{\mu}_t=\mu+(\mu_0-\mu)\,H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right) \tag{11}$$

Recall that $H_0=1$ and $H_t>H_{t+1}$, so $\bar{\mu}_0=\mu_0$ and $\bar{\mu}_t$ converges monotonically to μ as $t\to\infty$. If $\mu_0>\mu$ then the average belief in society converges to μ from above, while if $\mu_0<\mu$ then it converges to μ from below. The true distribution of types (and hence the limiting distribution of beliefs) follows the Bernoulli distribution with success parameter μ. As the newborn have identical beliefs, the variance of their beliefs is zero, while the limiting distribution has variance $\mu(1-\mu)$. Gathering the above results we get

Proposition 4 *As a cohort ages,*

(a) the wrongdoing rate increases and converges to the share of bad types $1-\mu$*,*

(b) the average belief converges monotonically to μ*, and*

(c) the variance of beliefs increases and converges to $\mu(1-\mu)$*.*

In particular, if initial beliefs are consistent with reality ($\mu_0=\mu$) then the average belief can never change. Regardless of how incorrect the initial beliefs may be, the wrongdoing rate keeps increasing as beliefs become more polarized. The reason is simple: good types do good regardless their awareness state, but bad types do wrong less often when unaware. This proposition also implies that, if the initial prior is pessimistic ($\mu_0<\mu$) then the average self-image will improve (as $\bar{\mu}_t$ increases towards μ) while the wrongdoing rate increases.

4.2 Wrongdoing rate of a society in steady state

In this section we show that, in a world where people eventually die and are replaced by births of new unaware individuals, two societies with the same share of bad types can have

different wrongdoing rates. Thus even long run corruption rates across countries do not necessarily and exclusively reflect "deep" moral fundamentals captured by the share of bad types.

Now interpret the parameter λ not as a discount factor stemming from impatience but as a constant survival probability facing each individual. Assume survival to be independent of all other features in the model. This interpretation of λ is immaterial for the individual decision and the wrongdoing rate within a cohort. Suppose also that a new cohort is born in every period, and that the size of newborn cohorts is constant. These simplifying assumptions allow for a tractable steady state analysis, as they mean that the size of every age group is constant over time.

Denote the steady-state population share of age$-t$ individuals by z_t. Entry and exit from each age group must balance out ($z_{t+1} = \lambda z_t$, $t = 1, 2, \ldots$), which requires the shares to be

$$z_t = (1 - \lambda)\, \lambda^{t-1} \quad \text{for } t = 1, 2, \ldots. \tag{12}$$

The steady-state rate of wrongdoing in society is the population-weighted average of cohort wrongdoing rates (10),

$$W \equiv \sum_{t=1}^{\infty} z_t w_t = (1 - \mu) \left\{ 1 - (1 - \lambda) \sum_{t=1}^{\infty} \lambda^{t-1} \phi^t H_t \left(\hat{x}_1, \ldots, \hat{x}_t \right) \right\}. \tag{13}$$

The proportion of bad types gives the worst-case potential for the wrongdoing rate in society so W must obviously be strictly below $1 - \mu$ since at least some bad types sometimes resist temptations. But just how much short of $1-\mu$ the steady state wrongdoing rate falls depends on the parameters of the model.

Proposition 5 *The steady state rate of wrongdoing in society W,*

(a) is lower when the initial beliefs μ_0 of the newly born are higher,

(b) is lower when the distribution of temptations F is lower in the first order stochastic dominance sense,

(c) is lower when the probability that free will works ϕ is higher (under exponential distributions of temptations).

Proof. Part (a) follows from Proposition 3(a); part (b) follows from Proposition 3(b); part (c) follows from

$$\frac{dW}{d\phi} = -(1 - \mu)(1 - \lambda) \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} t \phi^{t-1} H_t \left(\hat{x}_1, \ldots, \hat{x}_t \right) + \sum_{t=1}^{\infty} \lambda^{t-1} \phi^t \frac{dH_t \left(\hat{x}_1, \ldots, \hat{x}_t \right)}{d\phi} \right\} < 0,$$

where the sign follows from the fact that $dH_t/d\phi > 0$ from proposition 3(c). ■

It is clear from (13) that a higher true share of good types results in a lower wrongdoing rate. And it is also easy to see that regardless of the true population share of good types, the social wrongdoing rate is lower whenever μ_0 and ϕ are higher (even if these are incorrect beliefs), as well as when the distribution of temptations is lower. This follows from the results in Proposition 3, showing that such parametric changes make individuals more resistant to temptations. This suggests a useful social role for indoctrination in terms of inculcating favorable beliefs. Anything that gives rise to a widespread overly optimistic perception μ_0 could historically have been a factor in the "natural selection" between competing societies. Note that the steady-state patterns of wrongdoing are not qualitatively different even if the existence of "good types" is purely imaginary, i.e. if $\mu = 0$.

4.3 Response to shocks: wrongdoing across societies

Let us consider how a society responds to aggregate shocks in the distribution of temptations. For example, a period with adverse macroeconomic conditions would likely expose the population to higher temptations. Two otherwise similar societies who face different macroeconomic shocks may end up with different wrongdoing rates.

The case of a cohort Consider first two identical cohorts in similar environments, one of which encounters a temporary shock to its distribution of temptations. By shock we mean that, for one period, individual temptations are drawn from some distribution G instead of the usual F. Call the shock "bad" if G stochastically dominates F (i.e., $G(x) < F(x)$ for all $x > 0$) and "good" if the opposite is true. The shock comes as a surprise and is not expected to be repeated, so individuals use $\hat{x}_t$ from Section 3 as their optimal policy. Suppose that the shock takes place s periods after the birth of the cohorts. Obviously behavior before period s is identical across the two cohorts.

Proposition 6 *Of two otherwise similar cohorts, one that has encountered a bad (good) shock in the past has a permanently higher (lower) wrongdoing rate. The difference in wrongdoing rates converges to zero as the cohort becomes infinitely old.*

Proof. Using the expression for w_t in (10), and the definition of H_t from (3) where G replaces F at the time of the shock, the wrongdoing rate at ages $t \geq s$ for a cohort that experienced the shock at age $s \geq 1$ is

$$w_{t,s} = (1-\mu)\left\{1 - \phi^t H_t(\hat{x}_1, \ldots, \hat{x}_t)\frac{G(\hat{x}_s)}{F(\hat{x}_s)}\right\}. \tag{14}$$

Clearly $w_{t,s} > w_t$ for all $s \geq t$ if $G(\hat{x}_s) < F(\hat{x}_s)$, and vice versa if $G(\hat{x}_s) > F(\hat{x}_s)$. As $t \to \infty$, $\phi^t H_t \to 0$ so $w_{t,s} \to 1-\mu$. ■

The wrongdoing rates of the shocked cohorts converge to $1-\mu$ just as they do for a cohort that was not shocked, so eventually the effects of the shock wash out. Nevertheless, history matters, as wrongdoing rates are determined by a process that has memory. Bad shocks that prompted a higher share of people to give in to temptations in one period accelerate the polarization of beliefs and yield higher wrongdoing rates for every subsequent period. This underscores that moral capital at the level of society is not just about the average belief of individuals. It depends also on how beliefs are distributed across individuals.

The case of a society in demographic steady state Now consider a whole society that faces the shock G in some period; call that period zero without loss of generality. We are interested in the level of wrongdoing in society s periods after the shock. At that point all cohorts born less than s periods ago are not affected by the shock so their wrongdoing rate is given by (10), while those that were born during or before the shock have the wrongdoing rate given by (14). Combining the cohort wrongdoing rates with the population shares (12), the aggregate rate of wrongdoing s periods after the shock is

$$\begin{aligned} W_s &= \sum_{t=1}^{s} z_t w_t + \sum_{t=s+1}^{\infty} z_t w_{t,s} \qquad (15) \\ &= (1-\mu)\left\{1-(1-\lambda)\left(\sum_{t=1}^{s}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t) + \sum_{t=s+1}^{\infty}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\frac{G(\hat{x}_{t-s})}{F(\hat{x}_{t-s})}\right)\right\} \end{aligned}$$

where we define $\sum_{t=1}^{0}\text{term}_t \equiv 0$ for convenience to cover the case $s=0$. The direction of the shock depends on the ratios $G(\hat{x}_t)/F(\hat{x}_t)$ in the natural way. Clearly the wrongdoing rate must eventually return to the steady state value, as ever fewer survivors remain from the shocked period.

Proposition 7 *Of two otherwise similar societies, one that has encountered a bad (good) shock in the past has a permanently higher (lower) wrongdoing rate. The difference in wrongdoing rates converges asymptotically to zero over time.*

Proof. The difference of the wrongdoing rates in (15) and (13) is the deviation of society's wrongdoing rate from steady state s periods after the shock:

$$\Delta_s = W_s - W = (1-\mu)(1-\lambda)\sum_{t=s+1}^{\infty}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\left(1-\frac{G(\hat{x}_{t-s})}{F(\hat{x}_{t-s})}\right). \qquad (16)$$

This is positive if the shock is bad (i.e., if $G(x) < F(x)$), and negative if the shock is good. As $s \to \infty$, the terms $\lambda^{s-1}\phi^s H_s(\hat{x}_1, \ldots, \hat{x}_s)$ converge to zero. ■

If the shock is bad (i.e., $G(\hat{x}) < F(\hat{x})$ for all $\hat{x}$) then the deviation from steady state $W_s - W$ is positive. History matters through its impact on the stock of unaware individuals. Bad shocks accelerate learning, lower that stock, and augment wrongdoing. Tirole (1996) offers a model of corruption persistence based of stereotyping, where the extrinsic marginal cost of corruption to individuals (namely the marginal impact on the probability of getting caught) is assumed to be decreasing under repeated acts of corruption. His model generates a form of strong persistence of corruption in the form of multiple steady states. In our model self-reinforcing effects are permanent at the individual level and arbitrarily long lasting, but not eternal, at the social level. The effects of any shock will die out asymptotically because those who experience the shock eventually die and are replaced by new cohorts who did not experience the shock.

5 Applications

5.1 Enhanced punishment for repeat offenders

In this subsection we consider a planner who is interested in minimizing wrongdoing and who can offer incentives to agents. For concreteness, we will focus on punishments for bad behavior and assume the planner can detect bad behavior with some exogenous probability.

The margin we investigate here is whether punishment should change with offense history. In order to isolate the effect of interest we impose the following simplifications. The planner knows past behavior by all agents in a single cohort and has a one-time capability to impose punishment on those who do wrong in the current period. Denote with N_a and N_u the expected punishment to be imposed respectively on the aware and the unaware that do wrong. (In other words, N_a and N_u incorporate the probability of detection.) The net expected return from seizing a temptation x is therefore $x - N_a$ for the aware and $x - N_u$ for the unaware.

A planner who wants to minimize wrongdoing would have an easy task if punishment were costless. So we assume that expected punishments are costly to the planner, as captured by an increasing and convex function $c(N_a + N_u)$. This cost formulation captures a world where threatening with more likely and intense punishment is costly because it requires stronger

detection and punishment capabilities.[11] Lastly, we assume that the planner discounts the future according to the factor $\delta < 1$, while individuals have a survival rate λ and do not further discount time. Also, in order to guarantee the satisfaction of the second order conditions in the planner's problem, we assume that larger temptations are less common than small ones, i.e., that $f(x)$ is decreasing.[12]

To construct the objective of the planner, we first characterize the impact of punishment on wrongdoing. Because those who are good never do wrong, it is sufficient to concentrate on the behavior of the bad types; we normalize their mass to 1. We know from previous sections that, absent punishment, those who are bad and aware of it do wrong for sure. But threatened with a punishment N_a they would attempt to resist whenever the realized temptation satisfies $x < N_a$. Therefore, given a punishment N_a, the rate of wrongdoing among the aware will be $1 - \phi F(N_a)$. That means the punishment on the aware obtains a reduction in wrongdoing of exactly $\phi F(N_a)$ in the current period. As the punishment is for the current period only, and the aware learn nothing regardless of their action, N_a has no further impact on wrongdoing.

The impact of current period punishment on wrongdoing by the unaware is more complex and is captured in the following,

Lemma 1 *A one time punishment N_u attains a reduction in the expected wrongdoing of unaware individuals equal to $\phi\left(F\left(\hat{x}_1 + N_u\right) - F\left(\hat{x}_1\right)\right)\sum_{s=1}^{\infty}\left(\delta\lambda\right)^{s-1}\phi^{s-1}\frac{H_s}{H_1}$.*

Proof. See Appendix. ■

The proof shows that under punishment N_u the current period cutoff satisfies $\hat{x}_1^p = \hat{x}_1 + N_u$, so current punishment raises the current optimal cutoff of the unaware one for one. Thus punishment achieves a reduction in current wrongdoing equal to $\phi\left(F\left(\hat{x}_1 + N_u\right) - F\left(\hat{x}_1\right)\right)$. But because punishment complements the effects of moral capital it raises the share of unaware individuals who resist and remain unaware, leading to lower wrongdoing in future periods. Specifically, of those who are saved from temptation in the current period,

[11] Costs may also increase with the number of people who do wrong and who must eventually be punished. We abstract from this possibility which would introduce a form of increasing returns to punishment, as larger punishments could pay for themselves through a lower number of inmates. Our results in this subsection are robust in the face of those effects if we impose a technical condition on the distribution of temptations to ensure that overall punishment costs continue to be convex.

[12] This assumption is not strictly necessary. But as is well known, when densities of arbitrary shape are involved in an optimization problem second order conditions may not be satisfied. So assuming a monotonic density simplifies matters, and within this class a decreasing density becomes necessary to make sure that the second order conditions are satisfied.

$\phi\lambda F(\hat{x}_2)$ are saved again in period 2, and $(\phi\lambda)^2 F(\hat{x}_2) F(\hat{x}_3)$ are saved in period three, and so on, explaining the expression in the last lemma, where $\sum_{s=1}^{\infty} (\delta\lambda)^{s-1} \phi^{s-1} H_s/H_1 = 1 + \delta\lambda\phi F(\hat{x}_2) + (\delta\lambda)^2 \phi^2 F(\hat{x}_2) F(\hat{x}_3) + ...$ captures the present and future (discounted) reductions in wrongdoing. All future cutoffs are unchanged.

Social planner's problem

Using lemma (1), the planner's objective is to maximize,

$$\phi F(N_a) + \phi [F(\hat{x}_t + N_u) - F(\hat{x}_t)] Z_t - c(N_a + N_u) \tag{17}$$

with respect to N_a and N_u, where

$$Z_t = \sum_{s=1}^{\infty} (\delta\lambda)^{s-1} \phi^{s-1} \frac{H_s}{H_1}, \tag{18}$$

which only contains future cutoffs and does not involve $\hat{x}_1^p$. Given this program, and our stated assumptions, we obtain,

Proposition 8 *If the planner's patience or the agents' survival rate are low enough, then the planner imposes harsher punishment on repeat offenders relative to first-time wrongdoers. Formally, if δ or λ are low enough, then $N_a > N_u$.*

Proof. See Appendix. ■

An intrinsic disposition to resist temptations allows individuals to behave honestly even when there are no extrinsic incentives in place. And extrinsic incentives can obviously help to keep individuals behaving honestly. Proposition 8 in this appendix tells us that the design of extrinsic incentives should reflect the strength of intrinsic dispositions to avoid wrongdoing. In this extension of our model, a planner spends less resources trying to deter agents that already have intrinsic self-deterrent motives, and chooses to punish more harshly those who have lost their moral capital and are willing to take any temptation that comes their way. This design resembles the very common penal profile of heavier sentences on wrongdoers with a criminal record, and rules such as the "three strikes and you are out" that apply in many US states. Notably, in California there is a second strike provision according to which a second felony triggers a sentence twice as heavy (Clark, Austin, and Henry 1997). Note however that our last proposition does not support those institutions in an unconditional way. The planner should be sufficiently impatient, or agents die fast enough, so as to forgo an added benefit of imposing punishment on those who still have their moral capital. That

added benefit is the wider preservation of intrinsic incentives, which will lower wrongdoing in future periods.

This result carries over to the case where punishments are permanent. To see why, note first that future punishments make no difference to the decision of an aware person. Note next that higher permanent punishments N_a in the future would increase $\hat{x}_t$ today by making the life of wrongdoing less attractive (recall Proposition 3.a in the paper). This would further decrease the marginal deterrence value of N_u today by pushing the range of temptations where the punishment can affect individual behavior by the unaware even further to the tail of the distribution. This would reinforce the planner's incentives to increase the punishment on the aware.

This result shows that the design of extrinsic incentives should reflect the strength of intrinsic dispositions to avoid wrongdoing. Harsher punishment for repeat offenders can arise also in contexts of pure extrinsic deterrence.[13] Our point was not to characterize optimal deterrence in all generality, but to show that optimal extrinsic incentive schemes are affected by taking into account the endogeneity of moral standards.

5.2 Moral taboos and rituals

Moral taboos and rituals are sometimes sanctioned by religions or cultural norms and typically stipulate prohibitions to engage in certain acts. Very often, the taboos are against acts that convey satisfaction without imposing any obvious harm, such as eating and drinking certain things. For our purposes, a "taboo" can also be against deviations from some proscribed but avoidable inconvenience or "ritual", such as costly religious ceremonies, or other mandated behavior that deducts from otherwise available consumption utility. Here we analyze a rationale for such taboos.[14]

Suppose that individuals live for a period before they enter society and face the temptations we have considered so far. Before the initial period individuals have the possibility to consume a good (tea, say) that yields positive utility. Consider a tradition stipulating

[13] Polinsky and Rubinfeld (1991) and Polinsky and Shavell (1998) analyze conditions under which optimal fines may be higher for repeat offenders. In the first paper offense history tracks offense propensity. In the second it is shown that harsher punishment for repeat offenders may increase deterrence of first time offenders.

[14] For a different conception of taboos, see Fiske and Tetlock (1997), and Benabou and Tirole (2007). In the latter, the agent may decide to avoid information about the price of a "taboo" transaction, as part of a self-control strategy. See also the study of moral placebos in Prelec and Bodner (2003).

that consuming tea amounts to falling for a temptation. Now suppose that, as in our model, individuals who partake in the tradition consider such fall to reveal a bad type.

The size of the taboo temptation does not matter as long as individuals will attempt to resist it, so suppose the taboo is a temptation of size $x < \hat{x}_1$. Compared to a world without the taboo, the immediate benefit is that those who successfully resist the taboo will enter their first period with a resistance threshold $\hat{x}_2$ instead of $\hat{x}_1$. So, of all those bad types who had free will when facing the taboo (a fraction ϕ), a fraction $1 - \phi F(\hat{x}_2)$ will engage in wrongdoing in period 1 instead of a higher fraction $1 - \phi F(\hat{x}_1)$ which would engage in wrongdoing without the taboo (i.e., in a situation where consuming tea is not thought to convey information on one's type). The cost is that share $1 - \phi$ of individuals will fall to the temptation even before their first period because their free will fails them. Therefore, the gain from the taboo in terms of reduced wrongdoing in period 1 is,

$$1 - \phi F(\hat{x}_1) - [\phi(1 - \phi F(\hat{x}_2)) + (1 - \phi)] > 0,$$

which is positive whenever $\phi F(\hat{x}_2) > F(\hat{x}_1)$. The gain is increasing in the probability that a shock falls in between the original and the improved threshold. For the taboo to decrease wrongdoing the increase has to be sufficiently high to compensate for those who fall to the taboo temptation due to the failure of free will.

The taboo has a lasting impact on wrongdoing rates since survivors will carry with them a higher $\hat{x}_t$ in every subsequent period than what they would have had without the taboo. (Eventually this advantage fades away as $\hat{x}_t$ converges to its limiting value.) Assuming, for simplicity, that the breaking of the taboo does not count as actual wrongdoing, the wrongdoing rate of a cohort of age t that faced the taboo is

$$w_t^{'} = (1 - \mu)\left(1 - \phi^{t+1} H_t(\hat{x}_1, \ldots, \hat{x}_t) \frac{F(\hat{x}_{t+1})}{F(\hat{x}_1)}\right) \tag{19}$$

The impact of the taboo on steady-state wrongdoing in the society is,

$$W^{'} - W = -(1 - \mu)(1 - \lambda)\phi \sum_{t=1}^{\infty} (\phi\lambda)^{t-1} \left(\phi \frac{F(\hat{x}_{t+1})}{F(\hat{x}_1)} - 1\right) H_t(\hat{x}_1, \ldots, \hat{x}_t). \tag{20}$$

The taboo will lower the steady-state rate of wrongdoing in society when (20) is negative. Note that the choice of offering the taboo before the first period was mostly a normalization for the age index. A similar analysis would apply to an older cohort who could be exposed to a taboo in between ages $\tau - 1$ and τ, but with the above summation beginning at $t = \tau$.[15]

[15] Unadjusted, this formula would then mean that the artificial taboo period in the middle of the lifespan also comes with a risk of non-survival, and that the taboo was unanticipated by the individual.

5.3 Moral capital and career choice

How do individuals select into careers in an economy where individual beliefs vary and different careers offer different distributions of temptations? For concreteness, consider two occupations where one has a higher distribution of temptations, in the sense of first-order stochastic dominance. For example, one could consider politics as a high temptation activity and academia as a low temptation activity. The population consists of a continuum of individuals with heterogeneous initial beliefs $\mu \in [0, 1]$. We want to know how individuals self-select into different occupations depending on μ. We assume that the economy has a need for workers in both careers, hence compensation may have to adjust so that each career is preferred by some types. The mechanism of this adjustment is immaterial for our exercise; what is important is that in equilibrium individuals who require a lower compensating differential will self-select to the low-temptation career.

To make things simple, suppose individuals live for only one period. We then have

Proposition 9 *Consider an economy where individuals differ by initial self-image, and where two occupations offer different distributions of temptations, with one first-order stochastically dominant. In equilibrium, individuals with self-image above (below) an equilibrium cut-off enter the occupation with lower (higher) temptations.*

Proof. See Appendix. ■

For aware types the selection incentives are clear: An individual with $\mu = 1$ will be indifferent between the two careers, and will prefer the low-temptation career under any positive compensating differential favoring that career. An individual with $\mu = 0$ only cares about temptations and will choose the high-temptation activity unless there is a fairly large compensating differential in favor of the low-temptation activity. In between, the result is not obvious, because the unaware types have an incentive to protect their self-image by choosing a low-temptation activity. Low self-image individuals, judging themselves vulnerable, could be interested in protecting whatever little self-esteem they have by choosing a low temptation activity. As it turns out, the population can always be divided into just two segments by their beliefs μ so that types in the lower segment of self-beliefs will enter the high-temptation professions.

Are politicians more corrupt than academicians because they are inherently less moral or because they have more opportunities for corrupt behavior? In our model both arguments are correct. Even if people were divided randomly between occupations, the higher temptations would cause there to be more wrongdoing in the high-temptation sector, because

the opportunity cost of attempting to preserve a positive self-image is higher. However, the higher rate of wrongdoing in the high-temptation sector is further reinforced by the selection of types.

6 Conclusion

We propose a model of endogenous moral standards rooted in two ideas: that actions depend partly on unconscious drives and that people prefer to think they have the good type, i.e., that their unconscious drives are geared towards received morality. We characterize conditions under which self-restraint will emerge endogenously in the form of passing on enjoyable temptations for the sake of keeping a good introspective reputation. We also identify conditions for self-reinforcing patterns of virtue and corruption to emerge.

When intent does not fully determine actions, a history of resistance improves self-image and increases the disposition to resist temptations, yielding a view of morality as a cumulative process of habituation through action. This view of morality parallels Aristotle's account of the development of virtue. We view the improvement of the individual's self-image as a process of moral capital formation. When individuals perform actions that damage their self-image, durable damage is also done to their ability to resist such actions in the future, creating hysteresis in wrongdoing at the individual level.

Stronger initial beliefs about having a good type, lower expected temptations, a lower discount rate, and stronger confidence in one's ability to transform intentions into actions induce more stringent moral standards. At the social level, the wrongdoing rate is determined not just by the average self-image but more generally by its distribution across individuals. Societies with the same distribution of types but who have faced less fortunate histories involving larger temptation shocks will have to endure a more polarized distribution of individual self-images. This polarization will cause more wrongdoing even if the average self-image is the same across societies. Therefore, cross-country measures of wrongdoing and cultures of corruption may not reflect differences in deep moral fundamentals but simply different histories.

Our model offers some detail about the workings of identity (see also Bénabou and Tirole 2004). Akerlof and Kranton (2000) posit that identity affects behavior because it poses costs to an individual doing things that are deemed inappropriate for people with a given identity. Our model suggests that "identity-based costs" may not be constant, but respond to past actions and to the person's beliefs that such identity (e.g., that of a good person) is still

hers. The model can also rationalize taboos, why high temptation activities may attract the individuals least equipped to resist, thus magnifying wrongdoing differentials across activities, and a rationale for punishing repeat-offenders more harshly. This application illustrates that the optimal design of deterrence schemes may change when the disposition toward wrongdoing is endogenized.

Appendix

Proof of Proposition 1. We prove a series of lemmas (**2**, **3** and **4**), that jointly yield Proposition 1. The first lemma shows that optimal behavior is attached to a single sequence of cutoffs, the second one says that the first order conditions of the individual's problem identify the optimal cutoff sequence, and the third lemma says cutoffs will be positive iff $\rho > 0$.

Lemma 2 *There is a unique sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, ..$ characterizing optimal behavior.*

Proof of Lemma 2: From inspection of (7), each cutoff is uniquely determined as a sum of two terms: the first one captures the trade-off in the contemporary period ($\frac{g_s}{(1-\mu_0)\phi^s}$) and the second one captures the continuation value of the game up to a constant (the term $\sum_{t=1}^{\infty} \lambda^{t-s+1} \frac{H_{t+s-1}}{H_s} \times \left\{ \frac{F(\hat{x}_{t+s}) g_{t+s}}{(1-\mu_0)\phi^s} - \phi^t \int_0^{\hat{x}_{t+s}} x f(x) dx \right\}$ equals V_s minus a constant). Then the uniqueness of an optimal sequence characterized by (6) follows. To see this, suppose not. Then starting in some period $s \geq 1$ there is a number of periods where there is more than one cutoff forming part of a sequence satisfying the FOCs. Then all the optimal subsequences starting in period $s+1$ must yield the same continuation value. If not, following s the agent would choose the one subsequence yielding the highest expected payoff. But if all subsequences starting in $s+1$ yield the same continuation value, then there cannot be more than one cutoff in period s, because the FOC at s determines $\hat{x}_s$ uniquely as a function of the continuation value at $s+1$ and the term $\frac{g_s}{(1-\mu_0)\phi^s}$.■

One implication of this lemma is that the effects of changes in future thresholds (around the latter's optimal value) on the objective function cancel out and do not affect the optimal value of earlier thresholds.

Lemma 3 *A sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, \ldots$ satisfying the FOCs is a global maximizer of V_0.*

Proof of Lemma 3: First we show that a sequence $\{\hat{x}_i^*\}_{i=1}^{\infty}$ satisfying the FOCs constitutes a maximum. Later we show it is the only one.

Because the cross partial of V_0 with respect to any two cutoffs $\hat{x}_s, \hat{x}_t$ is zero (this can be shown through tedious but straightforward computation of the cross-partial), concavity of the objective function around each cutoff is sufficient for a maximum. Wlog we focus on the FOC for $\hat{x}_1$,

$$\frac{\partial V_0}{\partial \hat{x}_1} = f(\hat{x}_1) \left\{ \begin{array}{c} \{u_1[\mu_0 + (1-\mu_0)\phi] - \mu_0 u(1) - (1-\mu_0)\phi\hat{x}_1\} + \\ \frac{1}{F(\hat{x}_1)} \left(\sum_{t=1}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F_{t+1} u_{t+1} \left[\mu_0 + (1-\mu_0)\phi^{t+1}\right] \\ -\mu_0 F_{t+1} u(1) - (1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} \right) \end{array} \right\} = 0. \tag{21}$$

Inspection reveals that $V_0(\hat{x}_1^*, \hat{x}_2^*, ...)$ is concave in $\hat{x}_1$. First, $f(\hat{x}_1) > 0$. Second, the large product involving $\frac{1}{F(\hat{x}_1)}$ is independent of $\hat{x}_1$ (cancel $\frac{1}{F(\hat{x}_1)}$ out with the factor $F(\hat{x}_1)$ inside H_t). Then, at the optimum, any reduction in $\hat{x}_1$ below $\hat{x}_1^*$ would make $\{u_1[\mu_0 + (1-\mu_0)\phi] - \mu_0 u(1) - (1-\mu_0)\phi\hat{x}_1\}$ larger, making the entire left hand side of the FOC positive. A similar argument shows the entire LHS of the FOC would be negative for any $\hat{x}_1 > \hat{x}_1^*$.

To show that the sequence $\{\hat{x}_i\}_{i=1}^{\infty}$ constitutes a global maximum, note that this sequence is the unique interior extremum. So we just need to make sure it yields higher expected utility than some sequence where one or more thresholds take extreme values. Because the cross partials on cutoffs are zero, we can consider deviations in one threshold at a time. Can the agent gain by setting one threshold to the min in the support of x (or, analogously, by increasing the threshold without bound)? Suppose she can. Then the objective function attains another maximum at $\hat{x}_s = 0$. Because the objective function is increasing for $\hat{x}_s$ below but close to $\hat{x}_s^*$ and is continuously differentiable, the objective function must have a minimum somewhere in $(0, \hat{x}_s^*)$, a contradiction. A similar contradiction arises when considering the possibility of increasing $\hat{x}_s^*$ without bound.■

Lemma 4 *A necessary and sufficient condition for the sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, \ldots$ to be strictly positive and to converge asymptotically to a finite strictly positive limiting value is that $\rho > 0$.*

Proof of Lemma 4: We show first that the sequence $\{\hat{x}_t\}_{t=1}^{\infty}$ is positive iff $\rho > 0$. From Remark 1 all cutoffs are analogous up to μ_t. Thus, with no loss of generality, we focus now on showing that $\hat{x}_1 > 0$ iff $\rho > 0$. Recall that the solution for $\hat{x}_1^*$ is given by (8), which involves a lengthy second term that is the value of the objective function as of period 2 (up for the constant $\frac{\mu_0 u(1) + (1-\mu_0)Ex}{1-\lambda}$ which does not depend on any choice variable). That expression must must be nonnegative because by inspection it is clear one can always attain zero by setting all future thresholds to be zero. Therefore, it is sufficient that $g_1 > 0$ to get

$\hat{x}_1^* > 0$. Note that $g_1(\mu_0) > 0$ means that,

$$u(\mu_1)[\mu_0 + (1-\mu_0)\phi] - \mu_0 u(1) > 0 \Leftrightarrow \tag{22}$$

$$\left(\frac{\mu_0}{\mu_0 + (1-\mu_0)\phi}\right)^{1-\rho}[\mu_0 + (1-\mu_0)\phi] - \mu_0 > 0 \Leftrightarrow \mu_0\left(\frac{\mu_1^{1-\rho}}{\mu_1} - 1\right) > 0 \tag{23}$$

which is met if and only if $\rho > 0$. This does not show necessity, because the second term in $\hat{x}_1^*$ may be positive, so in principle $\hat{x}_1^*$ could be positive even if $\frac{g_1}{(1-\mu_0)\phi}$ is not. But note that for the second term of $\hat{x}_1^*$ to be positive it must have some positive terms $\frac{g_{t+1}}{(1-\mu_0)\phi}$. These have the same structure as $\frac{g_1}{(1-\mu_0)\phi}$, and also require $\rho > 0$ to be positive. If the second term of $\hat{x}_1^*$ is not positive then it is zero, and $\rho > 0$ becomes necessary for $g_1 > 0$.

To show that $\{\hat{x}_i^*\}_{i=1}^{\infty}$ converges to a positive limit whenever $\rho > 0$ we need to show two things: first, if $\{\hat{x}_i^*\}_{i=1}^{\infty}$ converges it does it to a unique finite limit, and second, that it really does converge. Note that as μ converges to unity the problem becomes stationary, so $\hat{x}^*$ should also be constant for all future periods. The limiting value of $\hat{x}^*$ must satisfy the following fixed point equation:

$$\hat{x}^* = G_1 + \sum_{t=1}^{\infty} \lambda^t F(\hat{x}^*)^t \left\{G_{t+1} - \phi^t E[x|x \leq \hat{x}^*]\right\} \tag{24}$$

where $E[x|x \leq \hat{x}^*] = \frac{1}{F(\hat{x}^*)}\int_0^{\hat{x}^*} xf(x)dx$ was used and $G_t = \frac{u\left(\frac{\mu_0}{\mu_0+(1-\mu_0)\phi^t}\right)\left[\mu_0+(1-\mu_0)\phi^t\right]-\mu_0 u(1)}{(1-\mu_0)\phi}$.The functional form of the utility function (as long as it is concave) affects $\hat{x}^*$ only via G_t.

Because $u(\mu) = \mu^{1-\rho}$, we have $\lim_{\mu \longrightarrow 1} G_t = \rho\phi^{t-1}$. We can simplify, from (24), the limiting value as the solution of $\hat{x}^* = \rho + \sum_{t=1}^{\infty} \lambda^t F(\hat{x}^*)^t \left(\rho\phi^t - \phi^t E[x|x \leq \hat{x}^*]\right)$. This can be written as $\hat{x}^* - \rho = \lambda\phi F(\hat{x}^*)\mathrm{E}[\hat{x}^* - x|x \leq \hat{x}^*]$. Note the right hand side of the last equation is nonnegative. Therefore, the left hand side yields $\hat{x}^* \geq \rho > 0$ leaving $\hat{x}^* > 0$. To see that this limit value $\hat{x}^*$ exists, is positive for all $\rho > 0$, and is unique, note that the left hand side in the last equality has slope equal to one, and the right hand side has slope $\lambda\phi F(\hat{x}^*) < 1$; hence the limit value for $\{\hat{x}_i^*\}_{i=1}^{\infty}$ exists and is unique. To see it converges, note that the sequence is bounded. This is clear from the fact that the continuation value is bounded for all t. Because the sequence is bounded, it has a convergent subsequence. Besides, because $\hat{x}^*$ is unique, every convergent subsequence converges to that point, and then the sequence converges to its unique limit value.■

Proof of Proposition 2: Note first that the resolution of the problem of determining the optimal sequence $\{\hat{x}_i^*\}_{i=s}^{\infty}$ is the same as solving for the sequence $\{\hat{x}_i^*\}_{i=1}^{\infty}$ up to the fact that one's beliefs will be higher in period s than they are in period 1. Therefore, we just need to show that $\hat{x}_1^*$ is increasing in the initial beliefs μ_0. As $\partial^2 V_0/\partial\hat{x}_1^*\partial\hat{x}_t^* = 0$, we are only

interested in $\frac{d\hat{x}_1}{d\mu_0}$ as given by the direct effects, plus the indirect effect that μ_0 has through its impact on the future values of $u(\mu_t)$, which depend on μ_0. Recall that $\hat{x}_1$ can be written as,

$$\hat{x}_1^* = \frac{g_1(\mu_0)}{(1-\mu_0)\phi} + + \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F_s \right) \left\{ F_{t+1} \frac{g_{t+1}(\mu_0)}{(1-\mu_0)\phi} - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \right\} \tag{25}$$

so we just need to show that $\frac{g_t(\mu_0)}{(1-\mu_0)\phi}$ is increasing in μ_0. So,

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{\left\{ \frac{du}{d\mu_t} \frac{d\mu_t}{d\mu_0} \left[\mu_0 + (1-\mu_0)\phi^t\right] + u_t\left(1-\phi^t\right) - u(1) \right\}}{(1-\mu_0)\phi} + \frac{g_t(\mu_0)}{(1-\mu_0)^2 \phi}. \tag{26}$$

The first term can be shown to equal $\frac{u_t\left[\mu_0+(1-\mu_0)\phi^t-\rho\phi^t\right]-u(1)}{\mu_0(1-\mu_0)\phi}$, so plugging this into $\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0}$ and using the definition for $g_t(\mu_0)$ we get,

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{\left[\mu_0 + (1-\mu_0)\phi^t - \rho\phi^t\right] u_t - \mu_0 u(1)}{\mu_0(1-\mu_0)\phi} + \frac{\left[\mu_0 + (1-\mu_0)\phi^t\right] u_t - \mu_0 u(1)}{(1-\mu_0)^2 \phi}, \tag{27}$$

and rearranging,

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{u_t\left\{\left[\mu_0 + (1-\mu_0)\phi^t\right] - (1-\mu_0)\rho\phi^t\right\} - \mu_0 u(1)}{\mu_0(1-\mu_0)^2 \phi}. \tag{28}$$

Therefore, we need to show $\left(\frac{\mu_0}{\mu_0+(1-\mu_0)\phi^t}\right)^{1-\rho} > \frac{\mu_0}{\mu_0+(1-\mu_0)(1-\rho)\phi^t}$. Tedious algebra shows that,

$$\left(\frac{\mu_0}{\mu_0 + (1-\mu_0)\phi}\right)^{1-\rho} > \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)\phi}, \ \phi \in (0,1), \rho \in (0,1), \mu_0 \in (0,1), \tag{29}$$

which implies the previous inequality.■

Proof of Proposition 3: (a) Follows from Remark 1 and the proof of Proposition 2.

(b) Again we can ignore indirect effects and compute only the partial derivative due to $\partial^2 V_0 / \partial \hat{x}_1^* \partial \hat{x}_t^* = 0$. Wlog we focus on $\hat{x}_1$, and compare its optimal value when the temptation in period k is expected to be drawn from G instead of F.

$$\begin{aligned}
\hat{x}_1^*(G) = {} & u_1\left[\frac{\mu_0 + (1-\mu_0)\phi}{(1-\mu_0)\phi}\right] - \frac{\mu_0}{(1-\mu_0)\phi} u(1) + \\
& \sum_{t=1}^{k-2} \lambda^t \left(\prod_{s=2}^{t} F_s \right) \left\{ \begin{array}{c} F_{t+1} u_{t+1} \frac{\left[\mu_0+(1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} F_{t+1} u(1) - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} \\
& + \lambda^{k-1} \left(\prod_{s=2}^{k-1} F_s \right) \left\{ \begin{array}{c} G_k u_k \frac{\left[\mu_0+(1-\mu_0)\phi^k\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} G_k u(1) - \phi^{k-1} \int_0^{\hat{x}_k} x f(x) dx \end{array} \right\} \\
& + \sum_{t=k}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F_s \frac{G_k}{F_k} \right) \left\{ \begin{array}{c} F_{t+1} u_{t+1} \frac{\left[\mu_0+(1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} F_{t+1} u(1) - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\}.
\end{aligned} \tag{30}$$

Note that $\hat{x}_1(F)$ is the same expression, only we should write F wherever we wrote G in the last expression. Then we can compute,

$$\hat{x}_1^*(G) - \hat{x}_1^*(F) = \lambda^{k-1}\left(\prod_{s=2}^{k-1} F_s\right)\left\{(G_k - F_k)\left[u_k\frac{\left[\mu_0 + (1-\mu_0)\phi^k\right]}{(1-\mu_0)\phi} - \frac{\mu_0}{(1-\mu_0)\phi}u(1)\right]\right\} +$$
$$\sum_{t=k}^{\infty}\lambda^t\left[\left(\prod_{s=2}^{t} F_s\frac{G_k}{F_k}\right) - \left(\prod_{s=2}^{t} F_s\right)\right]\left\{\begin{array}{c} F_{t+1}\left[u_{t+1}\frac{\left[\mu_0+(1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} - \frac{\mu_0}{(1-\mu_0)\phi}u(1)\right] - \\ -\phi^t\int_0^{\hat{x}_{t+1}} xf(x)dx \end{array}\right\}. \quad (31)$$

Note if a future threshold $\hat{x}_t$ is set to a positive value, it is because doing so must yield a positive payoff, which implies that all the terms in the summation inside $\hat{x}_1$ are nonnegative. This, together with $G_k > F_k$ implies that the last expression is positive.

c) This result is surprisingly hard to prove analytically. We have solved the model numerically covering the whole parameter space using exponential distributions for temptations and shown that the sequence of cutoffs increases in ϕ. These solutions are available upon request.

d) It is straightforward to show that V_0 is supermodular on $(\hat{x}_s, \lambda)$.■

Proof of Lemma 1: The unaware person facing punishment N_u (note the unaware person does not care about N_a because punishment only occurs in the current period) is to choose the sequence of cutoffs $\{\hat{x}_1^p, \hat{x}_2^p, ...\}$ to maximize,

$$\begin{aligned} V &= \frac{\mu_0 u(1) + (1-\mu_0)(Ex)}{1-\lambda} + F_1\left\{u_1\left[\mu_0 + (1-\mu_0)\phi\right] - \mu_0\right\} + \\ &+ (1-\mu_0)\left[\phi F_1 N_u - \phi\int_0^{\hat{x}_1^p} xf(x)dx - N_u\right] + \quad (32) \\ &+ \sum_{t=1}^{\infty}\lambda^t H_t\left\{\begin{array}{c} F_{t+1}u_{t+1}\left\{\left[\mu_0 + (1-\mu_0)\phi^{t+1}\right] - \mu_0\right\} + \\ -(1-\mu_0)\phi^{t+1}\int_0^{\hat{x}_{t+1}} xf(x)dx \end{array}\right\}, \quad (33) \end{aligned}$$

where F_t and H_t are functions respectively of $\hat{x}_t$ and the sequence of cutoffs $\{\hat{x}_1^p, \hat{x}_2^p, ...\}$. The first order condition for $\hat{x}_1^p$ is,

$$\begin{aligned} &\frac{\partial V}{\partial \hat{x}_1^p} = f(\hat{x}_1)\left\{u_1\left[\mu_0 + (1-\mu_0)\phi\right] - \mu_0\right\} - \quad (34) \\ &-(1-\mu_0)\phi\hat{x}_1 f(\hat{x}_1) + (1-\mu_0)\phi f(\hat{x}_1) pN_u + \quad (35) \\ + \quad &\frac{\partial}{\partial \hat{x}_1}\left(\sum_{t=1}^{\infty}\lambda^t H_t\left\{\begin{array}{c} F_{t+1}\left\{u_{t+1}\left[\mu_0 + (1-\mu_0)\phi^{t+1}\right] - \mu_0\right\} - \\ -(1-\mu_0)\phi^{t+1}\int_0^{\hat{x}_{t+1}} xf(x)dx \end{array}\right\}\right) = 0, \end{aligned}$$

from which, after some manipulation, we can solve for $\hat{x}_1^p$,

$$\hat{x}_1^{*p} = u_1\left[\frac{\mu_0 + (1-\mu_0)\,\phi}{(1-\mu_0)\,\phi}\right] - \frac{\mu_0}{(1-\mu_0)\,\phi} + pN_u \tag{36}$$

$$\frac{1}{(1-\mu_0)\,\phi}\sum_{t=1}^{\infty}\lambda^t\left(\prod_{s=2}^{t}F_s\right)\left\{\begin{array}{c} F_{t+1}\left\{u_{t+1}\left[\mu_0+(1-\mu_0)\,\phi^{t+1}\right]-\mu_0\right\}- \\ -(1-\mu_0)\,\phi^{t+1}\int_0^{\hat{x}_{t+1}} xf(x)dx \end{array}\right\}. \tag{37}$$

Comparing this expression with the FOC for $\hat{x}_1$,

$$\hat{x}_1^* = \frac{g_1}{(1-\mu_0)\,\phi} + \sum_{t=1}^{\infty}\lambda^t\left(\prod_{s=2}^{t}F\left(\hat{x}_s\right)\right)\left\{F\left(\hat{x}_{t+1}\right)\frac{g_{t+1}}{(1-\mu_0)\,\phi} - \phi^t\int_0^{\hat{x}_{t+1}} xf(x)dx\right\}. \tag{38}$$

tells us that $\hat{x}_1^{*p} = \hat{x}_1 + N_u$, implying that punishment N_u achieves a reduction in wrongdoing equal to $\phi\left(F\left(\hat{x}_1+N_u\right)-F\left(\hat{x}_1\right)\right)$ because current punishment raises the optimal cutoff of the unaware one for one in period 1. The cutoff solutions in equation (7) tell us that the cutoffs for all periods following the first depend on the static payoffs in each respective period, and on the continuation payoffs that depend on yet future cutoffs. Because punishment applies only to the current period, the cutoffs $\{\hat{x}_2^p, \hat{x}_3^p, ...\}$ are just like in the original problem. This does not mean however that one time punishment does not affect wrongdoing in future periods. But it does mean that the only effect that current punishment has on future wrongdoing is through its impact on the share of unaware individuals who resist and enter the future unaware. Specifically, of those who are saved from temptation in the current period, $\phi\lambda F\left(\hat{x}_2\right)$ are saved again in period 2, so $\phi\lambda F\left(\hat{x}_2\right)$ is the reduction of wrongdoing in period 2 as a result of punishment N_u having been present in period 1. Next, $\left(\phi\lambda\right)^2 F\left(\hat{x}_2\right)F\left(\hat{x}_3\right)$ are saved in period three, and so on. As a result, the one time punishment N_u leads to an expected wrongdoing reduction equal to $\phi\left(F\left(\hat{x}_1+N_u\right)-F\left(\hat{x}_1\right)\right)\left[1+\phi\lambda F\left(\hat{x}_2\right)+\left(\phi\lambda\right)^2 F\left(\hat{x}_2\right)F\left(\hat{x}_3\right)+...\right]$. And because the planner discounts future reductions in crime according to the factor δ, we obtain the expression in the lemma.■

Proof of Proposition 8: The first-order conditions for N_a and N_u are,

$$\phi f\left(N_a\right) - c'\left(N_a+N_u\right) = 0, \tag{39}$$

$$\phi f\left(\hat{x}_t+N_u\right)Z_t - c'\left(N_a+N_u\right) = 0. \tag{40}$$

Solving for $c'\left(N_a+N_u\right)$ and combining yields $f\left(N_a\right) = f\left(\hat{x}_t+N_u\right)Z_t$. Note from (18) that Z_t approaches 1 as λ or δ approach zero. Recall that $\hat{x}_1 > 0$. Therefore, in the neighborhood of $Z_t = 1$, $f\left(\hat{x}+N_u\right)$ is arbitrarily close to $f\left(N_a\right)$, which yields $N_u \simeq N_a - \hat{x}_t$ and hence $N_u < N_a$.

The second order conditions are,

$$\begin{aligned}
H_{11} &= f'(N_a) - c''(N_a + N_u) < 0, \\
H_{22} &= f'(N_u + \hat{x}_t) - c''(N_a + N_u) < 0, \\
|H| &= H_{11}H_{22} - H_{12}H_{21} = \\
&= (f'(x) - c'')(f'(y+k) - c'') - c''^2 = f'(x) f'(y+k) - c'' f'(x) - c'' f'(y+k) > 0.
\end{aligned}$$

The convexity of costs and the assumption $f'(x) < 0$ guarantee that these expressions are satisfied.■

Proof of Proposition 9: Using expression (8), the CRRA utility function, and considering a horizon of just one period, one obtains the optimal policy in the one-period case $\hat{x}^* = \frac{\mu_0}{\phi(1-\mu_0)}\left(\frac{\hat{\mu}_1^{1-\rho}}{\hat{\mu}_1} - 1\right)$. We now drop the star from the notation, so that $\hat{x}$ stands for the optimal cut-off. Notice that $\hat{x}$ is increasing in μ and ρ but independent of θ, and that $\lim_{\mu \to 1} \hat{x}(\mu) = \rho$.

The expected utility of an individual with belief μ going to a profession with mean temptation θ is

$$\begin{aligned}
V(\mu, \theta) &= F(\hat{x}|\theta)\left([\mu + (1-\mu)\phi]\, u(\hat{\mu}) + (1-\mu)(1-\phi)E[x|x < \hat{x}, \theta]\right) \\
&\quad + (1 - F(\hat{x}|\theta))\left(\mu u(1) + (1-\mu) E[x|x \geq \hat{x}, \theta]\right) \\
&= F(\hat{x}|\theta)\left([\mu + (1-\mu)\phi]\, u(\hat{\mu}) - \mu\right) + \mu \\
&\quad + (1-\mu)\left(\theta - \phi \int_0^{\hat{x}} x f(x|\theta)\, dx\right) \\
&= F(\hat{x}|\theta)\, \mu\left(\hat{\mu}^{-\rho} - 1\right) + \mu + (1-\mu)\left(\theta - \phi \int_0^{\hat{x}} x f(x|\theta)\, dx\right). \qquad (41)
\end{aligned}$$

The distribution with higher temptations is defined in terms of first order stochastic dominance, so $F_\theta(x|\theta) < 0$. Recall that $\hat{x}$ is independent of θ. Denote the mean temptation in the two careers by $\theta_H > \theta_L > 0$. The compensating differential for type μ for entering the low-temptation career is

$$\begin{aligned}
V(\mu, \theta_H) - V(\mu, \theta_L) &= \left(F(\hat{x}|\theta_H) - F(\hat{x}|\theta_L)\right)\mu\left(\hat{\mu}^{-\rho} - 1\right) + (1-\mu)(\theta_H - \theta_L) \qquad (42) \\
&\quad - (1-\mu)\phi \int_0^{\hat{x}} x\left[f(x|\theta_H) - f(x|\theta_L)\right] \mathrm{d}x.
\end{aligned}$$

Now hold any $\theta_L > 0$ as fixed and consider the difference $V(\mu, \theta_H) - V(\mu, \theta_L)$. To prove the proposition it suffices to show that this difference is decreasing in μ because then, for any $\theta_H > \theta_L$, the compensating differential required to attract individuals into the low-temptation sector is decreasing in μ. Denote $H(\mu) \equiv \mu\left(\hat{\mu}^{-\rho} - 1\right)$. Noting that the envelope

theorem helps us eliminate all terms involving $\hat{x}'(\mu)$, the differentiation of (42) with respect to μ yields

$$V_\mu(\mu,\theta_H) - V_\mu(\mu,\theta_L) = \tag{43}$$
$$(F(\hat{x}|\theta_H) - F(\hat{x}|\theta_L))H'(\mu) - (\theta_H - \theta_L) + \phi\int_0^{\hat{x}} x\left[f(x|\theta_H) - f(x|\theta_L)\right]\mathrm{d}x.$$

Using integration by parts to transform $\int_0^{\hat{x}} xf(x|\theta)\mathrm{d}x = \hat{x}F(\hat{x}|\theta) - \int_0^{\hat{x}} F(x|\theta)\mathrm{d}x$ then (43) becomes

$$(F(\hat{x}|\theta_H) - F(\hat{x}|\theta_L))(H'(\mu) + \phi\hat{x}) - (\theta_H - \theta_L) - \phi\int_0^{\hat{x}}\left[F(x|\theta_H) - F(x|\theta_L)\right]\mathrm{d}x.$$

The first term of (6) is negative if $H'(\mu) + \phi\hat{x}$ is positive. And since $\partial\hat{\mu}/\partial\mu = \phi(\hat{\mu}/\mu)^2$ we can write

$$H'(\mu) = \frac{\partial}{\partial\mu}\left[\mu\left(\hat{\mu}^{-\rho} - 1\right)\right] = \hat{\mu}^{-\rho} - 1 - \rho\mu\hat{\mu}^{-\rho-1}\frac{\partial\hat{\mu}}{\partial\mu} \tag{44}$$
$$= \hat{\mu}^{-\rho} - 1 - \rho\mu\hat{\mu}^{-\rho-1}\phi\left(\frac{\hat{\mu}}{\mu}\right)^2 = \hat{\mu}^{-\rho}\left(1 - \rho\phi\frac{\hat{\mu}}{\mu}\right) - 1. \tag{45}$$

Thus

$$H'(\mu) + \phi\hat{x} = \left[\hat{\mu}^{-\rho}\left(1 - \rho\phi\frac{\hat{\mu}}{\mu}\right) - 1\right] + \phi\left[\frac{\mu}{(1-\mu)\phi}\left(\hat{\mu}^{-\rho} - 1\right)\right] \tag{46}$$
$$= \left(\frac{1}{1-\mu}\right)\left[\hat{\mu}^{-\rho}\left(\frac{\mu + (1-\rho)(1-\mu)\phi}{\mu + (1-\mu)\phi}\right) - 1\right]. \tag{47}$$

This is always positive if

$$\frac{\mu + (1-\rho)(1-\mu)\phi}{\mu + (1-\mu)\phi} > \left(\frac{\mu}{\mu + (1-\mu)\phi}\right)^{\rho}, \tag{48}$$

which is implied by equation (29).■

6.1 The role of risk aversion

Here we include a formal explanation of the role of risk aversion in Proposition 1. To see make things simple, consider an individual that lives only for one period and who faces a temptation x. We first verify that expected beliefs after each possible intent are the same, and then examine expected payoffs.

Selecting no intent to resist means that the agent will only resist temptation if she is truly good, and because good types can only resist, her action will fully reveal her type.

Therefore, the expected utility from selecting a negative intent is $1 \times \mu_0 + 0 \times (1 - \mu_0) = \mu_0$, which is the same as the prior. Selecting a positive intent means that if she is bad but lucky to have free will she will also see herself pass on the temptation. Therefore, seeing herself resisting will be compatible both with a good type, and with a bad type who, having selected a positive intent, was lucky. The posterior she will have then is $\hat{\mu}_1 = \mu_0 / (\mu_0 + (1 - \mu_0)\phi)$. Seizing the temptation is only compatible with a bad type who, having selected a positive intent, was unlucky. The expected belief when selecting a positive intent is then $[\mu_0 + (1 - \mu_0)\phi] \times \mu_1 + (1 - \mu_0)(1 - \phi) \times 0 = \mu_0$, again the prior, as expected. So we have verified what we knew to be true from the martingale property of beliefs: expected beliefs cannot be affected by one's intent. Now we examine expected payoffs. Lack of intent buys the agent a lottery that generates a prize $u(1)$ with probability μ_0 and a prize $u(0) + x$ with probability $(1 - \mu_0)$. Selecting a positive intent buys her a lottery that yields a prize $u(\hat{\mu})$ with probability $\mu_0 + (1 - \mu_0)\phi$ (i.e., in the event that she is good, or in the event when she is bad, but, having selected a positive intent, is lucky and has free will determining a good action), and a prize $u(0) + x$ with probability $(1 - \mu_0)(1 - \phi)$. Selecting a positive intent is optimal if and only if,

$$[\mu_0 + (1 - \mu_0)\phi] u(\hat{\mu}) - \mu_0 u(1) > \phi(1 - \mu_0) x,$$

which in turn requires that,

$$\mu_0 \left(\frac{\hat{\mu}_1^{1-\rho}}{\hat{\mu}_1} - 1 \right) > \phi(1 - \mu_0) x. \tag{49}$$

This expression says that the agent, when selecting positive intent, decides to forgo the temptation x in case she is bad and has free will (which has probability $\phi(1 - \mu_0)$) in order to obtain a utility gain $\mu_0 (u(\hat{\mu}_1) / \hat{\mu}_1 - 1)$ in terms of thinking better or herself in that same instance. However, in case she is good she will now only enjoy a posterior equal to $\hat{\mu}_1$ rather than to 1. Therefore, the net utility gain, which is measured by $\mu_0 (u(\hat{\mu}_1) / \hat{\mu}_1 - 1)$, is exactly zero when the agent is risk neutral (i.e., when $\rho = 0$), because expected beliefs are invariant in the agent's intent. The utility gain from beliefs is only positive when the agent is risk averse. Note that the agent, although not improving her expected beliefs, is reducing the variance of such beliefs. When selecting no intent the variance over beliefs is $E(\hat{\mu}_1 - E(\hat{\mu}_1))^2 = \mu_0 (1 - \mu_0)$, but when selecting a positive intent that variance becomes smaller and equal to $\mu_0 (1 - \mu_0)(1 - \phi) \mu_0 / [\mu_0 + (1 - \mu_0)\phi]$.

The optimal cutoff in the one period problem is obtained by solving x from the equality

corresponding to (49):

$$\hat{x}^* = \frac{\mu_0}{\phi\left(1-\mu_0\right)}\left(\frac{\hat{\mu}_1^{1-\rho}}{\hat{\mu}_1} - 1\right). \tag{50}$$

It is easy to see that for any degree of risk aversion, as parameterized by $\rho > 0$, there is a positive cutoff $\hat{x}$ such that the agent will prefer to pass on temptations below that level because she prefers a lottery between beliefs zero and $\hat{\mu}_1$ (with an increased probability to get $\hat{\mu}_1$) rather than a lottery between beliefs 0 and 1. It is also easy to see from (50) that higher degrees of risk aversion will be associated with higher cutoffs: Individuals who are more averse to learning their type will be willing to forgo higher temptations.

References

Akerlof, G. and R. Kranton (2000), Economics and Identity, *Quarterly Journal of Economics* 115 (August), 715-53.

Aristotle (1998), Nichomachean Ethics, Dover.

Becker, G. (1968), Crime and Punishment: An Economic Approach, *Journal of Political Economy* 76(2), 169-217.

Bénabou, R. and J. Tirole (2004), Willpower and Personal Rules, *Journal of Political Economy* 112, 848-886.

Bénabou, R. and J. Tirole (2006), Incentives and Prosocial Behavior, *American Economic Review* 96(5), 1652-1678.

Bénabou, R. and J. Tirole (2007), Identity, Dignity and Taboos: Beliefs as Assets, *IZA discussion paper* 2583.

Bernheim, D. and A. Rangel (2004), Addiction and Cue-Triggered Decision Processes, *American Economic Review* 94(5), 1558-1590.

Brekke, K., S. Kverndokk, and K. Nyborg (2003), An Economic Model of Moral Motivation, *Journal of Public Economics* 87, 1967-1983.

Brocas, I. and J. Carrillo (2008), The Brain as a Hierarchical Organization, *American Economic Review* 98(4), 1312-1346.

Carrillo, J. and T. Mariotti (2000), Strategic Ignorance as a Self-Disciplining Device, *Review of Economic Studies* 67(3), 529-544.

Cervellati, M., J. Esteban and L. Kranich (2006), The Social Contract With Endogenous Sentiments, mimeo Institut d'Anàlisi Econòmica.

Clark, J., J. Austin and A. Henry (1997), Three Strikes and You're Out: A Review of State Legislation, National Institute of Justice Research in Brief Series (September), Department

of Justice of the United States.

Compte, O. and A. Postlewaite (2004), Confidence-Enhanced Performance, *American Economic Review* 94(5), 1536-1557.

Fehr, E. and S. Gächter (2002), Altruistic Punishment in Humans, *Nature* 415, 137-140.

Fiske, A. and P. Tetlock (1997), Taboo Trade-offs: Reaction to Transactions that Transgress the Spheres of Justice, *Political Psychology* 18, 255-297.

Fisman, R. and E. Miguel (2006), Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets, forthcoming *Journal of Political Economy.*

Fudenberg, D. and D. Levine (2006), A Dual-Self Model of Impulse Control, *American Economic Review* 96(5), 1449-76.

Gneezy, U. (2005), "Deception: The role of consequences," *American Economic Review*, 95(1), 384-394.

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

Hays, S. (1981), The Psychoendochrinology of Puberty and Adolescent Aggression. In Hamburg, D. and M. Trudeau (eds.) Biobehavioral aspects of aggression, Alan Liss Inc. New York.

Heinrich, J. and N. Smith (2004), Comparative Experimental Evidence From Machiguenga, Mapuche, Huinca, and American Populations, in Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (eds.), Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press.

Hermalin, B. and A. Isen (2008), A Model of the Effect of Affect on Economic Decision Making, *Quantitative Marketing and Economics* 6, 17-40.

Kaplow, L., and S. Shavell (2007), Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System, *Journal of Political Economy* 116(3), 494-514.

Kolm, S-Ch. (2004), Modern theories of justice. MIT Press.

Kőszegi, B. (2006), Ego-Utility, Overconfidence, and Task Choice, *Journal of the European Economic Association* 4(4), 673-707.

Loewenstein, G. (1996), Out of Control: Visceral Influences on Behavior, *Organizational Behavior and Human Decision Processes* 65(3), 272-92.

Nagin, D. and R. Paternoster (1993), Enduring Individual Differences and Rational Choice Theories of Crime, *Law & Society Review* 27(3), 467-496.

Ostrom, E., J. Walker, and R. Gardner (1992). Covenants With and Without a Sword: Self-Governance is Possible, *American Political Science Review* 86(2), 404–17.

Polinsky, M. and D. Rubinfeld (1991), A Model of Optimal Fines for Repeat Offenders, *Journal of Public Economics* 46(3), 291-306.

Polinsky, M. and S. Shavell (1998), On Offense History and the Theory of Deterrence, *International Review of Law and Economics* 18(3), 305-324.

Prelec, D. and R. Bodner (2003), Self-Signaling and Self-Control, in Loewenstein, G., D. Read and R. Baumeister (eds.) Time and Decisions. Russell Sage Foundation.

Rabin, M. (1994), Cognitive Dissonance and Social Change, *Journal of Economic Behavior and Organization* 23, 177-194.

Sussman, E., G. Inoff-Germain, E. Nottelmann, and D. Loriaux (1987), Hormones, Emotional Dispositions, and Aggressive Attributes in Young Adolescents, *Child Development* 58(4), 1114-1134.

Tabellini, G. (2007), The Scope of Cooperation: Values and Incentives, mimeo Bocconi.

Tirole, J. (1996), A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality), *Review of Economic Studies* 63(1), 1-22.

Weber, M. (2002 [1905]), The Protestant Ethic and the Spirit of Capitalism, Penguin.

Figure 1: Timeline for period *t*