Self-Esteem, Moral Capital, and Wrongdoing

Ernesto Dal Bó Marko Terviö*

February 25, 2008

Abstract

In order to help understand adherence to moral standards and the force of intrinsic motivation, we present an infinite-horizon model where an individual receives random temptations (such as bribe offers) and must decide which to resist. Temptations yield consumption value, but keeping a good self-image (a high belief of being the type of person that resists) yields self-esteem. Individual actions depend both on types and intent, so selecting a good intent does not guarantee good behavior and past resistance is informative of a good type. 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. 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 helps rationalize taboos, harsher punishment of repeat offenders, and a tendency of individuals with low moral capital to enter high-temptation activities.

Keywords: moral capital, intrinsic motivation, wrongdoing, moral costs, self-esteem, corruption, crime.

JEL codes: D83, K4, Z13.

[*] Ernesto Dal Bó: Stanford Graduate School of Business and NBER (edalbo@stanford.edu). Marko Terviö: Haas School of Business, UC Berkeley (marko@haas.berkeley.edu). We thank Roland Bénabou, Jeremy Bulow, Pedro Dal Bó, Erik Eyster, Botond Kőszegi, Keith Krehbiel, John Morgan, Santiago Oliveros, Matt Rabin, Tim Williamson, and seminar participants at Berkeley, Princeton, Stanford, and Universidad de San Andrés for useful conversations and comments. Juan Escobar provided excellent research assistance.

1 Introduction

We propose a theory of moral standards, understood as the inclination to adhere to some given principles of action at the cost of consumption-based utility. This can be seen as a dynamic theory of intrinsic motivation based upon two ideas: the first is that individuals obtain utility from their self-esteem, and the second is that an individual's self-esteem depends on his or her self-image, which in turn is affected by acts of the individual.

The idea that individuals may behave morally because they want to think of themselves as moral goes back at least to Adam Smith's Theory of Moral Sentiments, and has been highlighted in an important modern literature that we discuss below. The objective of our paper is two-fold. One is to provide a model that allows for the characterization of moral behavior in a fully dynamic setting, to provide a rationalization of Calvinist ethics, and a rationalization of why temporary economic shocks may lead to persistent increases in wrongdoing. The other objective is to enrich the analysis of crime and corruption (wrongdoing) from the perspective of a model that explicitly accounts for endogenous moral standards. Along these lines, our model can (i) help interpret the variation on crime and corruption across countries in relation to differences in self-image that have been caused by different national histories rather than by differences in deep moral fundamentals, (ii) help rationalize social regularities such as the presence of taboos and enhanced punishments for repeat offenders, (iii) allow an analysis of how low levels of moral capital may give some individuals a comparative advantage at entering into high temptation activities.

In our model, individuals face a sequence of consumption opportunities ("temptations") of random magnitudes. Individuals may have a good or a bad type, where good types invariably 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 can get tarnished by committing acts that deviate from what good types would do. Individuals maximize utility over an infinite horizon and have standard time preferences (exponential discounting). We identify conditions under which (i) individuals are interested in resisting actions that are deemed immoral even when they yield consumption value, (ii) individuals improve their self-image by resisting immoral actions, and (iii) an improvement in individual self-image leads to a stronger inclination to resist immoral actions. These results are not obvious nor necessary. Specific conditions must be met, and identifying these conditions illuminates the problem of why a person would want to develop a good self-image at all. After all, a life of mischief (unconstrained maximization of consumption utility) may be rewarding, too.

A crucial feature of moral behavior that is driven by an introspective reputation is that events that damage this reputation will end up lowering future incentives to act morally, leading to a self-reinforcing path of wrongdoing. 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 moral capital, such as taking a bribe. As a result of the damage to her self-image, the individual will have less of an incentive to behave morally after economic conditions have returned to normality.

Another important aspect of our model is the possibility of moral growth. The individual does not directly select her action, but her intent. Actions are not affected by intent deterministically, because circumstantial factors may alter effective actions that are exogenous to an individual's conscious designs. The stochastic shock affecting the map between intentions and actions is taken to be external to the individual (as external conditions that make it harder to resist a given temptation) or internal to her, as when a temporary organic disposition alters the ability to control a visceral impulse. A substantial literature in psychology has documented the role of visceral impulses and unconscious bias in decision-making (see, e.g., Loewenstein 1996, and Bernheim and Rangel 2004 for a model of addiction rooted in the neuroscience of impulse control). We will refer to the nondeterministic connection between intent and action as "imperfect free will" throughout the paper.[1] In each period the individual may or may not have free will; the realization of free will is random and unknown to the individual. In a period when she has free will, the individual's chosen intent will indeed determine her action, but in periods when she lacks free will her action will be determined by her type. In this way, the observation of her actions will be informative about her type.[2] As a result, past actions reveal information about a person's type. We will show that a history of good behavior—which is will be more likely for those who have the "luck" of facing low temptations—leads to improvements in the individual's self-image, which in turn strengthens the inclination to resist temptations. Self-image, then, is costly to improve (it requires forgoing temptations) and it enters the future "production" of good actions, so it works as a form of capital, which we call moral capital. The model accounts for the emergence of morality as a cumulative process of habituation through action, which parallels Aristotle's account of the attainment of virtue.

The model offers comparative statics results showing that individuals who are more patient and more confident about their ability to transform intentions into actions will attempt to resist more temptations. Thus, the personal traits that criminologists associate with

[1] The ability to transform intentions into actions could also naturally be associated with the idea of self-control. In economics the idea of self-control is mainly related to time-inconsistent preferences (which are absent from our model) while in criminology it is thought to encompass various traits, from pure impulse-control abilties to impatience. We keep these separate in our model.

[2] In a setup where the probability of free will goes to one, individuals may still resist temptations, but the history of actions does not affect moral dispositions.

higher self-control lead in our model to higher chances of avoiding wrongdoing, explaining interpersonal differences in criminal trajectories (see Gottfredson and Hirschi 1990). But our model allows for another explanatory factor for those differences: given personal traits, good behavior weeds out bad behavior and vice versa.

We extend the analysis of our one person model to consider a society populated of all age groups in demographic steady state. Wrongdoing across countries may differ even when the "deep" morality fundamentals in terms of proportions of good and bad types are the same. The cross-country wrongdoing rate differentials will reflect different historical fortunes in terms of the size of the temptations that have accrued to different societies. Moral capital at the level of a society does not merely depend on the average individual moral capital, but on its dispersion.

We provide a rationalization for certain types of taboo as a way to build moral capital. The function of taboo is to place restrictions that are not prohibitively costly to respect, and that, when respected, increase the introspective reputation of individuals, strengthening their subsequent inclination to avoid immoral actions.[3] The model can also be used to rationalize why repeat offenders face enhanced sentences (as is the case in most states in the US). A record of wrongdoing is associated with lower moral capital and lower intrinsic incentives to avoid future wrongdoing, which demands stronger extrinsic deterrents from a relatively impatient planner who wants to minimize wrongdoing. Furthermore, we consider activities characterized by different distributions of temptations, and show that individuals with low moral capital will sort themselves into the activities with relatively high expected temptations. If an activity like politics is seen as a high temptation activity, the disturbing implication is that those who do not think they have much to lose by way of introspective reputation will be more inclined to entering it. This would lead politics to display high levels of wrongdoing for two reasons: it has higher temptations, and it attracts the people who have the lowest intrinsic motivation to resist temptations.

Our setup offers a quite literal formalization of Weber's account of the Calvinist Ethic. According to that account, individuals are born saved or damned, but do not know their predestination status. Given that uncertainty, the account goes, individuals perform good deeds and resist mundane temptations. Doing good is a way for individuals to convince themselves that they were born saved. In other words, 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? And even if there is a way to improve one's confidence in one's type by doing good, a second question emerges: would a utility maximizing individual pass on enjoyable temptations in order to hold better

[3] For a different conception of taboos, see Fiske and Tetlock (1997) and Bénabou and Tirole (2007).

beliefs about herself? Our model answers these questions, which seems interesting in general, and in particular in the context of the Weberian account of the Calvinist Ethic, which has been called, even by critics, "the best known theory ever propounded by any sociologist" (Rubinstein, 1999). We find that the key condition for the individual to want to maintain a good self-image by choosing to resist temptations is that she be risk averse in terms of her self-image.[4] The possibility of developing a good self-image that improves over time requires, as said earlier, imperfect free will.

We do not attempt to explain the content of morality. We take it as given that individuals accept the structure of the model, where types who don't give in to temptations are deemed good, and where there is a self-esteem motivation to try to appear good to oneself. The content of morality may follow from evolutionary forces, and be transmitted by culture and parental authority. We study the determination of moral standards, seen as the degree of adherence to established moral principles. Why do this? We believe that moral standards are both endogenous and important for behavior. Many models in economics and politics give a prominent role to wrongdoing, seen as the commission of illegal or socially abusive acts. Examples include the analysis of corruption, tax evasion, crime, as well as the study of political influence. Such theories predict that the equilibrium level of wrongdoing will depend on the associated costs and benefits. The benefits are usually construed to follow from material gains, while the costs typically follow from potential sanctions and, often, an intrinsic distaste for doing wrong–the so-called "moral cost."[5] Economic theories have analyzed in detail how the costs and benefits of wrongdoing are affected by extrinsic incentive structures, while in contrast very little attention has been devoted to the determinants of moral costs.

One reason for this omission may be that moral costs are nonexistent or incomprehensible. However, 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.[6] Moreover, there is a "revealed preference" argument for the idea that moral costs are both important

[4] Different rationalizations of the Calvinist ethic are possible. The self-signaling setup of Bodner and Prelec (2003) can be used in such a way, but, as we explain below, self-signaling models differ from ours in important ways.

[5] A classic reference in the economic theory of crime is Becker (1968). His model considered an exogenous parameter u capturing the individual's general disposition to commit crimes. Many posterior models of crime and corruption share this feature.

[6] Experimental evidence indicates that people have a preference for avoiding telling lies (Gneezy 2005), and for imparting justice in the form of punishment against those who "misbehave" (Fehr and Gächter 2002). Considerations of fairness appear to vary across cultures, and affect behavior in settings where subjects have discretion to determine the distribution of resources (Heinrich and Smith 2004). Fisman and Miguel (2007) present evidence supporting the idea that traffic violations respond to cultural norms even when individuals share the same environment.

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, thus shaping desirable intrinsic motivations. This makes it worthwhile to attempt to enrich our understanding of wrongdoing from the perspective of a model where moral standards emerge endogenously.

The structure of the paper is as follows. The next section discusses 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. These include a rationalization of taboo, of harsher penalties on repeat offenders, and an analysis of why individuals with low moral capital would tend to prefer high temptation activities. Section 6 concludes.

2 Related literature

Two interesting exceptions to the absence of work on the shaping of moral standards are Kaplow and Shavell 2007, and Tabellini 2007. Kaplow and Shavell focus on the relative convenience of investing in instilling guilt and virtue versus using incentives to induce good behavior, while Tabellini studies investments in the transmission of cooperative values in an overlapping generations framework. Both studies address important aspects, but both abstract from the internal process that makes individuals want to adhere to received moral standards. In their models adherence to values responds directly to a given investment in their inculcation. We want to examine the degree of adherence to moral standards by linking good behavior to a self-discovery process where self-esteem plays a central role. Our work is not the first to study the development of discipline devices associated with the manipulation of beliefs - these features have precedent in an important literature, the vast majority of which involves hyperbolic discounting. In that context, the manipulation of beliefs serves an instrumental purpose, namely helping to overcome time-inconsistency. Important references in this line of work are Carrillo and Mariotti (2000) and Bénabou and Tirole (2004).[7] There is another strand of work that does not rely on time-inconsistent preferences and where the motivation for manipulating beliefs can be interpreted as intrinsic, as in our model. Within this strand, Prelec and Bodner (2003) propose the concept of "diagnostic self-signaling" to

[7] Another interesting instance of a model where beliefs are manipulated for instrumental reasons is due to Compte and Postlewaite (2004). In their model, an individual wants to stay optimistic because such psychological state will improve her future performance at a given task. Hermalin and Isen (2008) offer a model the where mood affects the choice of actions and viceversa, leading to potential multiple equilibria in individual behavior.

capture the phenomenon of an individual that attempts to signal to herself the possession of a desirable trait. Bénabou and Tirole (2006) study pro-social behavior as a signal (to oneself or to others) of the possession of pro-social preferences, and Bénabou and Tirole (2007) analyze the possibility that one may want to signal to oneself a strong preference for the type of assets one has accumulated, leading to forms of "excessive" behavior. All of these papers contain an important degree of intrapersonal conflict modeled as a non-cooperative game among different selves.[8] In contrast, our model features an individual that contains a single self, and optimal plans are time-consistent. Moreover, we characterize the full dynamics of individual behavior over an infinite horizon, while most of the pre-existing literature considers two- or three-period settings. This is important in relation to our analysis of the accumulation of moral capital, as finite horizon settings will introduce effects confounding the pure link between past good behavior and the incentives to continue to behave well.

Intrinsic motivation and the need to manipulate beliefs are strongly connected, as made clear in the literature on cognitive dissonance, which has provided ample evidence that the need to preserve a good image about self affects behavior. In such connection, Rabin (1995) offers a model where agents face exogenous moral constraints and engage in belief manipulation in order to mitigate the impact of such constraints in the pursuit of self-interest, thus explaining self-serving biases in the collection of information. The notion of self-image, which we rely upon, is key to the cognitive dissonance perspective. 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 and Cervellati, Esteban and Kranich (2006) in their model of moral sentiments and redistribution. Their conceptualization of self-image is however very different from ours, which follows Kőszegi's (2006) formulation of ego-utility. Kőszegi studies the emergence of overconfidence and the choice of tasks – he isolates conditions under which an agent may engage in an "ambitious" task depending on whether information on her type is welcome to the agent or not. The demand for information about self also plays a crucial role in our model, but our focus is on moral dispositions, not on the emergence of overconfidence. Also in Kőszegi's model the agent prefers to think she is of a type for whom extrinsic payoffs are higher, while in our model the opposite holds.

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

[8] Brocas and Carrillo (2005) and Fudenberg and Levine (2006) also adopt an explicitly non-cooperative approach to modeling intra-personal conflict in dynamic settings.

her, and she is born with an initial belief that she is good with probability μ_0. She has two 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 of 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 a harsher 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" 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. The individual has free will with probability $\phi \in (0, 1)$, drawn independently each time. 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.[9] The formulation we use resonates also with a long tradition in philosophy regarding the limits of free will.[10] The role of imperfect free will in

[9] There is a large literature in psychology emphasizing the impact of such impulses. And a growing literature in economics has incorporated insights from psychology and neuroscience to model personality as a result of an interplay of conscious and unconscious factors. On the precise issue of visceral impulses, see for example Loewenstein (1996).

[10] Kolm (1996, p. 38) characterizes the agent by saying that "An *agent* can perform *actions* and is endowed

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.

Besides the utility from temptations, the individual derives utility from self-esteem. The individual with belief μ_t will enjoy a self-esteem $u(\mu_t)$ in period t. We assume that

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

where $\rho \in [0, 1)$ is the coefficient of 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, reasons that are not connected to outcomes, but to the experience of living with a certain degree of self-worth as a result of confidence about one's attributes.

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 born with the initial belief μ_0, an individual who has successfully resisted t times remains unaware and has the belief

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

Even when knowing that one has selected a positive intent, beliefs about one's goodness improve when seeing oneself do good. 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.

The sources of self-esteem are outside our model: The question of why people derive utility from thinking that they share the type of those who reject temptations is beyond the scope of our enquiry. We do not set out to explain why the rejection of temptations, the trademark action of good types, is considered a good thing. Therefore, our account of the endogeneity of moral standards only covers their degree of stringency, and not the qualitative feature of why certain types of actions are morally condemned. In this regard, we want to use the temptations to capture opportunities for gain that, although enjoyable,

with a *will.* A will has *intention* and can influence–determine in part, be a cause of–certain *acts* of this agent... An agent's *action* is more *free* when it is more *caused by his will.* The other causes of an action constitute the *constraints* on it and determine the corresponding *domain of liberty*, or of *possibles* or of *choice.*" (Italics in the original).

would damage the individual's self-esteem. Our model may be seen as capturing self-control issues, and indeed we believe wrongdoing may be related to self-control. However, we do not attempt to explain self-control phenomena in general, but rather examine wrongdoing through a specific lens. This is the lens of the interactions between a source of self-esteem, however originated, and the propensity to commit acts that yield utility but damage that self-esteem.

3.1 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, and when we obtain the solution later we show that the optimal policy must indeed have such a form.

To set up the expected lifetime utility as a function of the cutoffs, it is useful to first consider the contribution of just one generic future period t. (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 previous to period 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.

We can now move towards writing the expected utility from a generic period $t \geq 1$ as perceived at the beginning of period 1, before the realization of x_1.

An agent who is aware of being good will enjoy the self-esteem rewards of her certainty, with value $u(1) = 1$. 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 $u\left(\hat{\mu}_t\right)$. Conditional on being good (which has prior probability μ_0), the two relevant states have probability

$$\begin{aligned}\Pr(\text{unaware}|\hat{x}_1,\ldots,\hat{x}_t) &= H_t(\hat{x}_1,\ldots,\hat{x}_t) &(4)\\ \Pr(\text{aware}|\hat{x}_1,\ldots,\hat{x}_t) &= 1-H_t(\hat{x}_1,\ldots,\hat{x}_t). &(5)\end{aligned}$$

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. However, someone who finds out in period t that she is bad will obtain a different expected utility depending on the circumstances. One possibility is that she faces a temptation above her cutoff $\hat{x}_t$, does not attempt to resist and sees herself fall for 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, these four alternatives have probabilities given by,

$$\begin{aligned}\Pr(\text{unaware}|\hat{x}_1,\ldots,\hat{x}_t) &= \phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t) &(6)\\ \Pr(\text{aware before}|\hat{x}_1,\ldots,\hat{x}_t) &= \left[1-\phi^{t-1}H_{t-1}(\hat{x}_1,\ldots,\hat{x}_{t-1})\right] &(7)\\ \Pr(\text{newly aware, high }x|\hat{x}_1,\ldots,\hat{x}_t) &= \phi^{t-1}H_{t-1}(\hat{x}_1,\ldots,\hat{x}_t)\,[1-F(\hat{x}_t)]. &(8)\\ \Pr(\text{newly aware, low }x|\hat{x}_1,\ldots,\hat{x}_t) &= \phi^{t-1}H_{t-1}(\hat{x}_1,\ldots,\hat{x}_t)\,[F(\hat{x}_t)(1-\phi)] &(9)\end{aligned}$$

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

$$\begin{aligned} EU_t &= \mu_0 \left[H_t u\left(\mu_t\right) + \left(1-H_t\right) u\left(1\right)\right] \\ &+ \left(1-\mu_0\right) \begin{pmatrix} \phi^t H_t u\left(\hat{\mu}_t\right) + \\ \left[1-\phi^{t-1} H_{t-1}\right] Ex + \\ \left(1-\phi\right) \phi^{t-1} H_{t-1} F\left(\hat{x}_t\right) E[x|x<\hat{x}_t] + \\ \phi^{t-1} H_{t-1}\left(1-F\left(\hat{x}_t\right)\right) E[x|x\geq \hat{x}_t] \end{pmatrix}. \end{aligned} \tag{10}$$

The sequence of utilities conditional on remaining unaware, $u\left(\hat{\mu}_1\right), u\left(\hat{\mu}_2\right), \ldots$, is just a known increasing sequence of numbers that converges to $u\left(1\right)$, hence we denote these numbers as u_t. Summing up and discounting the expected utilities (10) from all periods $t = 1, 2, \ldots$ gives (after rearrangement) the individual objective function

$$\begin{aligned} V_0\left(\hat{x}_1, \hat{x}_2, \ldots\right) &= \sum_{t=1}^{\infty} \lambda^{t-1} EU_t = \frac{\mu_0 u\left(1\right) + \left(1-\mu_0\right) Ex}{1-\lambda} + \\ &+ \sum_{t=1}^{\infty} \lambda^{t-1} \left\{ \begin{array}{c} H_t u_t \left[\mu_0 + \left(1-\mu_0\right)\phi^t\right] \\ -\mu_0 H_t u\left(1\right) - \left(1-\mu_0\right)\phi^t H_{t-1} \int_0^{\hat{x}_t} x f(x) dx \end{array} \right\}. \end{aligned} \tag{11}$$

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 (11). The cutoff $\hat{x}_t$ gives the highest temptation that she will intend to resist in period t conditional on remaining unaware at 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

$$\begin{aligned} \frac{\partial V_0}{\partial \hat{x}_s} &= \lambda^{s-1} H_{s-1} f\left(\hat{x}_s\right) \left\{ u_s \left[\mu_0 + \left(1-\mu_0\right)\phi^s\right] - \mu_0 u\left(1\right) - \left(1-\mu_0\right)\phi^s \hat{x}_s \right\} + \\ &+ \frac{f\left(\hat{x}_s\right)}{F\left(\hat{x}_s\right)} \sum_{t=s}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F\left(\hat{x}_{t+1}\right) u_{t+1} \left[\mu_0 + \left(1-\mu_0\right)\phi^{t+1}\right] - \\ -F\left(\hat{x}_{t+1}\right)\mu_0 u\left(1\right) - \left(1-\mu_0\right)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} = 0. \end{aligned} \tag{12}$$

Substantially rearranging this condition yields the extremum

$$\hat{x}_s^* = \frac{g_s}{\left(1-\mu_0\right)\phi^s} + \sum_{t=1}^{\infty} \lambda^{t+s-1} \frac{H_{t+s-1}}{H_s} \left\{ \frac{F\left(\hat{x}_{t+s}\right) g_{t+s}}{\left(1-\mu_0\right)\phi^s} - \phi^t \int_0^{\hat{x}_{t+s}} x f(x) dx \right\}, \tag{13}$$

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

This last expression (13) 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 $\hat{x}_1^*$:

$$\hat{x}_1^* = \frac{g_1}{(1-\mu_0)\,\phi} + \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F(\hat{x}_s) \right) \left\{ F(\hat{x}_{t+1}) \frac{g_{t+1}}{(1-\mu_0)\,\phi} - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \right\}. \quad (14)$$

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 entering period t has the updated beliefs $\hat{\mu}_{t-1}$. Therefore $\hat{x}_1^*$ is identical to that of $\hat{x}_s^*$ except for 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'.

A number of important questions arise: Does the sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, ...$ constitute a maximum, and if so, is this maximum unique? Are any of those thresholds strictly positive? To get at these questions, we begin by stating a series of results that will allow us to characterize the optimal policy of the individual.

Lemma 1 *There is a unique sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, ..$ characterizing optimal behavior.*

Proof. See Appendix. ■

This means that the effects of any 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 2 *A sequence $\hat{x}_1^*, \hat{x}_2^*, \hat{x}_3^*, \ldots$ satisfying the FOCs is a global maximizer of V_0.*

Proof. See Appendix. ■

Lemma 3 *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. See Appendix. ■

The last lemma implies that there indeed exists an interior solution to the agent's problem, and therefore the FOCs do characterize the individual's optimal policy. This implies that the agent will only want to resist temptations for reasons of self-esteem as long as $\rho > 0$, i.e., when she is risk averse regarding her own beliefs about her type. From now on we will assume $\rho > 0$. Before discussing the role of risk aversion in the next subsection, we now complete the characterization of optimal individual behavior. The previous results imply

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 $x_t \leq \hat{x}_t^*$.*

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 shows 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. Conversely, if the agent reaches period t unaware of her type, it doesn't matter what cutoffs she chose in the past.

Note that we did not simply 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 thus entirely due to their optimizing behavior.

3.3 Characteristics of individual behavior

3.3.1 The role of risk aversion

As shown above, 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.

To see this in the most clear way possible, 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,

$$\mu_0 \left(\frac{\hat{\mu}_1^{1-\rho}}{\hat{\mu}_1} - 1 \right) > \phi (1 - \mu_0) x. \tag{15}$$

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 (15):

$$\hat{x}^* = \frac{\mu_0}{\phi (1 - \mu_0)} \left(\frac{\hat{\mu}_1^{1-\rho}}{\hat{\mu}_1} - 1 \right). \tag{16}$$

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 (16) 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.

Are people really risk averse regarding their beliefs? We believe risk aversion is plausible, and simply point out that risk aversion is needed in this type of setup to obtain resisting behavior. 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.[11] 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 most of them refuse is suggestive of risk aversion over beliefs.

Our model can capture well known patterns of behavior. An important example is the Weberian description of the personal struggle of the Protestant believer. This is a person who, not knowing whether she has been born saved or damned, engages in good acts to maintain or increase her conviction of having been born saved. An immediate problem with the Weberian view is that it is hard to see why one would prefer to maintain any conviction one may have at a cost in terms of consumption. An attractive alternative could be to just find out the truth about one's type and then live accordingly. In this subsection we have clarified that a condition for an individual to want to pass on temptations in order to protect an internal reputation is that she be risk averse over her beliefs. 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. If an individual remembers her motivation to produce just that evidence, she could attribute the good acts to these deliberate attempts, and not to any underlying unknown type. The next subsection discusses the role of imperfect free will in allowing for just that type of learning.

3.3.2 The role of imperfect free will

If intents always turn into actions ($\phi = 1$) then individuals cannot learn about their type when they choose a positive intent. As long as they always choose a positive intent, they remain unaware and have the prior belief μ_0. However, by choosing no intent, they expose

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

themselves to a gamble whereby they find out their type. A high enough temptation can lure them to accept the gamble. The agent now faces an optimal stopping problem in a stationary environment. As there is no growth in self-esteem, $\hat{x}^*$ is the same in every period as long as the individual remains unaware. Therefore it is defined by a stationary version of (14), 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.

$$\hat{x}^* = \frac{g}{1-\mu_0} + \sum_{t=1}^{\infty} \lambda^t \left(F(\hat{x}^*)^{t-1}\right) \left\{ F(\hat{x}^*) \frac{g}{1-\mu_0} - \int_0^{\hat{x}^*} x f(x) dx \right\} \Longleftrightarrow \tag{17}$$

$$\hat{x}^* = \frac{g}{1-\mu_0} + \left(\frac{1}{1-\lambda F(\hat{x}^*)}\right)\left(\frac{g}{1-\mu_0} - E[x|x < \hat{x}^*]\right) \tag{18}$$

This fixed point equation defines the optimal solution. The LHS is a 45-degree line. The RHS begins at a positive value $2g/(1-\mu_0)$ and grows towards

$$\frac{g}{1-\mu_0} + \left(\frac{1}{1-\lambda}\right)\left(\frac{g}{1-\mu_0} - Ex\right) \tag{19}$$

which is finite. Therefore there has to be at least one solution. This indicates that risk aversion is necessary to have individuals pass on temptations, but that imperfect free will is not. Imperfect free will is however necessary for people to learn from past actions of resistance.

3.3.3 The life-cycle of endogenous moral standards, moral capital, and Aristotelian virtue

The individual that behaves as described in Proposition 1 is someone who will attempt to pass on temptations that are low enough. An important question is whether a person who begins by selecting a positive intent has more or less reasons to do that as time goes by and she sees herself resist. In his treatment of moral virtues in Nichomachean Ethics,[12] 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, a person 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. On the other hand, 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 $u(1)$. Another way of

[12] See especially Book II.

seeing the problem is as follows: a risk averse individual will pass on low enough temptations if this buys a reduction in the variance of beliefs. However, those reductions will become very small when the agent becomes close to certain of having a good type. As a result, it is not obvious that individuals who resist temptations make their moral standards, as captured by $\hat{x}_t^*$, more stringent over time. We now state,

Proposition 2 *Individuals who are successful in resisting temptations become more predisposed to resist further temptations.*

Proof. I.e., the sequence $\hat{x}_1^*, \hat{x}_2^*, \ldots$ is increasing. See Appendix. ■

This proposition relies on the fact that individuals who are successful in resisting temptations become more confident about having the good type. This higher confidence, which we call *individual moral capital*, in turn predisposes them to resist even larger temptations, which is expressed in the form of their setting higher cutoffs over time. As we said before, when beliefs get close to certainty the further gains from reducing the variance of beliefs are very small. Why would the agent be interested in setting ever higher cutoffs? The basic reason is that the cost of choosing a higher cutoff decreases faster than the gains from reducing the variance over beliefs. To see that costs must decrease, note that the cost of setting a very high cutoff is that one may miss possibly large temptations. But what is this expected cost? 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. In her mind, 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 the individual's life, depending on her 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 and luck at being in control of their actions. Any of those individuals 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. In the following section we analyze a cohort of individuals and comment on how the distribution of individual beliefs and moral standards evolves over time.

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. First, 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 predispositions 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 it would be confounding by the shortening horizon of the stream of self-esteem returns. 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.

Second, we assumed that good types always behave, while bad types may not. 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 this model. This is due to the fact that the conditions usually invoked in order to characterize policy functions when using dynamic programming are much stronger than necessary and are not met in our model. 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.

Third, we assume 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 enters with symmetric opposite effects depending on intent, and hence may cause an attempt to misbehave to fail. Our results go through in this formulation provided one condition is met, namely 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, the role of confidence in a bright future defined as a lower distributions of temptations, and the role of free will ϕ. Examining changes in beliefs is straightforward: simply consider an alternative initial μ_0 or ϕ. To analyze the role of a brighter future, consider an alternative distribution of temptations G that is first order stochastically dominated by F, i.e., tends to generate lower temptations. Therefore, G may for instance 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.

(b) initial belief μ_0 is higher.

(c) belief about the effectiveness of free will ϕ is higher (shown under exponential distribution of temptations).

Proof. See Appendix. ■

Part (a) tells us that when the individual expects lower temptations in the future she will choose more stringent moral standards today. Thus, a better environment reduces the probability that the individual has done wrong by a given date through two channels: 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. In addition to this, the expectation of a better environment leads the individual to resist even larger shocks, thus magnifying the direct effect. This positive feedback suggests that small differences in the environment could generate relatively large departures in the propensity to do wrong.

Part (b) tells us 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) tells us that when individuals believe that they have more control over their actions they will choose more stringent standards.

4 Moral capital and wrongdoing in a society

In this section we consider a society consisting of individuals who each face 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 the 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 based on 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. After all, only truly bad types can fall for the temptation. We denote the actual share of good types by μ and, for now, assume that ϕ is a correct belief, i.e., the actual probability that free will works.

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 during 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 as aware of being bad are those who did wrong at age t. (This includes individuals who only became aware during age t, i.e., after doing wrong for the first time). 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:

$$\begin{aligned} w_t &= (1-\mu)\left(1 - \Pr\left(\text{unaware}|\hat{x}_1, \ldots, \hat{x}_t, \text{bad}\right)\right) \\ &= (1-\mu)\left(1 - \phi^t H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\right). \end{aligned} \tag{20}$$

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 those who have the bad type eventually become aware of it – either because a very high temptation eventually materializes, or because their free will fails them in some period. Recall that once bad types become aware they will take every temptation that comes along.

The evolution of wrongdoing rates is linked to the evolution of the distribution of beliefs, which we now characterize. 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. All of the unaware people 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

$$\begin{aligned}\bar{\mu}_t &= \left[\mu + (1-\mu)\,\phi^t\right] H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right) \hat{\mu}_t + \mu\left[1 - H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\right] \\ &= \mu + \left(\mu_0 - \mu\right) H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right).. \qquad (21)\end{aligned}$$

Recall that $H_0 = 1$ and $H_t > H_{t+1}$, so clearly $\bar{\mu}_t$ starts from $\bar{\mu}_0 = \mu_0$ and 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 limiting distribution of beliefs is the true distribution of types: A share μ of individuals will have beliefs $\mu_t = 1$, and a share $1 - \mu$ have beliefs $\mu_t = 0$. The variance of beliefs at age t is

$$\begin{aligned}&S_t = \left[\mu + (1-\mu)\,\phi^t\right] H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\left(\hat{\mu}_t - \bar{\mu}\right)^2 + \\ &\mu\left[1 - H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\right](1-\bar{\mu})^2 + (1-\mu)\left[1 - \phi^t H_t\left(\hat{x}_1, \ldots, \hat{x}_t\right)\right]\bar{\mu}^2.\end{aligned}$$

By inspection, the variance of beliefs starts at $S_0 = 0$ and converges to $\mu\,(1-\mu)$ as t goes to infinity. 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 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 more 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) at the same time while the wrongdoing rate increases.

In light of Proposition 3(b), and recalling (21)

Proposition 5 *Inculcating a higher confidence in individuals' own type by inducing a higher initial belief μ_0 leads to lower wrongdoing rates at all ages. This benefit disappears asymptotically as the wrongdoing rate of an infinitely old cohort converges to the share of bad types.*

A successful inculcation requires individuals to not observe the behavior of more than a finite number of other people. If individuals observed a large sample of others' behavior while knowing the structure of the model, they could back out the true share of good types μ and should then use that as the initial prior μ_0. The population could then only be inculcated if they can all be convinced that they have a higher-than-average chance of having the good type.

4.2 Wrongdoing rate of a society in steady state

So far we have showed that, within a single cohort, the wrongdoing rate eventually converges to the share of bad types, regardless of the other model parameters (free will, distribution of temptations, initial beliefs). In this section we show that, in a world where people have finite lifetimes and are replaced by births of new unaware individuals, the wrongdoing rates of two societies with the same share of bad types can have permanently different wrongdoing rates. Thus long run corruption rates across countries do not necessarily and exclusively reflect "deep" moral fundamentals captured by the share of bad types.

Interpret now 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 makes no difference to the wrongdoing rate within a cohort. Suppose also that a new cohort is born in every period, and that the size of newborn cohorts has always been the same. 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 population's share of age$-t$ individuals by z_t. In steady state, entry and exit from each age group must balance out. The steady state relations are

$$z_1 = (1-\lambda)\sum_{t=1}^{\infty} z_t \tag{22}$$

$$z_t = \lambda z_{t-1} \quad \text{for } t = 2, 3, \ldots . \tag{23}$$

The first equation balances out the "currently born" and the "currently dying," while the second equation takes into account that the mass of individuals in all age groups $t \geq 2$ is equal to the mass of survivors from the previous age. Taking into account that $\sum_t z_t = 1$, these steady state relations can be solved for

$$z_t = (1-\lambda)\,\lambda^{t-1} \quad \text{for } t = 1, 2, 3, \ldots. \tag{24}$$

The steady-state rate of wrongdoing in society (call it W) is the weighted average of wrongdoing rates w_t with the weights given by the population shares of the cohorts.

$$W = \sum_{t=1}^{\infty} z_t w_t = (1-\lambda)\sum_{t=1}^{\infty} \lambda^{t-1} w_t, \tag{25}$$

Using the expression for w_t from (20), the steady-state rate of wrongdoing is

$$\begin{aligned} W &= (1-\lambda)\sum_{t=1}^{\infty}\lambda^{t-1}(1-\mu)\left(1-\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\right) \\ &= (1-\mu)\left\{1-(1-\lambda)\sum_{t=1}^{\infty}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\right\} \end{aligned} \tag{26}$$

The proportion of bad types $1-\mu$ 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 6 *The steady state rate of wrongdoing in society W is lower when*

(i) the initial beliefs μ_0 of the newly born are higher,

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

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

Proof. Part (i) follows from Proposition 3(a); part (ii) follows from Proposition 3(b); and part (iii) follows from

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

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

It is obvious that a higher true share of good types results in a lower wrongdoing rate. However, regardless of the true population share of good types, a higher initial belief μ_0 makes individuals more resistant to temptations and thus lowers the wrongdoing rate. This is because, with higher μ_0 the resistance cutoffs are higher so the bad types, on average, go on longer before facing an irresistible temptation. This suggests a useful role for indoctrination.

A society that displays higher corruption may just have had worse luck in previous history, while the deep "moral" make up in terms of types is the same, and even the average belief is the same.

4.3 Response to shocks: wrongdoing across societies

Now let's 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 in utility terms. Two otherwise similar societies who face different macroeconomic shocks may end up with different wrongdoing rates.

The case of a cohort Consider first two initially 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 2 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 7 *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 (20), 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

$$w_{t,s} = (1-\mu)\left\{1 - \phi^t H_t(\hat{x}_1, \ldots, \hat{x}_t)\frac{G(\hat{x}_s)}{F(\hat{x}_s)}\right\}. \tag{27}$$

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 the society is not about the average belief of individuals. Instead, it depends 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 the cohorts that were born less than s periods ago are not affected by the shock so their wrongdoing rate is given by (20), while those that were born during or after the shock have the wrongdoing rate given by (27). Combining the cohort wrongdoing rates with the population shares (24), the aggregate rate of wrongdoing s periods after the shock is

$$\begin{aligned} W_s &= \sum_{t=1}^{s} z_t w_t + \sum_{t=s+1}^{\infty} z_t w_{t,s} \qquad (28) \\ &= (1-\mu)\left\{1-(1-\lambda)\left(\sum_{t=1}^{s}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)+\sum_{t=s+1}^{\infty}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\frac{G(\hat{x}_{t-s})}{F(\hat{x}_{t-s})}\right)\right\} \end{aligned}$$

where we define $\sum_{t=1}^{0}\text{term}_t \equiv 0$ for convenience to cover also 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 shock period.

Proposition 8 *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 (28) and (26) is the deviation of society's wrongdoing rate from steady state s periods after the shock:

$$\Delta_s = W_s - W = (1-\mu)(1-\lambda)\sum_{t=s+1}^{\infty}\lambda^{t-1}\phi^t H_t(\hat{x}_1,\ldots,\hat{x}_t)\left(1-\frac{G(\hat{x}_{t-s})}{F(\hat{x}_{t-s})}\right). \qquad (29)$$

This is positive if the shock is bad (i.e., if $G(x) < F(x)$), and negative if the shock is good. As the shock becomes more distant past, $s \to \infty$, even the smallest of the factors $\lambda^s\phi^{s-1}H_s(\hat{x}_1,\ldots,\hat{x}_s)$ converges to zero. ■

If the shock is "bad" then the deviation from steady state $W_s - W$ is positive. The effects of any shock will die out asymptotically for two reasons: First, and more obviously, because the dying are replaced by new cohorts who did not experience the shock, and second, because among the survivors the remaining unaware types eventually find out the truth which some of their unlucky peers found out prematurely due to the bad shock.

History matters through the stock of unaware bad types: Their ranks are diminished more than usual in a period when shocks are high worse than usual. And those who find out they have the bad type are locked-in in wrongdoing until they die. But this irreversibility of behavior at the individual level is progressively washed away by the entry of new generations. If the survival rate λ is high then the appearance of new cohorts is very slow and the shock dies out slowly. Similarly, by inspection of (29), a high free will parameter ϕ slows the erosion of the stock of unaware bad types within any cohort, thus leading to a longer-lasting effect from the shock because it wiped out more resisting behavior.

5 Applications

5.1 Two policy instruments: taboo and punishment

5.1.1 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. These incentives could adopt the form of payments contingent on good behavior or punishments contingent on bad behavior. For concreteness, we will focus on the latter case assuming that the planner can detect bad behavior with some exogenous probability. We do not try to characterize optimal incentive schemes in all generality, but simply to show that optimal incentives are shaped by the fact that wrongdoing has an impact on moral capital and predispositions toward future wrongdoing.

An important margin that we investigate here relates to whether punishment should differ between those who do wrong for the first time and those who are repeat offenders. In order to isolate the effect of interest we will impose the following simplifications. We assume that the planner faces a population of mass 2 constituted by aware and unaware individuals in equal proportions. Without loss of generality we make the age of the population equal to 1 and assume that in each period they receive temptations which are independently drawn from the distribution $F(x)$. The planner knows past behavior by all agents. Further, the planner 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 punishment to be imposed respectively on the aware and the unaware that are caught doing wrong. These punishments should be interpreted as expected punishments - 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 that wants to minimize wrongdoing would certainly have an easy task if punishment were costless. So we assume that increasing expected punishment is costly to the planner as captured by the increasing and convex cost function $c(N_a + N_u)$. Our cost formulation captures a world where threatening with more likely and intense punishment is costly because it requires stronger detection and punishment capabilities that must be deployed ex ante to make such threats credible.[13] Lastly, we assume that the planner discounts the future according to the factor δ.

[13] Costs may also increase with the number of people who do wrong and who must eventually be punished. We abstract from this possibility which essentially introduces a form of increasing returns to punishment, because larger punishments pay for themselves through a lower number of offenders who must be punished. 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.

To construct the objective of the planner, we first characterize the impact of punishment on wrongdoing. Because those who are good never do wrong, the wrongdoing rates that matter to the planner are those by the bad types. Therefore in what follows we focus only on bad types and to simplify notation we normalize their measure 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 prefer to attempt to resist temptation 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. Because 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 4 *A one time punishment N_u attains a reduction in the expected wrongdoing rate 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 of this lemma explains that under punishment N_u the cutoff of the current period satisfies $\hat{x}_1^p = \hat{x}_1 + N_u$, so current punishment raises the optimal cutoff of the unaware in the current period 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 successfully, leading to further reductions in wrongdoing in future periods. Specifically, of those who are saved from temptation in the current period, $\phi F\left(\hat{x}_2\right)$ are saved again in period 2, and $\phi^2 F\left(\hat{x}_2\right) F\left(\hat{x}_3\right)$ are saved in period three, and so on, explaining the expression in the last lemma, where $\sum_{s=1}^{\infty}\left(\delta\lambda\right)^{s-1}\phi^{s-1}\frac{H_s}{H_1} = 1 + \delta\lambda\phi F\left(\hat{x}_2\right) + \left(\delta\lambda\right)^2\phi^2 F\left(\hat{x}_2\right)F\left(\hat{x}_3\right) + ...$ captures the present and future (discounted) reductions in wrongdoing. All future cutoffs are unchanged.

The social planner's problem

Using lemma (4), the planner's objective is to maximize,

$$\phi F\left(N_a\right) + \phi\left[F\left(\hat{x}_t + N_u\right) - F\left(\hat{x}_t\right)\right]Z_t - c\left(N_a + N_u\right) \tag{30}$$

with respect to N_a and N_u, where $Z_t = \sum_{s=1}^{\infty}\left(\delta\lambda\right)^{s-1}\phi^{s-1}\frac{H_s}{H_1}$, which only contains future cutoffs and does not involve $\hat{x}_1^p$. Given this program, we can now state,

Proposition 9 *If the planner's patience is low enough and larger temptations are less frequent than smaller ones, then the planner imposes harsher punishment on repeat offenders* *relative to first-time wrong-doers. Formally, if δ is low enough and $f(x)$ is decreasing, then $N_a > N_u$.*

Proof. The first-order conditions for N_a and N_u are,

$$\begin{aligned} \phi f(N_a) - c'(N_a + N_u) &= 0, \qquad (31) \\ \phi f(\hat{x}_t + N_u) Z_t - c'(N_a + N_u) &= 0. \qquad (32) \end{aligned}$$

Solving for $c'(N)$ and combining yields

$$f(N_a) = f(\hat{x}_t + N_u) Z_t. \qquad (33)$$

Note that Z_t approaches 1 as λ approaches zero. Recall that $\hat{x}_1 > 0$. Therefore, in the neighborhood of $Z_t = 1$, $f(x)$ being decreasing yields the result. ■

An intrinsic disposition to resist temptations allows individuals to behave honestly even when there are no extrinsic incentives in place. And as we know from the literature on crime, extrinsic incentives can work to keep individuals behaving honestly. Proposition 9 tells us that the design of extrinsic incentives should reflect the presence 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. A first condition is that larger temptations should be relatively more rare than small temptations. A second condition is that the planner should be sufficiently impatient 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 should carry 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). 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.

5.1.2 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 wrongdoing is socially harmful and that there is a social planner who would like to reduce wrongdoing rates. Assume that the planner can indoctrinate individuals by instilling in them a taboo against some inherently harmless behavior. Would she find it worthwhile to do so?

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 (beer, say) that yields positive utility. Suppose that the planner can convince individuals, before they make this choice, that consuming beer amounts to falling for a temptation, and that doing so reveals that they have the bad type. Given the structure of the model, this is a lie, but we assume that the planner has the ability to make this lie about the structure of the model and be believed. Assume that as in our model, individuals care about the self-esteem associated to the probability of being a good type. Lastly, assume that the planner cares about minimizing the steady state wrongdoing rate.

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 measure 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. 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

[14] Benabou and Tirole (2007) also study taboos. In their setup the agent herself may decide to avoid information about the price of a "taboo" transaction (e.g., for selling an organ or a sexual service), as part of a self-control strategy. In our setup, agents cannot avoid knowing the size of the current temptation, and the classification of what counts as a taboo is beyond their control.

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 is

$$w_t^{'} = (1-\mu)\left(1-\phi^{t+1}H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right)\frac{F\left(\hat{x}_{t+1}\right)}{F\left(\hat{x}_1\right)}\right) \tag{34}$$

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

$$W^{'}-W = -\left(1-\mu\right)\left(1-\lambda\right)\phi\sum_{t=1}^{\infty}\left(\phi\lambda\right)^{t-1}\left(\phi\frac{F\left(\hat{x}_{t+1}\right)}{F\left(\hat{x}_1\right)}-1\right)H_t\left(\hat{x}_1,\ldots,\hat{x}_t\right). \tag{35}$$

The taboo will lower the steady-state rate of wrongdoing in society when (35) 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]

5.2 Moral capital and career choice

In an economy where individual beliefs vary and different careers offer different distributions of temptations, how do individuals select into careers? 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 holding different initial beliefs μ that they have the good type. We want to know how individuals will self-select into different occupations depending on their μ. When individuals can choose between careers with different mean temptations, they will require compensation to enter a career that would otherwise promise them a lower expected utility. We assume that the economy has a need for workers in both careers, so compensation has 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 will live for only one period and must make their occupational choice as soon as they are born. We next show that types with high beliefs about their own type select into careers with lower temptations.

[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.

Proposition 10 *In an economy where two occupations offer different distributions of temptations, and one first-order stochastically dominates the other, individuals are divided into two segments such that those with a higher belief about being a moral type will select into the occupation with lower temptations.*

Proof. See appendix. ■

For aware types the selection is obvious: An individual with $\mu = 1$ will be indifferent between the two careers, and will prefer the low-temptation career under any positive compensating differential. An individual with $\mu = 0$ only cares about temptations and will choose the high-temptation activity even in the absence of compensating differential. In between, the result is not obvious, because the unaware types also have an incentive to protect their self-image by choosing a low-temptation activity. In fact, 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 types 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. The higher rate of wrongdoing in the high-temptation sector is further reinforced by the selection of types.

6 Conclusion

We propose a model where an individual faces a sequence of temptations which, if taken, would yield positive payoffs. In a standard model, a rational individual would always seize those temptations. However, in our model the individual also obtains self-esteem from her self-image, modeled as a flow utility from her beliefs about her type. When this type is associated to the performance of certain actions (like resisting temptations), it is possible that the individual may perform those actions even if they yield strictly lower extrinsic payoffs because they maintain the individual's introspective reputation, or self-image. Not every individual will want to resist temptations: a necessary and sufficient condition is that the individual be risk-averse regarding her self-image.

The model can be cast to deal with the issue of intrinsic motivation in general, but our main application is to the development of moral dispositions and the propensity to do wrong. We do not explain the content of morality, but a mechanism by which individuals formulate their moral standards. When individuals are risk averse regarding their self-image, they will want to resist some temptations that are enjoyable if the content of morality condemns those temptations. When lacking perfect free will, 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. Criminologists understand self-control broadly, encompassing both the ability to control impulses and the ability to take into account the future (see, inter alia, Gottfredson and Hirschi 1990 and Nagin and Paternoster 1993). The model predicts that both traits of self-control, namely a higher ability to transform intentions into actions and a lower discount rate will increase individuals' endogenous moral dispositions.

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-image. This polarization will cause more wrongdoing even if the average beliefs are the same. 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. We explain the determination of those costs in connection to the value the person places in sharing such identity, and to the evolution of the person's beliefs that the identity is truly hers. The model can also rationalize taboos and why societies punish repeat-offenders more harshly. Lastly, we consider the problem of who will be attracted to high temptation activities, of which politics may be a good example. We find that individuals with low moral capital have a comparative advantage at high temptation activities and will tend to self-select into them. This implies that high temptation activities will generate high wrongdoing for two reasons that compound each other: they generate higher temptations on average, and they attract the people least interested in resisting them.

Appendix

Proof of Lemma 1: Inspection of the equation (13) reveals that each cutoff is uniquely determined as a sum of two terms: the first one captures the trade-off facing an individual 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 $\sum_{t=1}^{\infty} \lambda^{t-s+1} \frac{H_{t+s-1}}{H_s} \left\{ \frac{F(\hat{x}_{t+s})g_{t+s}}{(1-\mu_0)\phi^s} - \phi^t \int_0^{\hat{x}_{t+s}} xf(x)dx \right\}$ equals V_s minus a constant). Then the uniqueness of an optimal sequence characterized by the FOCs in (12) follows. To see this, suppose not. Then starting in some period $n \geq 1$ there is a number of periods in which there is more than one cutoff forming part of a sequence satisfying the FOCs. Take any period s where there is more than one cutoff. If there are future periods with more than one cutoff, 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 as said earlier the FOC at s determines $\hat{x}_s$ uniquely as a function of the continuation value at $s+1$ and the term $\frac{g_s}{(1-\mu_0)\phi^s}$.■

Proof of Lemma 2: 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 $\hat{x}_1$,

$$\frac{\partial V_0}{\partial \hat{x}_1} = f(\hat{x}_1) \left\{ \begin{array}{c} \{u_1 [\mu_0 + (1-\mu_0)\phi] - \mu_0 u(1) - (1-\mu_0)\phi \hat{x}_1\} + \\ \frac{1}{F(\hat{x}_1)} \left(\sum_{t=1}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F_{t+1} u_{t+1} \left[\mu_0 + (1-\mu_0)\phi^{t+1}\right] \\ -\mu_0 F_{t+1} u(1) - (1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} xf(x)dx \end{array} \right\} \right) \end{array} \right\} = 0. \tag{36}$$

Inspection reveals that $V_0(\hat{x}_1^*, \hat{x}_2^*, ...)$ is concave in $\hat{x}_1$: first, because the density is positive everywhere in the support of x we have that $f(\hat{x}_1) > 0$. Second, the large product involving $\frac{1}{F(\hat{x}_1)}$ can in fact be shown to be independent of $\hat{x}_1$ by canceling $\frac{1}{F(\hat{x}_1)}$ out with the factor $F(\hat{x}_1)$ inside H_t. Therefore, 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 rendered negative by an increase of $\hat{x}_1$ above $\hat{x}_1^*$.

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 by increasing the threshold without bound? Suppose she can. That would mean that either when a threshold $\hat{x}_s$ is getting close to zero or when getting arbitrarily large the objective function would be increasing. Consider the first case when 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 because it 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.■

Proof of Lemma 3: 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 (14), 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,

$$u(\mu_1)[\mu_0 + (1-\mu_0)\phi] - \mu_0 u(1) > 0, \tag{37}$$

or, in other words, that

$$\left(\frac{\mu_0}{\mu_0 + (1-\mu_0)\phi}\right)^{1-\rho}[\mu_0 + (1-\mu_0)\phi] - \mu_0 > 0, \tag{38}$$

or,

$$\mu_0\left(\frac{\mu_1^{1-\rho}}{\mu_1} - 1\right) > 0,$$

which is clearly met if and only if $\rho > 0$. This does not show necessity, however, 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 $g_1 > 0$.

Now we show $\{\hat{x}_i^*\}_{i=1}^{\infty}$ converges to a positive limit whenever $\rho > 0$. We need to show two things. First, that if $\{\hat{x}_i^*\}_{i=1}^{\infty}$ converges it does it to a unique limit that exists. We then show it converges. To see the first point, note that as μ converges to unity the problem becomes stationary, so $\hat{x}^*$ should also be stationary and equal in all future periods. The limiting value of $\hat{x}^*$ must satisfy the following fixed point equation:

$$\hat{x}^* = G_1 + \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F(\hat{x}^*)\right)\left\{F(\hat{x}^*)G_{t+1} - \phi^t \int_0^{\hat{x}^*} xf(x)dx\right\} \tag{39}$$

$$\hat{x}^* = G_1 + \sum_{t=1}^{\infty} \lambda^t F(\hat{x}^*)^t \left\{G_{t+1} - \phi^t E[x|x \leq \hat{x}^*]\right\} \tag{40}$$

where $E[x|x \leq \hat{x}^*] = \frac{1}{F(\hat{x}^*)} \int_0^{\hat{x}^*} xf(x)dx$ was used and

$$G_t = \frac{u\left(\frac{\mu_0}{\mu_0+(1-\mu_0)\phi^t}\right)\left[\mu_0 + (1-\mu_0)\phi^t\right] - \mu_0 u(1)}{(1-\mu_0)\phi}. \tag{41}$$

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 (40), the limiting value as the solution of

$$\hat{x}^* = \rho + \sum_{t=1}^{\infty} \lambda^t F(\hat{x}^*)^t \left(\rho\phi^t - \phi^t E[x|x \leq \hat{x}^*]\right) \tag{42}$$

$$\hat{x}^* = \rho + \sum_{t=1}^{\infty} \lambda^t F(\hat{x}^*)^t \phi^t \mathrm{E}[\rho - x|x \leq \hat{x}^*]. \tag{43}$$

Alternatively this can be written as,

$$\hat{x}^* - \rho = \lambda\phi F(\hat{x}^*) \mathrm{E}[\hat{x}^* - x|x \leq \hat{x}^*]. \tag{44}$$

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 $\lambda\phi F(\hat{x}^*) < 1$. This establishes that the limit value for $\{\hat{x}_i^*\}_{i=1}^{\infty}$ exists and is unique and therefore that if the sequence converges it does it to a unique limit.

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.■

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. In other words, the problem of finding the optimal $\hat{x}_1$ is analogous to the problem of finding the optimal $\hat{x}_s$ for any $s > 1$ up to the change in beliefs. Therefore, if we can show that $\hat{x}_1^*$ is increasing in the initial beliefs μ_0, then we will know that the sequence $\{\hat{x}_i^*\}_{i=1}^{\infty}$ is increasing over time.

As said earlier, $\partial^2 V_0/\partial\hat{x}_1^*\partial\hat{x}_t^* = 0$, the indirect effect of μ_0 on $\hat{x}_1$ through changes in future thresholds $\hat{x}_s$ is zero. This means that we are 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. Now recall that $\hat{x}_1$ can be written as,

$$\begin{aligned}\hat{x}_1^* &= \frac{g_1(\mu_0)}{(1-\mu_0)\phi} + \\ &+ \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F_s \right) \left\{ F_{t+1} \frac{g_{t+1}(\mu_0)}{(1-\mu_0)\phi} - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \right\}. \end{aligned} \tag{45}$$

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

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{\left\{ \frac{du}{d\mu_t}\frac{d\mu_t}{d\mu_0}\left[\mu_0 + (1-\mu_0)\phi^t\right] + u_t\left(1-\phi^t\right) - u(1)\right\}}{(1-\mu_0)\phi} + \frac{g_t(\mu_0)}{(1-\mu_0)^2\phi}. \tag{46}$$

The first term can be shown to equal,

$$\frac{u_t\left[\mu_0 + (1-\mu_0)\phi^t - \rho\phi^t\right] - u(1)}{\mu_0(1-\mu_0)\phi}, \tag{47}$$

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,

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{\left[\mu_0 + (1-\mu_0)\phi^t - \rho\phi^t\right]u_t - \mu_0 u(1)}{\mu_0(1-\mu_0)\phi} + \frac{\left[\mu_0 + (1-\mu_0)\phi^t\right]u_t - \mu_0 u(1)}{(1-\mu_0)^2\phi}, \tag{48}$$

and rearranging,

$$\frac{d\left(\frac{g_t(\mu_0)}{(1-\mu_0)\phi}\right)}{d\mu_0} = \frac{u_t\left\{\left[\mu_0 + (1-\mu_0)\phi^t\right] - (1-\mu_0)\rho\phi^t\right\} - \mu_0 u(1)}{\mu_0(1-\mu_0)^2\phi}. \tag{49}$$

Therefore, we need to show

$$\left(\frac{\mu_0}{\mu_0 + (1-\mu_0)\phi^t}\right)^{1-\rho} > \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)\phi^t}. \tag{50}$$

Tedious algebra shows us that,

$$\left(\frac{\mu_0}{\mu_0 + (1-\mu_0)\phi}\right)^{1-\rho} > \frac{\mu_0}{\mu_0 + (1-\mu_0)(1-\rho)\phi},\ \phi \in (0,1), \rho \in (0,1), \mu_0 \in (0,1), \tag{51}$$

which is identical to the expression we need to prove, except for the fact that the latter expression contains ϕ where we should have ϕ^t. Because the latter expression is true for any value of ϕ in $(0,1)$, it must also be true for ϕ^t.■

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

(b) Again we ignore indirect effects and compute only the partial derivative due to $\partial^2 V_0 / \partial \hat{x}_1^* \partial \hat{x}_t^* = 0$. Without loss of generality 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.

$$\begin{aligned}
&\hat{x}_1^*(G) = u_1 \left[\frac{\mu_0 + (1-\mu_0)\,\phi}{(1-\mu_0)\,\phi} \right] - \frac{\mu_0}{(1-\mu_0)\,\phi} u\,(1) + \\
&\sum_{t=1}^{k-2} \lambda^t \left(\prod_{s=2}^{t} F_s \right) \left\{ \begin{array}{c} F_{t+1} u_{t+1} \frac{\left[\mu_0 + (1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} F_{t+1} u\,(1) - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} + \\
&+ \lambda^{k-1} \left(\prod_{s=2}^{k-1} F_s \right) \left\{ \begin{array}{c} G_k u_k \frac{\left[\mu_0 + (1-\mu_0)\phi^{k}\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} G_k u\,(1) - \phi^{k-1} \int_0^{\hat{x}_{k}} x f(x) dx \end{array} \right\} \\
&+ \sum_{t=k}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F_s \frac{G_k}{F_k} \right) \left\{ \begin{array}{c} F_{t+1} u_{t+1} \frac{\left[\mu_0 + (1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} \\ -\frac{\mu_0}{(1-\mu_0)\phi} F_{t+1} u\,(1) - \phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\}. \qquad (52)
\end{aligned}$$

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,

$$\begin{aligned}
\hat{x}_1^*(G) - \hat{x}_1^*(F) = \lambda^{k-1} \left(\prod_{s=2}^{k-1} F_s \right) \left\{ (G_k - F_k) \left[u_k \frac{\left[\mu_0 + (1-\mu_0)\,\phi^k\right]}{(1-\mu_0)\,\phi} - \frac{\mu_0}{(1-\mu_0)\,\phi} u\,(1) \right] \right\} + \\
\sum_{t=k}^{\infty} \lambda^t \left[\left(\prod_{s=2}^{t} F_s \frac{G_k}{F_k} \right) - \left(\prod_{s=2}^{t} F_s \right) \right] \left\{ \begin{array}{c} F_{t+1} \left[u_{t+1} \frac{\left[\mu_0 + (1-\mu_0)\phi^{t+1}\right]}{(1-\mu_0)\phi} - \frac{\mu_0}{(1-\mu_0)\phi} u\,(1) \right] - \\ -\phi^t \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\}. \quad (53)
\end{aligned}$$

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 extremely 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.

Proof of Lemma 4: The optimization problem for an 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 maximize, The unaware person wants to choose a sequence of cutoffs $\{\hat{x}_1^p, \hat{x}_2^p, ...\}$ to maximize,

$$
\begin{aligned}
V &= \frac{\mu_0 u(1) + (1-\mu_0)(Ex)}{1-\lambda} + F_1 \{u_1 [\mu_0 + (1-\mu_0)\phi] - \mu_0\} + \\
&\quad + (1-\mu_0)\left[\phi F_1 N_u - \phi \int_0^{\hat{x}_1^p} x f(x) dx - N_u\right] + \qquad (54) \\
&\quad + \sum_{t=1}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F_{t+1} u_{t+1} \{[\mu_0 + (1-\mu_0)\phi^{t+1}] - \mu_0\} + \\ -(1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\}, \qquad (55)
\end{aligned}
$$

where the functions 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,

$$
\begin{aligned}
&\frac{\partial V}{\partial \hat{x}_1^p} = f(\hat{x}_1) \{u_1 [\mu_0 + (1-\mu_0)\phi] - \mu_0\} - \qquad (56) \\
&- (1-\mu_0)\phi \hat{x}_1 f(\hat{x}_1) + (1-\mu_0)\phi f(\hat{x}_1) p N_u + \qquad (57) \\
+ \quad &\frac{\partial}{\partial \hat{x}_1} \left(\sum_{t=1}^{\infty} \lambda^t H_t \left\{ \begin{array}{c} F_{t+1} \{u_{t+1} [\mu_0 + (1-\mu_0)\phi^{t+1}] - \mu_0\} - \\ -(1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\} \right) = 0,
\end{aligned}
$$

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

$$
\begin{aligned}
&\hat{x}_1^{*p} = u_1 \left[\frac{\mu_0 + (1-\mu_0)\phi}{(1-\mu_0)\phi}\right] - \frac{\mu_0}{(1-\mu_0)\phi} + p N_u \qquad (58) \\
&\frac{1}{(1-\mu_0)\phi} \sum_{t=1}^{\infty} \lambda^t \left(\prod_{s=2}^{t} F_s\right) \left\{ \begin{array}{c} F_{t+1} \{u_{t+1} [\mu_0 + (1-\mu_0)\phi^{t+1}] - \mu_0\} - \\ -(1-\mu_0)\phi^{t+1} \int_0^{\hat{x}_{t+1}} x f(x) dx \end{array} \right\}. \qquad (59)
\end{aligned}
$$

Comparing this expression with the FOC for $\hat{x}_1$ in (14) tells us that $\hat{x}_1^{*p} = \hat{x}_1 + N_u$, implying that punishment N_u achieves a reduction in wrongdoing equal to $\phi(F(\hat{x}_1 + N_u) - F(\hat{x}_1))$ because current punishment raises the optimal cutoff of the unaware one for one in period 1. The first order conditions in (13) 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 increasing the share of unaware individuals who resist successfully and enter the future in a continuing state of unawareness. Specifically, of those who are saved from temptation in the current period, $\phi F(\hat{x}_2)$ are saved again in period 2, so $\phi F(\hat{x}_2)$ is the reduction of wrongdoing in period 2 as a result of punishment N_u having been present in period 1. Next, $\phi^2 F(\hat{x}_2) F(\hat{x}_3)$ are saved in period three, and so on. As a result, the one time punishment N_u leads to an expected number of wrongdoing reduction equal to

$\phi\left(F\left(\hat{x}_1+N_u\right)-F\left(\hat{x}_1\right)\right)\left[1+\phi F\left(\hat{x}_2\right)+\phi^2 F\left(\hat{x}_2\right)F\left(\hat{x}_3\right)+...\right]$. If, moreover, the planner discounts more heavily reductions in crime that take place farther into the future at according to a factor δ, we obtain the expression in the lemma.■

Proof of Proposition 10: Recall the optimal policy in the one-period setup (16). 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

$$\begin{aligned}
V\left(\mu,\theta\right) &= F\left(\hat{x}|\theta\right)\left(\left[\mu+\left(1-\mu\right)\phi\right]u\left(\hat{\mu}\right)+(1-\mu)(1-\phi)E[x|x<\hat{x},\theta]\right) \\
&\quad +\left(1-F\left(\hat{x}|\theta\right)\right)\left(\mu u\left(1\right)+\left(1-\mu\right)E[x|x\geq\hat{x},\theta]\right) \\
&= F\left(\hat{x}|\theta\right)\left(\left[\mu+\left(1-\mu\right)\phi\right]u\left(\hat{\mu}\right)-\mu\right)+\mu \\
&\quad +\left(1-\mu\right)\left(\theta-\phi\int_0^{\hat{x}}xf\left(x|\theta\right)dx\right) \\
&= F\left(\hat{x}|\theta\right)\mu\left(\hat{\mu}^{-\rho}-1\right)+\mu+\left(1-\mu\right)\left(\theta-\phi\int_0^{\hat{x}}xf\left(x|\theta\right)dx\right). \qquad (60)
\end{aligned}$$

The distribution with higher temptations is defined in terms of first order stochastic dominance, so $F_\theta\left(x|\theta\right)<0$. 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

$$\begin{aligned}
V\left(\mu,\theta_H\right)-V\left(\mu,\theta_L\right) &= \left(F\left(\hat{x}|\theta_H\right)-F\left(\hat{x}|\theta_L\right)\right)\mu\left(\hat{\mu}^{-\rho}-1\right)+\left(1-\mu\right)\left(\theta_H-\theta_L\right) \quad (61)\\
&\quad -\left(1-\mu\right)\phi\int_0^{\hat{x}}x\left[f\left(x|\theta_H\right)-f\left(x|\theta_L\right)\right]\mathrm{d}x.
\end{aligned}$$

Now hold any $\theta_L>0$ as fixed and consider the difference $V\left(\mu,\theta_H\right)-V\left(\mu,\theta_L\right)$. 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 $H\left(\mu\right)\equiv\mu\left(\hat{\mu}^{-\rho}-1\right)$. Noting that the envelope theorem helps us eliminate all terms involving $\hat{x}'\left(\mu\right)$, the differentiation of (61) with respect to μ yields

$$\begin{aligned}
&V_\mu\left(\mu,\theta_H\right)-V_\mu\left(\mu,\theta_L\right)= \qquad (62)\\
&\left(F\left(\hat{x}|\theta_H\right)-F\left(\hat{x}|\theta_L\right)\right)H'\left(\mu\right)-\left(\theta_H-\theta_L\right)+\phi\int_0^{\hat{x}}x\left[f\left(x|\theta_H\right)-f\left(x|\theta_L\right)\right]\mathrm{d}x.
\end{aligned}$$

Using integration by parts to transform $\int_0^{\hat{x}}xf\left(x|\theta\right)\mathrm{d}x=\hat{x}F\left(\hat{x}|\theta\right)-\int_0^{\hat{x}}F\left(x|\theta\right)\mathrm{d}x$ then (62) becomes

$$\begin{aligned}
& \left(F\left(\hat{x}|\theta_H\right)-F\left(\hat{x}|\theta_L\right)\right)H'\left(\mu\right)-\left(\theta_H-\theta_L\right) && (63)\\
& \quad +\phi\hat{x}\left(F\left(\hat{x}|\theta_H\right)-F\left(\hat{x}|\theta_L\right)\right)-\phi\int_0^{\hat{x}}\left[F\left(x|\theta_H\right)-F\left(x|\theta_L\right)\right]\mathrm{d}x && (64)\\
= & \left(F\left(\hat{x}|\theta_H\right)-F\left(\hat{x}|\theta_L\right)\right)\left(H'\left(\mu\right)+\phi\hat{x}\right)-\left(\theta_H-\theta_L\right)-\phi\int_0^{\hat{x}}\left[F\left(x|\theta_H\right)-F\left(x|\theta_L\right)\right]\mathrm{d}x && (65)
\end{aligned}$$

The first term of (65) is negative if $H'\left(\mu\right)+\phi\hat{x}$ is positive. And since $\partial\hat{\mu}/\partial\mu=\phi\left(\hat{\mu}/\mu\right)^2$ we can write

$$\begin{aligned}
H'\left(\mu\right) &= \frac{\partial}{\partial\mu}\left[\mu\left(\hat{\mu}^{-\rho}-1\right)\right]=\hat{\mu}^{-\rho}-1-\rho\mu\hat{\mu}^{-\rho-1}\frac{\partial\hat{\mu}}{\partial\mu} && (66)\\
&= \hat{\mu}^{-\rho}-1-\rho\mu\hat{\mu}^{-\rho-1}\phi\left(\frac{\hat{\mu}}{\mu}\right)^2=\hat{\mu}^{-\rho}\left(1-\rho\phi\frac{\hat{\mu}}{\mu}\right)-1. && (67)
\end{aligned}$$

Thus

$$\begin{aligned}
H'\left(\mu\right)+\phi\hat{x} &= \left[\hat{\mu}^{-\rho}\left(1-\rho\phi\frac{\hat{\mu}}{\mu}\right)-1\right]+\phi\left[\frac{\mu}{\left(1-\mu\right)\phi}\left(\hat{\mu}^{-\rho}-1\right)\right] && (68)\\
&= \left(\frac{1}{1-\mu}\right)\left[\hat{\mu}^{-\rho}\left(\frac{\mu+\left(1-\rho\right)\left(1-\mu\right)\phi}{\mu+\left(1-\mu\right)\phi}\right)-1\right]. && (69)
\end{aligned}$$

This is always positive if

$$\frac{\mu+\left(1-\rho\right)\left(1-\mu\right)\phi}{\mu+\left(1-\mu\right)\phi}>\left(\frac{\mu}{\mu+\left(1-\mu\right)\phi}\right)^{\rho}, \qquad (70)$$

which is implied by (51).■

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 (forthcoming), The Brain as a Hierarchical Organization, *American Economic Review.*

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.

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.

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.

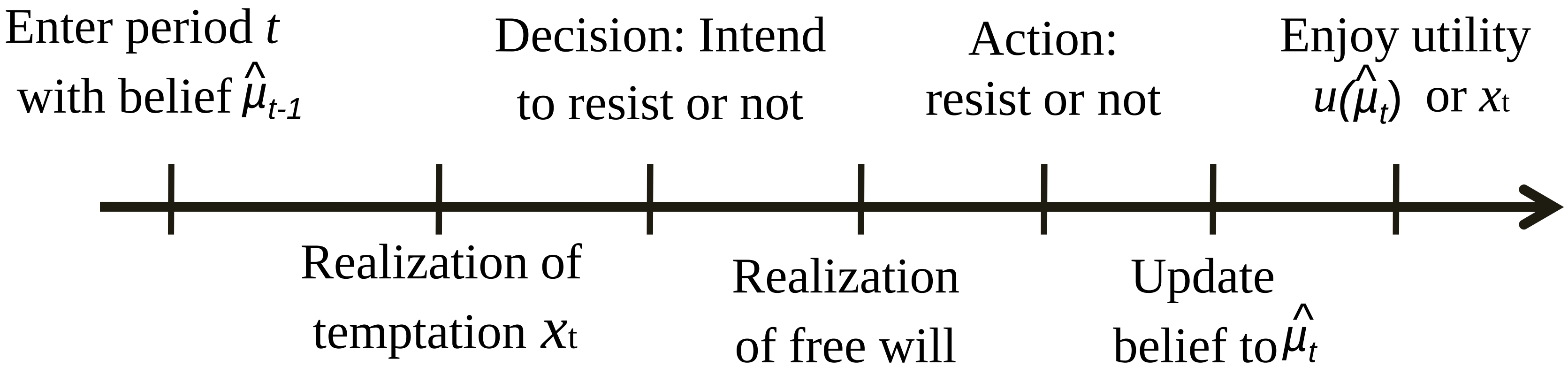
Rabin, M. (1995), Moral Preferences, Moral Constraints, and Self-Serving Biases, mimeo UC Berkeley.

Rubinstein, W.D. (1999), The Weber Thesis and the Jews, in Brezis, E. and P. Temin (eds.), Elites, Minorities, and Economic Growth. North-Holland.

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

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

Figure 1: Timeline for period *t*


Enter period t with belief $\hat{\mu}_{t-1}$
Realization of temptation x_t
Decision: Intend to resist or not
Realization of free will
Action: resist or not
Update belief to $\hat{\mu}_t$
Enjoy utility $u(\hat{\mu}_t)$ or x_t