This is the last working paper version of the article that appeared in the Journal of the European Economic Association 2013, 11(3), p. 599-663.

Self-Esteem, Moral Capital, and Wrongdoing*

Ernesto Dal Bó
UC Berkeley and NBER

Marko Terviö
Aalto University and HECER

June 25, 2012

Abstract

We present an infinite-horizon planner-doer model of moral standards, where individuals receive random temptations (such as bribe offers) and must decide which to resist. Individual actions depend both on conscious deliberation and on a type reflecting unconscious drives. Temptations yield consumption value, but confidence in being the type of person who resists temptations yields self-esteem. We identify conditions for individuals to build an introspective reputation for goodness ("moral capital") and for good actions to lead to a stronger disposition to do good. Bad actions destroy moral capital and lock-in further wrongdoing. Economic shocks that result in higher temptations have long-lasting effects on wrongdoing. We show how optimal deterrence can change under endogenous moral costs and how wrongdoing can be compounded as high temptation activities attract individuals with low moral capital.

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

JEL codes: D83, K4, Z13.

---

*Ernesto Dal Bó: Haas School of Business, UC Berkeley and NBER (dalbo@haas.berkeley.edu). Marko Terviö: Aalto University (marko.tervio@aalto.fi). We thank Roland Bénabou, Jeremy Bulow, Pedro Dal Bó, Erik Eyster, Mitri Kitti, Botond Kőszegi, Keith Krehbiel, Don Moore, John Morgan, Santiago Oliveros, Demian Pouzo, Matt Rabin, Tim Williamson, and seminar participants at Arizona State, Berkeley, Birmingham, Essex, HECER, Princeton, Stanford, UCSD, and Universidad de San Andrés for useful conversations and comments. Juan Escobar provided excellent research assistance.

1 Introduction

Norms defining socially acceptable behavior play a large role in socioeconomic outcomes by discouraging opportunistic behavior. But what determines adherence to received social norms such as moral rules? If norms are inculcated as a stable part of tastes, Beckerian models of crime will predict an individual's adherence to norms to depend steadily on variations in the extrinsic payoffs to opportunism. However, the propensity for opportunistic behavior may change with a personal history of misbehavior due to changes in intrinsic motivation. If personal history matters to individual incentives for observing norms, then variation in "cultures of corruption" across countries or organizations may reflect adverse shocks to past behavior, rather than just deep moral fundamentals.[1] For example, an individual who becomes corrupt during an economic crisis may persist in corrupt tendencies even after the economy has recovered. Conversely, someone who has behaved well may have more of a stake in maintaining good conduct.

We propose a theory focused on the dynamics of virtue and corruption and study the possibility of persistent patterns of self-reinforcing behavior. The theory also helps understand how the power of extrinsic incentives (e.g., law enforcement) depends on timing, and how small differences in the environment may lead to large differences in behavior. In our model an individual lives for ever and faces in each period a stochastic "temptation," which we think of as an opportunity to increase consumption utility by dishonest means. For concreteness, think of a policeman who in each period faces a bribe offer of random size. The officer receives utility from consumption goods bought with bribe money, but he also values the possibility of maintaining the notion of having "a good heart," i.e., that he is the kind of a person whose nature steers him towards honesty. Another example fitting the model is the Weberian account of the Calvinist Ethic. In that account, a person who does not know his predestination status (saved or doomed) may enjoy being profligate but would also like to maintain or even increase his confidence of having been born saved.

We use a planner-doer model a la Shefrin and Thaler (1981) with a few modifications. The planner is the only deliberate decision-maker, and believes the doer to have one of two types, "good" or "bad", each with a hardwired tendency for either a "good" or a "bad" action (e.g., reject or accept the bribe, respectively). The type of the doer is unknown to the planner, but the planner cares directly about it—a form of self-esteem tied to received values or norms. The planner must then decide in each period whether to attempt to steer the doer towards the good action or to give up and let the doer alone drive behavior; the planner makes this decision knowing that he will rationally update his beliefs about the doer's type upon observing behavior. As done in earlier work (Bénabou and Pycia 2002, Fudenberg and Levine 2006) we conceptualize the planner as the conscious element in the mind, usually linked to the notion of "executive function" and the top-down coordinating role of the prefrontal cortex (PFC). The doer captures subcortical parts of

[1] Fisman and Miguel (2007) present evidence illustrating different national "cultures of corruption."

the brain that exert subconscious influences on behavior.

There are several precedents in the literature of linking the emergence of self-discipline to self-image management. Economic theory has incorporated the idea that people care about self-image (e.g., Rabin 1994, Brekke, Kverndokk and Nyborg 2003, Kőszegi 2006, Bénabou and Tirole 2006, 2011, Cervellati, Esteban and Kranich 2006), and there is experimental evidence that self-concept maintenance limits dishonesty (Fischbacher and Heusi 2008). Economics has also incorporated the notion that behavior is influenced by subconscious impulses (e.g., Bénabou and Pycia 2002, Prelec and Bodner 2003, and Berhneim and Rangel 2004), and that past actions may contain information about the self, creating a role for introspective reputation (Prelec and Bodner 2003, Bénabou and Tirole 2002, 2004, 2011).

Two factors enable past actions to contain information about the doer's type in our model. First, the planner's control over the doer is imperfect, meaning that the planner's attempt to steer the doer towards a "good" action may not succeed. This is consistent with work in neuroscience showing that the PFC carries out its tasks by "biasing signals" that enhance some impulses and inhibit others (Miller and Cohen 2001). This function is not perfect nor acts unopposed; other areas also affect the signals available to circuits that implement actions.[2] Second, the planner has imperfect information regarding the "authorship" of actions: if the action taken by the doer matches the planner's override, the planner cannot be sure whether this was due to the override or to the doer's own drive. The inability of an individual to assign authorship over his externally observable action is consistent with work in psychology and neuroscience. It may appear obvious that we do things "because we want to." However, determining the causal role of conscious deliberation and cognitive override in the actions we take is a formidable inference process for the planner (Wegner and Wheatley 1999). After reviewing the various ways in which the mind may frame its control over actions (and be variably successful), Vallacher and Wegner (1987) conclude that exact attribution of "*this simple input for self-conception—action—is inherently uncertain.*"[3]

Given these ingredients, our first step is to study conditions for the self-esteem motive to induce adherence to a received moral norm for a single decision-maker with time-consistent preferences in a stationary, infinite horizon environment. In our model an override by the planner lowers the variance over future self-image. Thus, a planner with concave self-esteem payoffs will want to override the doer if temptations are low enough. This emergence of self-restraint mirrors results in Kőszegi (2006) (see also Bénabou and Tirole 2002). But it is not obvious how virtue is to respond to the presence of a future comprising an ongoing learning and decision-making process. We show

[2]For example, the Sub Hippocampal area affects PFC influence (Grace et al. 2007). The notion that conscious cognitive control may lose out to automatic subconscious responses to cues is well established, as explained by Camerer, Loewenstein and Prelec (2005). See also Bargh and Chartrand (1999) and Berridge (2003).

[3]See also Nisbett and Wilson (1977) for a discussion on the evidence that there is very poor conscious access to the determinants of decisions.

that the presence of a future increases incentives to resist temptations. We then move on to ask whether good behavior will build its own demise, or whether good acts, by improving self-image, will lead to ever stronger incentives to do good. Such a persistently self-reinforcing path would lend support to the Weberian account of the Calvinist Ethic as an explanation for sustained good behavior.

The key contribution in our paper is to show that, under certain assumptions, a stationary environment must result in a (literally) virtuous circle: good actions improve the self-image, and this in turn further strengthens incentives to resist temptations.[4] This resembles Aristotle's description of virtue as a process of habituation through action. As the beliefs about the self enter the problem as a state variable that is costly to improve (it requires forgoing temptations), they act as a form of capital, which we term "moral capital." We explore the limits of our virtuous circle result. We show that moral capital has non-monotonic effects on behavior when the environment is not stationary due to a finite horizon, or when doers can tremble and as a result changes in moral capital affect the informational content of actions.

A key mechanism in our model is that the incentive to adhere to a social norm stems from a desire to diminish the arrival of information. An individual cannot change his beliefs in expectation, but an attempt to override the doer diminishes the variance of future self-image.[5] The concept of self-esteem utility closest to ours is found in the task choice model of Kőszegi (2006), where individuals derive utility from the self-image of being a highly productive type and where individuals may be risk averse with respect to that self-image.

We offer three applications. One shows how the distribution of moral capital has an impact on aggregate outcomes. In particular, in a society in demographic steady state, shocks in the distant past have long-lasting effects on wrongdoing by affecting the polarization of individual beliefs. In another extension we show that the presence of moral capital affects optimal deterrence schemes through channels absent from traditional models of crime, a la Becker (1963), where "moral costs" are a constant taste parameter. Since individuals with lost self-esteem no longer have an intrinsic motive to self-deter, a myopic social planner will select harsher punishments for repeat offenders. In another extension we study how individuals with different levels of self-image sort themselves into activities with different levels of temptation. We show that high temptation activities (e.g., politics) display more prevalent wrongdoing than low temptation activities (e.g., academia) not

[4] Thus, our work on the role of beliefs as a state variable complements the work of Bénabou and Tirole (2011) analyzing the link between beliefs about the self and incentives for prosocial behavior in a finite horizon setting.

[5] A similar variance-shifting incentive to manipulate information has been shown to arise in different contexts, for instance in Carrillo and Mariotti (2000), where the individual manipulates information for instrumental reasons. Other papers where the individual manipulates information for instrumental reasons are Bénabou and Tirole (2004) and Compte and Postlewaite (2004). The models by Bénabou and Tirole (2006, 2011) are interpretable as capturing both instrumental and intrinsic motives.

only due to the higher temptations but also because they attract those least equipped to resist temptations.

We take both self-esteem and temptations to be genuine sources of utility, unlike the subconscious elements that may alter externally observed actions. This mirrors the approach in Bernheim and Rangel (2004), as does our view that subconscious factors affect actions in ways that are independent of consciously perceived utility.[6] Observable actions may constitute a mistake from the perspective of the planner, placing limits on a revealed preference approach. One implication is that non-choice data (from the neuroscientific to happiness reports) may be relevant to the study of adherence to moral norms. In an online appendix we offer foundations for our model through a more general framework of (imperfect) conscious control over actions that is related to Bernheim and Rangel's (2004) set up.

2 The model

2.1 Basic setup

An individual lives in an infinite horizon world with discrete time, and discounts the future by a factor $\lambda \in (0,1)$. The conscious aspect of the individual is referred to as the "planner," and the subconscious part as the "doer." The planner is the only decision maker in the model, and the one whose welfare we equate with that of the individual. The doer is the default driver of externally observable actions by the individual, but the planner has the (imperfect) ability to override the doer. The planner's override decision entails choosing a value for the control $a_t \in [0,1]$; we will be interested in whether a_t takes the value of 0, in which case we will say the planner "gives up," or 1, in which case we will say the planner attempts to override the doer. It is technically convenient to allow for mixing and let a_t take any value in the unit interval. The doer is characterized by a type $\theta \in \{\theta_g, \theta_b\}$, good or bad, and we will refer to the type of the doer as that of the individual, in the understanding that the type characterizes the doer only. The individual is born with an initial belief that her type is good with probability μ_0.

In each period t the individual faces a nonnegative temptation x_t, drawn independently from a distribution with cumulative density F in the nonnegative real numbers. We assume that F is continuous and strictly increasing with $Ex < \infty$. For example, think of a bureaucrat facing an opportunity for taking a bribe each period. The temptation is the additional consumption utility obtained by consuming the bribe. Given the lack of restrictions on the shape of F, we can assume without loss of generality that utility is linear in x. To see what our reduced-form temptation means,

[6] The authorship inference problem (and the role of self-esteem) differentiates our theory from other planner-doer models, such as those by Bénabou and Pycia (2002), Bernheim and Rangel (2004), Fudenberg and Levine (2006) and Ali (2011).

denote the consumption utility function by v, the consumption available by honest means by c_h, and the additional consumption available by dishonest means by c_d. Then $x \equiv v(c_h + c_d) - v(c_h)$ 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 towards higher temptations reflects an environment where wrongdoing opportunities are relatively more attractive, but not necessarily one where total consumption (including that bought with bribe money) is higher.

The planner consciously perceives an additively separable payoff comprising both the utility from temptations x_t and a utility from beliefs for being good (self-esteem). A planner holding a belief μ by the end of a period enjoys a self-esteem utility $u(\mu)$ in that period. We assume that u is concave, strictly increasing, and bounded for $\mu \in [0, 1]$. This formulation where the individual obtains utility from beliefs in a direct way follows an ego-utility formulation as in Kőszegi (2006). This differs from the usual expected utility formulation where individuals have a von Neumann-Morgenstern utility from outcomes and where beliefs only count as weights associated with different outcome realizations. Beliefs may affect utility directly because they yield a sense of self-worth, or because the person derives utility from anticipatory feelings about future outcomes. Consumption utility x_t is not dependent on type: individuals realize that goods obtained by dishonest means would yield as much consumption utility as those gained honestly. Thus, while individuals enjoy the thought of being good, they draw utility from temptations consumed.

An individual can take one of two *externally observable* actions in a given period: yield to the temptation ($r_t = 0$), or resist ($r_t = 1$). We now explain how externally observable actions are determined.[7] If, after observing the temptation x_t the planner gives up ($a_t = 0$) the doer is left to drive behavior alone, and behavior will match the doer's type: resist if good ($r_t = 1$ if $\theta = \theta_g$), give in if bad ($r_t = 0$ if $\theta = \theta_b$).[8] If the planner attempts an override ($a_t = 1$), then this increases the probability, from zero to $\phi > 0$, that when the doer is bad the externally observable action is resistance. (The override is inconsequential if the doer is good, because the good doer resists for sure.) The parameter ϕ is the probability that the override attempted by the planner overcomes the tendencies of the doer.[9] It captures a causal effect of a deliberate choice by the planner on the action taken externally, and as such it is a manifestation of free will as defined by some psychologists (e.g., Vallacher and Wegner 1987). We will refer to ϕ as the "free will" parameter.

The timing of the problem is depicted in Figure 1.

[7] In the online appendix we show how this model can be obtained from a more general one that explicitly considers the process by which the PFC can bias the signals available to circuits engaged in the implementation of actions.

[8] As in other models with behavioral types (e.g., Kreps and Wilson 1982), the abundance or even objective existence of a good type is not essential; what matters is that players assign it a positive probability.

[9] We can also assume that the planner, rather than giving up, can actively attempt an override in the direction of taking the temptation. The results do not change as long as an override in the direction of good behavior leads to a reduction in the variance over beliefs.

[**FIGURE 1 HERE** - Timeline]

2.2 One-period problem

First consider an individual who lives only for one period and faces a temptation of size x. If the planner chooses to give up, the temptation is taken only if the doer is bad and resisted only if the doer is good. The externally observable action r will then fully reveal the doer's type. With probability μ_0 the individual will become certain of having a good type ($\mu_1 = 1$), and with complementary probability $1 - \mu_0$ certain of having a bad type ($\mu_1 = 0$). The expected posterior is, of course, equal to the prior μ_0.

Attempting an override means that if the doer is bad, but the override works, then the person will see himself pass on the temptation. Then, successfully resisting would be compatible both with having a good type, and with having a bad type that was successfully overridden. The planner will then have the posterior belief $\mu_1 = b(\mu_0)$, where

$$b(\mu) \equiv \frac{\mu}{\mu + (1-\mu)\,\phi} \tag{1}$$

is the Bayesian update after attempting override and resisting the temptation. Beliefs about self only improve with a history of good behavior under imperfect override ($\phi < 1$); if $\phi = 1$ then $b(\mu_0) = \mu_0$ and override only preserves existing beliefs.

The expected belief when attempting an override is also, of course, equal to the prior, so attempting an override cannot improve the expected posterior. Although the expected belief cannot be manipulated, expected utility can be, because an override attempt yields a "gamble" over future self-image that is less risky. Attempting an override results in a lower variance over posterior beliefs, $\mu_0^2 (1-\mu_0)(1-\phi) / [\mu_0 + (1-\mu_0)\phi]$, than leaving the doer alone, in which case the variance is $\mu_0 (1-\mu_0)$.

We denote expected utility as $U(a, x, \mu_0)$. If the planner chooses gives up ($a = 0$) the realized action will reveal the true type. This means receiving the maximum self-esteem utility with probability μ_0, and the minimum self-esteem utility combined with the consumption utility from the temptation with probability $1 - \mu_0$. Resulting expected utility is

$$U(0, x, \mu_0) = \mu_0 u(1) + (1-\mu_0)(u(0) + x). \tag{2}$$

Attempting override ($a = 1$) succeeds in keeping the individual uncertain with probability $\mu_0 + (1-\mu_0)\phi$ (i.e., in the event that the doer is good, and also in the event that the doer is bad but the override is successful); with complementary probability $(1-\mu_0)(1-\phi)$ the doer takes the temptation and reveals a bad type. Expected utility is

$$U(1, x, \mu_0) = (\mu_0 + \phi(1-\mu_0))\, u(b(\mu_0)) + (1-\mu_0)(1-\phi)(u(0) + x). \tag{3}$$

Attempting an override is optimal iff $U(1, x, \mu_0) \geq U(0, x, \mu_0)$. The level of temptation enters both sides linearly, and at different slopes, so we can solve for a unique threshold value

$$x < \frac{[\mu_0 + (1 - \mu_0)\,\phi]\,u\,(b(\mu_0)) - [\mu_0 u\,(1) + \phi\,(1 - \mu_0)\,u(0)]}{\phi\,(1 - \mu_0)} \equiv x_0^*. \tag{4}$$

The planner will attempt to steer the doer away from the temptation if the latter is below the cutoff x_0^*. This cutoff is decreasing in the probability $(1 - \mu)\,\phi$ that the override causes the temptation to be resisted (captured by the denominator) and increasing in the expected utility gain in terms of self-esteem utility (as captured by the numerator). The first effect might be slightly surprising: a more confident planner (one with a higher μ_0) attempts overrides of larger temptations because he is more confident that the override is redundant and unlikely to cause him to forgo the temptation. This reflects the fact that the planner, although mindful of the benefits of self-esteem, values consumption as well.

The numerator of (4) shows that the cutoff x_0^* is strictly positive if and only if u is strictly concave. Some temptations will be resisted if and only if the individual is risk averse over her self-image. In the risk neutral case the cutoff is exactly zero, and the planner never attempts an override. Furthermore, it is straightforward to show that the cutoff has limiting value

$$\lim_{\mu \to 1} x_0^*\,(\mu) = u(1) - u(0) - u'(1) \tag{5}$$

which, by the concavity of u, is in $[0, u(1) - u(0)]$. We later rely on the constant relative risk aversion (CRRA) utility function (11) with a coefficient of relative risk aversion $\rho \in [0, 1)$. With this assumption the cutoff (4) reduces to

$$x_0^* = \frac{\mu_0}{\phi\,(1 - \mu_0)} \left(b(\mu_0)^{-\rho} - 1\right). \tag{6}$$

A planner who is more risk averse with respect to beliefs will try to resist higher temptations.

The idea that individuals may want to manipulate the higher moments of a distribution of beliefs has emerged in various settings. In Carrillo and Mariotti (2000) an individual has commitment problems and knows that her behavior, current and future, depends on beliefs about the level of future costs that current consumption creates. This individual cannot alter her expected beliefs, but changing the distribution of beliefs may be beneficial when actions depend on the position of beliefs relative to specific, decision-relevant, thresholds. Related ideas emerge in Kamenica and Gentzkow's (2011) study of persuasion. In Kőszegi's (2006) paper on overconfidence and task choice the individual manipulates the arrival of information for intrinsic reasons. Risk aversion over future beliefs about competence drives incentives for information manipulation. In Kőszegi's study the objective is to understand the emergence of overconfidence, while we study adherence to norms. In addition, we assume a smoothly concave ego-utility rather than a step function. The link between risk aversion and the demand for information is present also in Bénabou and Tirole

2002 (see especially p. 906-7). The known insight that risk aversion leads to information aversion foretells the driving force behind our result that the shape of self-esteem utility drives adherence to norms. But the question remains open as to whether the insight remains valid in the dynamic version of our problem, where the future matters in non-trivial ways.

Does the presence of a future affect current adherence to norms, and is adherence necessarily strengthened by the accumulation of moral capital? We study this problem with a dynamic model. We show later that a finite horizon confounds the effect of changes in the individual's beliefs with those of an approaching terminal date. Therefore we use an infinite horizon model for the main analysis.

2.3 Repeated problem

We call individuals with a belief $\mu_t \in (0,1)$ *uncertain*, while those who know their type for sure, $\mu_t \in \{0,1\}$, are *certain*. A planner facing a temptation x_t who is certain about the doer's type has no meaningful choice. If the type is good there is no need, nor point, to attempt an override: the payoff will be $u(1)$ in every period. If the type is bad, there is no point in attempting to resist temptations, as there is no self-esteem to protect, so the period payoff will always be $u(0) + x_t$. Using $\lambda \in (0,1)$ to denote the planner's discount factor, we can pin down the expected value facing individuals who are certain,

$$\begin{array}{c} EV(0,x) = \frac{u(0)+Ex}{1-\lambda} \equiv V_0, \\ EV(1,x) = V(1,x) = \frac{u(1)}{1-\lambda} \equiv V_1. \end{array} \tag{7}$$

For the uncertain planner, choosing to give up brings beliefs to an absorbing state of certainty, where expected value is constant. Bad types still face uncertainty over the realized temptation in each period (V_0, unlike V_1, depends on expected temptations). The choice is between period utilities (2) and (3) plus the associated continuation values. While x is stochastic, the planner exerts partial control over the evolution of the belief μ, in the same way as in the one-period problem. Written in recursive form, the dynamic problem is

$$V(\mu,x) = \max \left\{ \begin{array}{c} U(1,x,\mu) + \lambda \left(\begin{array}{c} [\mu + (1-\mu)\phi]\, EV\left(b(\mu), x'\right) + \\ + (1-\mu)(1-\phi)\, V_0 \end{array} \right), \\ U(0,x,\mu) + \lambda \left[\mu V_1 + (1-\mu) V_0\right] \end{array} \right\}, \tag{8}$$

where the first line inside the max operator reflects attempted resistance, the second one giving up, and x' is the as yet unknown temptation next period. Writing out expressions (1), (2), (3), and (7), and gathering terms, the dynamic problem can also be written as

$$V(\mu,x) = \max \left\{ \begin{array}{c} \left(\begin{array}{c} [\mu + \phi(1-\mu)] \left[u\left(\frac{\mu}{\mu+(1-\mu)\phi}\right) + \lambda EV\left(\frac{\mu}{\mu+(1-\mu)\phi}, x'\right) \right] \\ + (1-\mu)(1-\phi)\left[u(0) + x + \lambda EV(0,x')\right] \end{array} \right), \\ \mu\left[u(1) + \lambda EV(1,x')\right] + (1-\mu)\left[u(0) + x + \lambda EV(0,x')\right] \end{array} \right\}. \tag{9}$$

The continuation value is evaluated at the same updated beliefs that yield current period self-esteem utility; hence their isomorphic appearance in the recursion. The two choices in the max operator show the trade-off facing the planner. As in the one-period problem, the override attempt creates a gamble over an interior belief and a zero belief, with respective probabilities $\mu+\phi(1-\mu)$ and $(1-\mu)(1-\phi)$. Giving up creates a higher variance gamble between extreme beliefs, with respective probabilities μ and $1-\mu$. The difference now is that the gamble involves continuation payoffs, also dependent on future beliefs.

It is straightforward to show that the map in (9) satisfies the conditions for existence of a unique (continuous) value function.[10] This dynamic formulation yields the first result,

Lemma 1 *An optimal policy a^* exists and can be represented by a cutoff function $x^*(\mu_t)$ such that if $x_t \leq x^*(\mu_t)$ the planner attempts to override the doer ($a_t = 1$) and gives up otherwise ($a_t = 0$).*

Proof. The map in (9) is of the form $V = \max\{\varpi, \varsigma\} = \max_{a\in\{0,1\}}\{a\varpi + (1-a)\varsigma\}$, where ϖ and ς are continuous functions of (μ, x). The objective is also continuous in the control and the choice set is compact, so the Theorem of the Maximum implies that an optimal policy correspondence a^* exists (and is upper hemicontinuous). Straightforward algebra shows that override ($a_t = 1$) is optimal iff

$$x \leq x^*(\mu) \equiv \begin{array}{l} \frac{[\mu+\phi(1-\mu)]u(b(\mu))+(1-\mu)(1-\phi)u(0)-[\mu u(1)+(1-\mu)u(0)]}{(1-\mu)\phi} \\ +\lambda\frac{[\mu+\phi(1-\mu)]EV(b(\mu),x')+(1-\mu)(1-\phi)V_0-[\mu V_1+(1-\mu)V_0]}{(1-\mu)\phi} \end{array}, \tag{10}$$

otherwise giving up ($a_t = 0$) is optimal. The term $x^*(\mu)$ depends on parameters and on expectations involving the unique value function. Thus $x^*(\mu)$ is constant in the space of current temptations $[0, \infty)$, and represents the unique (cutoff) policy function for the individual. ■

The cutoff $x^*(\mu)$ is strictly positive if the lower risk gamble (over both the current and the continuation payoff) stemming from the override attempt yields strictly higher expected utility. In the one period problem (captured by the first line in (10)) the cutoff is, as shown before, positive whenever u is concave. If the value function V were concave it would be immediate that cutoffs must be positive in the dynamic problem as well, since the added term (captured by the second line in (10)) is isomorphic to the one expressing the static trade-off. However, the map in (9) does not preserve concavity, so we rely on an alternative argument based on the sequential optimality of cutoffs to obtain the following,

Proposition 1 *If self-esteem utility u is strictly concave, then optimal policy $x^*(\mu)$ is strictly positive for all $\mu \in (0, 1)$.*

[10] See the online appendix.

Proof. The first line in (10) gives the optimal cutoff in the one-period problem (4), which is strictly positive if u is strictly concave. Thus it suffices to show that the second line in (10) cannot be negative. Consider the continuation value $b(\mu) V_1 + (1 - b(\mu)) V_0$ under a (putatively) non-optimal policy of always giving up; it is straightforward to check that the numerator in the second line in (10) is zero when evaluated at this continuation value. Optimizing behavior implies $EV(b(\mu), x') \geq b(\mu) V_1 + (1 - b(\mu)) V_0$, guaranteeing $x^*(\mu) > 0$. ■

While the planner remains uncertain he attempts an override to resist every temptation x_t such that $x_t \leq x^*(\mu_t)$. If the individual is risk averse over beliefs about type, then as long as he remains uncertain the personal history of cutoffs or "moral standards" will evolve according to $x^*(\mu_0), x^*(\mu_1), \ldots$. The resistance continues until the first time he either faces a temptation above $x^*(\mu_t)$ or, if the doer is bad, the first time the override fails, which has the independent probability $1 - \phi$ every period.

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

Remark. Individuals do not develop infinitely high standards It can be seen from (10) that the maximum of $x^*(\mu)$ is finite. As long as the planner is uncertain, some temptations must be high enough that he would not attempt to override the doer. Discounting and $Ex < \infty$ together guarantee that the expected value EV is bounded above by the present value of getting the best possible expected period utility forever. The immediate payoff differential from not attempting override, $U(0, x, \mu) - U(1, x, \mu)$, is linear and increasing in x for all $\mu \in (0, 1)$, so a sufficiently high x will make it optimal to not override.

Three additional observations are in order. First, wrongdoing provides conclusive evidence of bad type, so the stochastic process for beliefs differs from many learning setups where revisions become smaller as more information is obtained. Here, as long as there is any uncertainty over the type, a sudden and ever higher "fall from grace" always remains a possibility, although such a fall keeps getting more unlikely as the record of good behavior gets longer.

Second, in a world where the average temptation is sufficiently high, having a good type is bad news for expected utility. The benefit of having a good type is the self-esteem, but having a bad type has the benefit of increasing the opportunities for consumption. Thus, both period utility and present value V may be increasing or decreasing in μ, depending on whether the average temptation Ex is large or small relative to the maximum utility gain from self-esteem, $u(1) - u(0)$.

Third, further characterizing the policy function in the context of the recursive formulation is hard. In addition to the map in (9) not preserving concavity, intuitive approaches that could help characterize the optimal policy in the context of dynamic programming are not applicable to our

problem.[11] Therefore, our strategy for analyzing the optimal policy will rely on the less elegant approach of dealing with discounted sums of utility flows. We relegate the proofs to the appendix.

Optimal policy

From now on we impose a CRRA functional form for self-esteem utility,

$$u(\mu) = \frac{\mu^{1-\rho}}{1-\rho}, \tag{11}$$

where $\rho \in [0, 1)$ is the coefficient of relative risk aversion. We now establish a central result of our paper.

Proposition 2 *Policy monotonicity.* $x^*(\mu)$ *is strictly increasing in* $\mu > 0$ *and has a finite limit as* $\mu \to 1$.

Proof. See Appendix. ■

This result can be understood by considering how the expected costs and benefits of attempted resistance vary with μ, while holding x fixed. The benefit-cost ratio can be seen in (10), separately (but in the same functional form) for current and future periods. Given that the uncertain planner faces qualitatively the same problem every period, the intuition is best explained by reference to the trade-off in current period alone. The benefit of override is the expected utility gain from risk reduction, as reflected in the numerator of the benefit-cost ratio for the current period—see the first line in (10). This benefit is at first increasing, but eventually decreases in μ, because uncertainty is highest at intermediate values of μ. The cost, in turn, is proportional to the probability $\phi(1-\mu)$ that the attempt to resist causes the individual to forgo the temptation. This cost is reflected in the denominator of the benefit-cost ratio, and is linearly decreasing in μ. For low values of μ the result is obvious: higher μ means higher benefit (more risk reduction) and lower cost (fewer forgone temptations), so the benefit-cost ratio of an override can only increase. Eventually both benefits and costs are decreasing in μ, so the result is non-obvious. The proof shows that the rate ϕ at which cost decrease is faster than the rate at which the benefits of risk reduction can decrease for all values of μ.

While the optimal policy $x^*(\mu)$ is stationary, the temporal evolution of personal standards has a direction. As long as the planner remains uncertain of the doer's type, the belief μ_t will keep increasing according to (1), so, by virtue of Proposition 2, the effective cutoff $x^*(\mu_t)$ will also increase over time. This yields,

[11] With a mononote increasing value function, an approach to show the policy function must be increasing would be to establish the supermodularity in (x, μ) of the per period payoff and the transition function describing the probabilities over future beliefs. However, our value function is not necessarily monotonic and it is easy to show the transition is not supermodular.

Remark. Moral growth. Externally observable actions of resistance to temptation increase confidence in having the good type, and in turn increase the subsequent likelihood of resistance.

This implication of Proposition 2 echoes Aristotle's characterization of virtue as a process of habituation through action, where the exercise of virtue makes it more likely that virtuous behavior obtains subsequently (see Nichomachean Ethics 1998); a literal virtuous circle. The chronological age implications of the result should not be taken literally – age in our model reflects the number of temptations a person has faced before, so the interpretation of periods as calendar time implicitly assumes that temptations arrive at the same rate during the lifetime.

Note that the result in Proposition 2 obtains regardless of whether the value function is increasing or even monotonic. The top panel of Figure 2 depicts the expected value function, $V(\mu) \equiv E_x V(\mu, x)$, and the associated optimal policy function, for three cases.[12] In the first case temptations are relatively small, so good types have a higher expected utility than bad types, $V_1 > V_0$. The resulting value function is increasing. The second case is the opposite, with relatively high temptations, $V_1 < V_0$. This case results in a decreasing value function. In the third case expected utilities are balanced, $V_1 = V_0$. In this case the value function is non-monotonic, with an interior maximum (the same is true for the roughly balanced case, $V_1 \approx V_0$). The bottom panel of Figure 2 depicts the (always increasing) policy function x^* for the same cases.

[**FIGURE 2 HERE** - Value function and optimal policy]

It is worth noting the role of the assumption of imperfect override. With perfect override ($\phi = 1$) beliefs do not change ($b(\mu) = \mu$) and there can be no moral growth. However, the planner would still attempt to override the doer when the temptation is low enough, due to the risk aversion over beliefs. The individual would keep resisting until a sufficiently high temptation is encountered, at which point the planner accepts the "final gamble" and finds out the doer's true type.

2.3.1 Comparative statics

Next we study the dependence of the optimal policy on the level of anticipated temptations and on individual time preference. To introduce higher anticipated temptations, consider a uniformly shifted distribution such that $F_\varepsilon(x) = F(x - \varepsilon)$ for $x \geq \varepsilon$ and zero otherwise. The parameter ε defines an increase in temptations that implies a shift in the sense of first order stochastic dominance.

Proposition 3 *$x^*(\mu)$ is decreasing in the level of anticipated temptations, and increasing in the discount factor λ, at every $\mu \in (0, 1]$.*

Proof. See Appendix. ■

[12] For the numerical examples we assume that temptations are distributed exponentially.

When the planner expects higher temptations in the future he will choose less stringent moral standards today. This is because falling to a temptation today would wipe out the person's moral capital and direct him to "a life of crime." The higher the distribution of temptations, the more attractive is that life relative to slowing learning for the sake of self-esteem.

Note that a more benign environment in the sense of lower temptations will impact behavior in two ways. A direct effect is that, given the individual's cutoffs, a less tempting environment makes it less likely that a high enough temptation will materialize so as to induce the planner to give up. The indirect effect is that the expectation of a more benign environment leads the individual to resist even larger shocks, complementing the direct effect. This positive feedback suggests that small differences in the environment can generate larger departures in the overall level of wrongdoing.

Higher moral standards can be interpreted as a type of investment; hence the term "moral capital." The temptation would be available immediately, but a successful resistance improves (in expectation) the entire future path of self-esteem utility flows. It is therefore natural that a higher discount factor increases cutoffs.[13]

By contrast, the relation of the efficacy of the planner's intervention ϕ and moral standards is not clear cut. On the one hand, if the planner attempts an override, then a higher ϕ reduces the probability of finding out the true type this period, thus reducing the variance of updated beliefs. On the other hand, there is also an effect in the opposite direction, because the attempt is now more likely to preclude the enjoyment of a temptation.

2.4 Finite horizon

We now explore forces that may create non-monotonic effects of moral capital on individual moral standards. Consider a finite horizon setting. Figure 3 shows the numerically obtained optimal policy for the last three periods of a life with a known end period. (The one period case can be interpreted as the last period of a finite horizon life.) Just like in the infinite horizon case, the optimal cutoff is monotone increasing in beliefs in any given period. At the same time, the cutoff is higher the longer is the remaining lifetime, for any given level of beliefs μ. This is consistent with our comparative static result on time preference.

Now let us consider how moral standards may evolve as we follow an individual who remains uncertain and keeps resisting temptations during the final periods of a finite life. As he moves forward in time, a shrinking horizon reduces the relative weight of the future, and this induces a weakening in personal moral standards for any given level of beliefs. At the same time, his

[13] The result that higher patience results in less wrongdoing matches the finding by criminologists that the inability to take the future into account plays key a role in the disposition toward crime. See Gottfredson and Hirschi 1990, and Nagin and Paternoster 1993.

consistent rejection of temptations increases his confidence in having a good type, and this works towards higher standards. With t closer to final period T, $x_t^*(\mu)$ is lower at any μ, but μ_t is higher than μ_{t-1}. The direction of the temporal change in moral standards is therefore ambiguous. Figure 3 shows two examples for the evolution of beliefs, in accordance with (1), depicted by the successive vertical dashed lines. For the individual who enters the final three periods with a relatively low initial belief the impact of "Aristotelian" moral growth dominates at first, and the effective cutoff $x_t^*(\mu_t)$ is at first increasing. For another individual, who enters with a relatively high belief, the impact of the shrinking horizon dominates and the cutoff is decreasing over time.

[**FIGURE 3 HERE** - Finite horizon]

In sum, the stationary environment offered by the infinite horizon is crucial for obtaining unambiguous results about personal moral growth, because adherence to standards is an investment in moral capital that yields returns in the future. The effect of an approaching terminal date reduces the return to such investment and acts as a confounder.

2.5 Fallible types

So far we have assumed that individuals have a dichotomous and absolute view of the nature of good and bad: they do not believe that it is possible for a person to be "a little bit corrupt." This implies that wrongdoing, if observed even once, provides conclusive evidence of a bad type. This Manichaean interpretation of good and bad is not just a useful simplifying assumption (which it also is) but has a long history in religion and philosophy. In this subsection we investigate the impact of allowing for a less stark formulation, by assuming that individuals believe that types are fallible: under no override, bad types may take good actions, and good types may take bad actions.

In particular, we now assume that in the absence of a successful override by the planner, the good doer selects the good action with some probability $\gamma_g < 1$ and the bad type takes the good action with a lower probability $\gamma_b \in [0, \gamma_g)$. Thus $1-\gamma_g$ and γ_b are "error rates" for the good and bad doers, respectively. We assume that the deviations from typical behavior are independent across time and that the planner's override prevails independently with probability ϕ. Thus, when the planner selects $a = 1$ then the probability of a type $j \in \{g, b\}$ doer taking the good action is $\gamma_j + \left(1-\gamma_j\right) a\phi$. The planner's override increases the probability of a good action.

The updating rules become slightly more complicated. No evidence is conclusive of either type, but bad actions will cause the beliefs to be revised downwards, and good actions upward, and revisions are larger when the error rates are smaller.[14] This expanded model nests our baseline

[14] In particular, the update rule after observing a bad action is now $\mu_{t+1}|(\mu_t, r_t = 0, a_t) = \frac{\mu_t\left[1-\gamma_g(1-a\phi)-a\phi\right]}{\mu_t\left[1-\gamma_g(1-a\phi)-a\phi\right]+(1-\mu_t)[1-\gamma_b(1-a\phi)-a\phi]}$ and after observing a good action $\mu_{t+1}|(\mu_t, r_t = 1, a_t) = \frac{\mu_t\left[\gamma_g(1-a\phi)+a\phi\right]}{\mu_t\left[\gamma_g(1-a\phi)+a\phi\right]+(1-\mu_t)[\gamma_b(1-a\phi)+a\phi]}$.

model as a special case where $\gamma_g = 1$ and $\gamma_b = 0$. In any case, beliefs will eventually converge to be arbitrarily close to the truth.

Figure 4 shows the numerically obtained optimal policy when individuals believe that good types can "tremble," for several cases of the "error rate." The policy functions reveal that when the belief in one's goodness gets sufficiently close to certainty, then the cutoff selected by the planner begins to decrease. The reason is intuitive. If the planner is almost certain that the doer is good, but also believes that the good doer can make mistakes, then the planner also anticipates that after observing a bad action he will attribute the bad action to a mistake by the good doer rather than to the doer being bad. This plausible deniability reduces the incentives to maintain adherence to high standards. However, if good types are believed to make mistakes very rarely (γ_g close to one) then the optimal policy is close to that of the basic model, and the level of beliefs beyond which moral standards begin to decrease gets very close to one. Interpreting the discount factor as including a hazard rate for survival, a good type following such a policy could have a very low chance of ever making it to the decreasing part of the policy function.

[**FIGURE 4 HERE** - Fallible types]

It is worth noting that very high self-confidence may also lead to more lax moral standards for reasons other than those captured here, but which have received attention in psychology, such as feelings of invincibility or a loss of perspective.

2.6 Discussion

By focusing on conditions that ensure stationarity, we are able to isolate a result where moral capital always reinforces intrinsic motivation. That result breaks down with a finite horizon or due to factors that alter the informational content of actions. With fallible types, changes in moral capital change the relative role the planner assigns to luck versus the doer's type. This complements the recent analysis of "beliefs as assets" by Bénabou and Tirole (2011). We now discuss other implications of our model and further discuss related literature.

Rationalizing the Calvinist ethic The Weberian account of the Calvinist Ethic (Weber 1905) has been extremely influential at shaping views on differential socioeconomic performance, and been called "*the most important sociological thesis of all time*" (Rubinstein, 1999).

In the Weberian account, a person not knowing his predestination status (saved or doomed, the types of the doer) may enjoy being profligate but would also like to maintain or even increase his confidence of having been born saved. The presumably inevitable consequence is that everyone has an incentive to try and live like a saved person would. There are two problems with that account. The first is that, if anyone can mimic a "saved" person, it is unclear how a good introspective

reputation can be developed. The second problem is that if past good actions can improve self-image, but a better self-image could weaken the incentives to adhere to norms, then it is difficult to explain *sustained* good behavior. The study of the Calvinist Ethic from a modeling perspective was pioneered by Prelec and Bodner (2003) and Bénabou and Tirole (2004), in papers where individual behavior results from non-cooperative interactions between fully-strategic temporal selves. They study conditions under which the person may improve his confidence of being saved even when knowing that he has a motive to try to act like one.

Our model allows an interpretation of the Calvinist ethic from a dynamic perspective, based on different assumptions about the sources of impulse control. Our approach involves a single decision-maker whose visceral impulses are captured as behavioral types with no strategic intent. Our approach can explain the possibility of increased confidence in salvation through the planner's imperfect override capability and inability to assign authorship to actions. But where our model helps more distinctly is with the second problem—when do gains in moral capital reinforce good behavior—which is eminently dynamic, and which requires a clean isolation of the effects of moral capital.

Our analysis shows that the reinforcing effects can be persistent rather than self-defeating. The self-reinforcement of virtue is more likely with a longer time horizon and with a lower chance that a bad action can be attributed to a "mistake" by the good type. When the environment is stationary (infinite horizon) and when the informational content of actions remains constant across levels of the state variable (infallible good types), then moral capital has unambiguously reinforcing effects. Thus, the improvement in introspective reputation following good actions, and the reinforcing effect of introspective reputation on further incentives to behave like a "saved" type would, provide a way to understand how the Calvinist ethic could yield sustained good behavior.

Introspective reputation, time consistency, and planner-doer approaches Our model joins various others in the study of introspective reputation. Learning about some aspect of the self is possible when the person updates by conditioning on an "incomplete" set of events. In Bénabou and Tirole's (2004) model, for instance, the individual forgets the motivations that led to an act. For updating purposes, this has the same effect as the planner not knowing whether an override attempt has worked. In their model, the person forgets, but made a fully conscious decision that reveals the person's preference at the point in time when the decision was made. In our case, as in Bernheim and Rangel (2004), externally observable actions may constitute a mistake from the person's perspective, which places limits on a revealed preference approach to study motivation.

It is worth remarking that in our model preferences are time-consistent. Planner-doer models with time-consistent preferences can generate behavior that resembles what obtains in models where time inconsistency is built-in by assumption, a point made by Bénabou and Pycia (2002)

and Fudenberg and Levine (2006). Not all predictions will be similar however. Settings where time-inconsistency is assumed tend to generate a demand for commitment. The individual would typically like to face lower temptations or have a lower vulnerability to them. This is not necessarily true in our set up. First, when the expected value of temptations is sufficiently high, then having a good type (which is good for "commitment") results in lower expected utility. Second, in our model the induced preferences over environments are different. For example, if two activities entail different parameters for the salience of the present, as captured by the hyperbolic discounting parameter β, then individuals with time-inconsistent preferences would always choose the activity with the higher β. Another possibility, in the context of Bénabou and Tirole (2004), is to capture temptations with the cost of forgoing a craving. In their model individuals would always prefer those costs to be lower, while in our model individuals may prefer the distribution of temptations to be higher. As we will discuss later in the career choice application, it is precisely the individuals who are less confident of having a good type (and who should have the stronger demand for commitment) who turn out to have a stronger preference for a high temptation activity.

The modelling approach closest to ours in the literature is perhaps that in Bernheim and Rangel (2004), so it may be helpful to further clarify the connections between our theory and theirs. Bernheim and Rangel posit an individual who faces cues from the environment, and has a level of susceptibility M to these cues. When M lies above a given threshold M^T the individual falls in a "hot" state where he consumes a substance regardless of any conscious attempts to do otherwise. The susceptibility M in each period depends on a probabilistic state ω as well as on a "lifestyle" action a_t taken at the beginning of each period. The essential role of the lifestyle action a_t is to shift susceptibility and make it less likely that the individual enters the hot state. In their model, override fails completely in the hot state. With this similarity comes a related difference: in their model the override failure (having entered a hot state) is known to the individual while in ours it is beyond the reach of consciousness.[15] Thus, the individual in our model is less cognizant of the possible values of each argument affecting M, although he does know his own choice of a. Conditional on that difference, our action a can also be interpreted as a lifestyle choice.[16] The most obvious difference is that the history of consumption plays no role in our model, and we focus instead on belief-driven sources of persistence. (This also sets our theory apart from early models of addiction like that by Becker and Murphy (1988) where past consumption alters future marginal

[15] While for a phenomenon like addiction knowledge of being in a hot state is plausible, for actions that may be affected by subtle cues the possibility of unconscious operation of those cues is plausible as well. It is well known that experimenters in psychology can prime subjects and induce variations in behavior even when the cues utilized are below the threshold of consciousness. Also, although the cue itself may at times be consciously perceived, the individual may not understand how the cue affects behavior relative to a counterfactual situation where the cue is absent. Such cues can play a role in relation with dishonesty. For example, Gino and Pierce (2008) show that money-related visual cues induce cheating behavior.

[16] We thank an anonymous referee for pointing out this intepretation of the model.

utilities). A less obvious, but important, difference is that in Bernheim and Rangel's theory the individual lacks any power to select his desired action in the hot state, but has full power in the cold state. But this does not address the question of how we ever select actions. In the online appendix we present a framework that unpacks what it means to choose an externally observable action, by conceptualizing the initial, and in our model unobservable, action a, as an exertion of the will aimed at overriding impulses. This helps rationalize the role of the types and override invoked in our basic model.

3 Applications

3.1 Wrongdoing in the Aggregate

With a small extension, the individual level model can be used to analyze the rate of wrongdoing in a large population, and its sensitivity to past shocks. Consider first a single cohort of mass one, with individuals born into age $t = 0$ with initial belief $\mu_0 \in (0, 1)$ (that may or may not be equal to the true population share of good types). Tracking such a cohort allows us to characterize what would happen if we could "populate" all the possible histories a person can experience, and show how aggregate wrongdoing depends on the dispersion in the distribution of individual moral capital. Assume that temptations are independent both across time and across individuals. The share of certain individuals can only increase over time, as more and more individuals encounter a temptation above the cutoff or a failed override. The only ones to resist temptations at age t are those who either have the good type, or those who, despite being bad, end the period with interior belief $\mu_{t+1} \in (0, 1)$. Therefore the wrongdoing rate of the cohort at age t is the probability that an individual has become certain of being bad by the end of age t:

$$\Pr(r_t = 0) = (1 - \mu)\left(1 - \phi^t \prod_{\tau=0}^{t-1} F(x^*(\mu_t))\right). \tag{12}$$

As the cohort ages, the age-dependent term approaches zero and the wrongdoing rate converges to the share of bad types, $1 - \mu$. The beliefs of infinitely lived individuals converge inevitably to the truth. Thus, the distribution of beliefs goes from being degenerate at μ_0 to entirely polarized at 0 and 1, and this polarization occurs along with an increase in the wrongdoing rate. Endogenous moral standards slow the process.

Now we want to study the effect on aggregate wrongdoing of past shocks to the distribution of temptations. As time goes by, however, two things occur simultaneously: shocks recede in the past, and individuals age. To isolate the role of past shocks it is necessary to approximate an "ageless" cohort, and to achieve this we consider a society made of infinitely many generations in demographic steady state. Assume that the discount factor reflects a constant death rate, and that there is a constant birth rate of new uncertain individuals. As we show in our working paper, there

exists a demographic steady state in which older cohorts have exponentially smaller population shares. With finite lifespans (of uncertain length) a substantial fraction of bad types never fall, because they die first.

Now consider two societies with the same fundamentals (μ, ϕ, λ, F), one of which suffers an unexpected shock to the distribution of temptations for one period. For example, a macroeconomic shock lowers baseline consumption and thereby increases the additional consumption utility from any given stealing opportunity. In our setup this means that the distribution of temptations is higher, in the sense of stochastic dominance. The immediate impact of the shock is that an unusually large fraction of each living cohort encounters a temptation above their cutoff. The bad types who fall because of the shock lose their moral capital and do not return to good ways after the shock is over. Bad shocks that caused a higher share of people to give in to temptations in one period yield higher wrongdoing rates for every subsequent period in finite time. Wrongdoing rates in the shocked society only converge to the baseline levels of the "normal" society in infinite time, once all cohorts alive at the time of the shock are replaced. Thus, discrepancies in wrongdoing across societies reflect bad luck in the past, rather than differences in moral fundamentals.

If the birth prior μ_0 equals the true probability of having a good type, a bad shock does not affect the average belief, but makes its distribution more polarized for all subsequent periods in finite time–thus, the higher wrongdoing rate in the shocked society is related to the higher second moment of the distribution of individual moral capital.

According to Proposition 2, higher initial beliefs lead to stronger resistance to temptations, which suggests a useful social role for indoctrination. Suppose the aim is to minimize wrongdoing. The strength of individual resistance to temptations is increasing in the confidence of having a good type, so inculcating a high initial belief μ_0 on the youth would reduce the aggregate wrongdoing rate regardless of the true μ. It is obvious that wrongdoing would be reduced by a reduction in the available temptations, but, more surprisingly, wrongdoing rates could also be reduced by introducing an additional contrived temptation on the youth. Suppose that the society is able to label some consumption opportunity as a temptation to be avoided (i.e., a taboo good). To be useful in building moral capital, the size of this temptation has to be below the cutoff of the inexperienced individual $x^*(\mu_0)$, and and be socially innocuous in the sense that consuming it does not constitute part of the wrongdoing that is being minimized. The benefit of this contrived temptation is that those who successfully resist it will be stronger when they meet their first real temptation (they have a cutoff $x^*(\mu_1) > x^*(\mu_0)$). The cost is that some bad types now fall earlier than they otherwise would have. Depending on parameters, this can reduce the steady state wrongdoing rate in the society.[17]

[17] This idea is developed in the working paper version, see subsection "Taboos and Rituals."

3.2 Punishing repeat offenders

We now investigate whether and how the presence of moral capital may affect optimal extrinsic deterrence schemes. The main idea is that the presence of endogenous moral capital affects the effectiveness of deterrence depending on its timing relative to criminal history, so optimal deterrence will vary with individuals' criminal record; in addition, and perhaps more interestingly, the effect depends on the time horizon of the social planner. In order to get the essential message across as cleanly as possible, we impose a number of simplifications.

Consider a social planner facing a single cohort of a given age, the members of which survive from one period to the next with probability $\lambda \in (0,1)$. The planner knows the past behavior by all agents and wants to minimize aggregate wrongdoing going forward. The planner has a one-time capability to impose punishment on those who do wrong in the current period, an event that is detectable with some probability. Should the social planner make punishments contingent on the offender's personal history?

For simplicity,we subsume the probability of detection and the intensity of punishment in a composite expected punishment variable that the planner controls. N_r and N_f denote the expected punishment to be imposed respectively on the repeat and the first-time offenders that seize a temptation in the current period. The net expected return from seizing a temptation x is therefore $x - N_r$ for the repeat offender and $x - N_f$ for the first time offender. Note that a repeat offender is already certain of having a bad type, while the first time offender is, before committing a crime, uncertain. Making punishment contingent on an offense being the first is equivalent to making it contingent on the offender being uncertain. We restrict attention to bad types (the only ones that do wrong) and normalize their mass to 1.

To make things interesting, assume that raising expected punishments is costly to the social planner, as captured by an increasing and convex function $c(N_r + N_f)$. This captures a world where threatening with more likely and intense punishment is costly because it requires stronger detection and punishment capabilities.[18] To cleanly separate the time preference of the planner from that of the population, we assume that the planner discounts the future according to the factor $\delta < 1$, while individuals have a survival rate λ and (to save on notation) do not further discount time. Also, for technical reasons, we assume here that larger temptations are less common than small ones, i.e., that $f(x)$ is decreasing.[19]

To construct the objective of the social planner, we first characterize the impact of punishment

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

[19] A decreasing density ensures that the social planner's second order conditions are satisfied.

on wrongdoing. We know from previous sections that, absent punishment, those who are certain of being bad give up and do wrong for sure. Threatened with a punishment N_r they would attempt to resist whenever the realized temptation is smaller than the punishment, i.e., whenever $x < N_r$, and in that case resist with probability ϕ. Therefore, given a punishment N_r, the rate of wrongdoing among the certain will be $1 - \phi F(N_r)$. That means the punishment on repeat offenders obtains a reduction in wrongdoing of exactly $\phi F(N_r)$ in the current period. As punishment applies in the current period only, and the certain learn nothing regardless of their action, N_r has no further impact on wrongdoing.

Recall that $x_t^* = x^*(\mu_t)$, and set the current period to $t = 1$. The impact of current period punishment on wrongdoing by first-time offenders is then captured by the following

Lemma 2 *A one time punishment N_f attains a reduction in the expected present value of wrongdoing by individuals who are uncertain of their type at age 1 given by:*

$$\phi\left(F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right)\left\{1 + \sum_{s=1}^{\infty}\left(\delta\lambda\phi\right)^s \frac{1}{F\left(x_1^*\right)}\prod_{j=1}^{s+1}F\left(x_j^*\right)\right\}.$$

Proof. See online Appendix. ■

The proof shows that under punishment N_f the current cutoff is $x_1^{*p} = x_1^* + N_f$ (the superscript "p" denotes a solution under the punishment regime), so current punishment raises the current optimal cutoff of the uncertain one for one. So punishment achieves a reduction in current wrongdoing equal to $\phi\left(F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right)$, and raises the share of individuals who resist and remain uncertain of their type, leading to lower wrongdoing in future periods. Specifically, of those who are saved from temptation in the current period, $\phi\lambda F\left(x_2^*\right)$ are saved again in period 2, and $\left(\phi\lambda\right)^2 F\left(x_2^*\right) F\left(x_3^*\right)$ in period 3, and so on, explaining the expression in the last lemma, where the summation captures the present and future (discounted) reductions in wrongdoing. All future cutoffs are unchanged by the one-time punishment.

Social planner's problem Using lemma (2), the social planner's objective is to choose N_r and N_f to maximize,

$$\phi F\left(N_r\right) + \phi\left[F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right]Z - c\left(N_r + N_f\right), \tag{13}$$

where

$$Z = 1 + \sum_{s=1}^{\infty}\left(\delta\lambda\phi\right)^s \prod_{j=2}^{s+1}F\left(x_j^*\right), \tag{14}$$

which only contains future cutoffs and does not involve x_1^{*p}. Given this objective, and our stated assumptions, we obtain,

Proposition 4 *If the social planner is sufficiently patient or the agents' survival rate is sufficiently low, repeat offenders are punished more harshly than first-time offenders. Formally, if δ or λ are close enough to zero, then $N_r > N_f$.*

Proof. See online Appendix. ■

An intrinsic disposition to resist temptations allows individuals to behave honestly even when there are no extrinsic incentives in place. But extrinsic incentives can obviously help keep individuals behaving honestly. Proposition 4 tells us that the design of extrinsic incentives should reflect the strength of intrinsic dispositions to avoid wrongdoing. An optimizing social 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.[20] 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– harsher punishment for repeat offenders may not make sense if the planning horizon is long, which can occur either because the planner is patient, or because agents live for a long time. In this case there is an option value to keeping the uncertain honest, in order to preserve their self-deterrence for the future. Under a long horizon, this effect may dominate and the planner would prefer to direct resources to deterring crime by first time offenders.[21]

3.3 Career choice, moral capital, and moral adverse selection

We have shown that wrongdoing rates will be higher when individuals are impatient, when their confidence of having the good type is low, and when temptations are higher. The latter feature induces both a mechanical increase in wrongdoing through higher temptations being drawn, as well as a decrease in the endogenous resistance threshold chosen by individuals. But these effects could be mitigated if individuals who are less confident about being good were to optimally select low temptation environments. But how do individuals select into careers in an economy where individual beliefs vary and different careers offer different distributions of temptations?

For concreteness, consider two occupations, "politics" and "academia." Assume that a person who enters politics faces a higher distribution of temptations, in the sense of first-order stochastic dominance. The population consists of a continuum of individuals with heterogeneous initial beliefs

[20] But harsher punishment for repeat offenders can arise also in contexts of pure extrinsic deterrence. Polinsky and Rubinfeld (1991) and Polinsky and Shavell (1998) analyze conditions under which optimal fines may be higher for repeat offenders.

[21] Our proposition involves a condition that is sufficient, but perhaps not necessary.

$\mu \in [0,1]$. We want to know how individuals self-select into different occupations depending on μ. One might imagine that individuals with a low prior μ may have an incentive to choose a low temptation activity, given that they have lower cutoffs and therefore a higher chance of giving up. Indeed, as mentioned earlier, "shelter-seeking" behavior can arise in models with time-inconsistent preferences.

Assume the economy needs workers in both careers, so compensation adjusts to ensure both careers attract a positive measure of types. The mechanism of this adjustment is immaterial for our exercise; what matters is that in equilibrium individuals who require a lower compensating differential will enter the low-temptation career. To isolate the effect of interest in the simplest way, suppose individuals live for one period. We then have

Proposition 5 *Consider an economy where individuals differ by initial self-image $\mu \in [0,1]$, and where two occupations offer different distributions of temptations, with one first-order stochastically dominant. There exists $\bar{\mu} \in (0,1)$ such that in equilibrium individuals with self-image $\mu \geq \bar{\mu}$ enter the occupation with lower temptations with the rest entering the alternative occupation.*

Proof. See online Appendix. ■

Rather than seek "shelter" in the low temptation occupation, individuals who are more vulnerable to temptation choose activities with higher temptations. This is a moral adverse selection pattern whereby the activities with temptations attract those least equipped to resist them. In the extreme case of one activity with temptations and one without, this sorting must lead to an increase in wrongdoing. This result informs the literature on political selection which is concerned with the incentives of able and honest individuals to enter a public life where corruption opportunities may abound.[22]

For individuals who know their type the selection incentives are clear: if we abstract from wages, an individual with $\mu = 1$ will obtain $u(1)$ in either occupation and be indifferent between the two careers. It follows he will prefer the low-temptation career under any positive compensating differential favoring that career. On the other end of the type spectrum, an individual with $\mu = 0$ only cares about temptations and will choose the high-temptation activity unless there is a fairly large compensating differential in favor of the low-temptation activity. In between, the result is not obvious, because of the said incentive facing the uncertain types to protect their self-image by choosing a low-temptation activity. Moreover, as shown before, for a given distribution of temptations the value function is not necessarily monotonic in μ, nor is it clear which types would place more value in a given shift in temptations. The value added of the last proposition (and the non trivial aspect of the proof) is that the compensating differential that must be paid to attract a type μ into the low temptation activity is monotonically decreasing in μ. Therefore, the population

[22] See, e.g., Caselli and Morelli (2004), Dal Bó, Dal Bó and Di Tella (2006).

can always be divided into just two segments by their beliefs μ so that types in the lower segment of self-image will enter the high-temptation profession.

Are politicians more corrupt than academics because they are inherently less moral or because they have more opportunities for corrupt behavior? In our model both arguments are correct. Even if people were divided randomly between occupations, the higher temptations would cause there to be more wrongdoing in the high-temptation sector, because of the forces highlighted by the main model: a mechanical effect and, in a dynamic setting, because cutoffs are endogenously lower when expected temptations are higher. But a third effect is present: the high temptation activity will attract the weakest types: so they will choose even lower cutoffs, and even holding fixed the temptations, fall more often.

4 Conclusion

We propose a planner-doer model of endogenous moral standards rooted in three ideas: that actions depend partly on unconscious drives subsumed in the doer, that the planner cannot easily attribute authorship of actions between himself and the doer, and that people prefer to think they have a good type of doer, i.e., that their unconscious drives are geared towards received morality. We characterize conditions under which self-restraint will emerge endogenously in the form of passing on enjoyable temptations for the sake of keeping a good introspective reputation. Our emphasis is on studying a stationary dynamic environment in order to identify conditions for persistently self-reinforcing patterns of virtue and corruption.

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

Stronger initial beliefs about having a good type, lower expected temptations and a lower discount rate increase endogenous adherence to moral standards. At the societal 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 histories involving larger temptation shocks will have a more polarized distribution of individual self-images and higher wrongdoing. Therefore, cross-country measures of wrongdoing and cultures of corruption may not reflect differences in deep moral fundamentals but simply different shock histories.

A valid critique of our basic model, and of other models in the literature where changing higher

moments of a distribution is valuable, is that these models are not fully identified empirically. The general point is made in a perceptive paper by Eliaz and Spiegler (2006), and the implication for our setup is one could obtain a positive function $x^*(\mu)$ if postulating (in a suitable manner) that the planner has a self-esteem that decreases in μ. This does not render these theories vacuous–they may be rooted in the behaviorally correct notions but fail to make predictions that are distinct enough. To be sure, more needs to be done to further validate these theories, for example by examining auxiliary predictions empirically. An additional critique is that the model focuses exclusively in ex ante incentives to manipulate information under the assumption that the individual is fully Bayesian afterwards, when in reality individuals may engage in actions to protect their self-concept ex post, for example by not updating in a Bayesian manner after doing something immoral.

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 deemed inappropriate for people with that identity. Our model suggests that "identity-based costs" may not be constant, but respond to past actions and to the person's beliefs that such identity (e.g., that of a good person) is still hers. The model can also rationalize why high temptation activities may attract the individuals least equipped to resist–a pattern we call moral adverse selection–thus magnifying wrongdoing differentials across activities, and a rationale for punishing repeat-offenders more harshly. This application illustrates that the optimal design of deterrence schemes may change when the disposition toward wrongdoing is endogenized.

Appendix

The (non-recursive) infinite horizon formulation Consider an individual in an arbitrary period, say $t = 1$, with prior belief μ_1. In the future, conditional on remaining uncertain, his belief will increase deterministically after repeated application of the update b, so that in period t self-esteem utility is based on the updated belief $\mu_{t+1} = \mu_1 / \left(\mu_1 + (1-\mu_1)\phi^t\right)$. The associated period utility consists only of self-esteem and also increases deterministically; denote these utilities $u_t = u(\mu_{t+1})$. A certain agent has a trivial problem with a known terminal expected present value $V_0 = (u(0) + Ex)/(1-\lambda)$ or $V_1 = u(1)/(1-\lambda)$. The problem of the individual is to select a policy threshold for all t, conditional on still remaining uncertain, in order to maximize the present value expected lifetime utility (or "value" for short). Denote these cutoffs $\hat{x}_1, \hat{x}_{2,...}$and their optimal values as $x_1^*, x_2^*, \ldots$ Denote the probability of receiving temptations below the cutoff for the first t periods by

$$H_t(\hat{\mathbf{x}}_t) = \prod_{s=1}^{t} F(\hat{x}_s) \quad \text{for } t > 1, \tag{A.1}$$

where $\hat{\mathbf{x}}_t \equiv (\hat{x}_1, \ldots, \hat{x}_t)$, and define $H_0 \equiv 1$ for convenience. For a good type H_t is also the probability of remaining uncertain after t periods. Conditional on bad type, the payoff relevant

states at the end of period t are i) being uncertain of type, ii) having been certain already in $t-1$, and iii) becoming certain of type in period t. The probabilities are

$$\begin{array}{ll} \text{(i)} & \phi^t H_t\left(\hat{\mathbf{x}}_t\right), \\ \text{(ii)} & 1-\phi^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right), \\ \text{(iii)} & \phi^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right)\left[1-\phi F\left(\hat{x}_t\right)\right]. \end{array} \tag{A.2}$$

Value can be decomposed into the contributions conditional on being bad and being good over all periods. For a good type, the contribution of any future period t into present value consists of a reduction below the maximum possible self-esteem utility, which occurs with the probability of still being uncertain in period t:

$$V_{\text{good}}\left(\hat{\mathbf{x}}_\infty\right) = \sum_{t=1}^{\infty} \lambda^{t-1}\left(H_t\left(\hat{\mathbf{x}}_t\right) u_t + \left(1-H_t\left(\hat{\mathbf{x}}_t\right)\right) u\left(1\right)\right) \tag{A.3}$$

$$= V_1 - \sum_{t=1}^{\infty} \lambda^{t-1} H_t\left(\hat{\mathbf{x}}_t\right)\left(u\left(1\right) - u_t\right) \tag{A.4}$$

For a bad type, the contribution from being certain at t is different depending on whether period t is the period of first falling for a temptation or not, because at the first falling the cutoff policy affects the expected value of the temptation consumed:

$$V_{\text{bad}}\left(\hat{\mathbf{x}}_\infty\right) = \sum_{t=1}^{\infty} \lambda^{t-1} \left(\begin{array}{c} \phi^t H_t\left(\hat{\mathbf{x}}_t\right) u_t + \\ \left(1-\phi^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right)\right)\left[Ex + u\left(0\right)\right] + \\ \phi^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right)\left[1-\phi F\left(\hat{x}_t\right)\right]\left[z\left(\hat{x}_t\right) + u\left(0\right)\right] \end{array} \right) \tag{A.5}$$

$$= V_0 + \sum_{t=1}^{\infty} \left(\lambda\phi\right)^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right)\left(\phi F\left(\hat{x}_t\right)\left[u_t - u\left(0\right) - z\left(\hat{x}_t\right)\right] + z\left(\hat{x}_t\right) - Ex\right),$$

$$\text{where } z\left(x_t^*\right) = \frac{\left(1-\phi\right) F\left(\hat{x}_t\right) E[x|x \le x_t^*] + \left(1-F\left(\hat{x}_t\right)\right) E[x|x > \hat{x}_t]}{1-\phi F\left(x_t^*\right)} \tag{A.6}$$

$$= \frac{\left(1-\phi\right) Ex + \phi \int_{\hat{x}_t}^{\infty} x f\left(x\right) \mathrm{d}x}{1-\phi F\left(\hat{x}_t\right)} \tag{A.7}$$

is the expected value of temptation in period t conditional on using decision threshold x_t^* and on t being the period of becoming certain of bad type. Given policy $\mathbf{x}_\infty^*$, the value is

$$V\left(\hat{\mathbf{x}}_\infty\right) = \mu_1 V_{\text{good}}\left(\hat{\mathbf{x}}_\infty\right) + \left(1-\mu_1\right) V_{\text{bad}}\left(\hat{\mathbf{x}}_\infty\right). \tag{A.8}$$

Using (A.3) and (A.5), value can be expressed as

$$V\left(\hat{\mathbf{x}}_\infty\right) = \mu_1 V_1 + \left(1-\mu_1\right) V_0 \tag{A.9}$$

$$+ \sum_{t=1}^{\infty} \lambda^{t-1} H_t\left(\hat{\mathbf{x}}_t\right)\left(\left[\mu_1 + \left(1-\mu_1\right)\phi^t\right] u_t - \mu_1 u\left(1\right)\right) \tag{A.10}$$

$$+ \left(1-\mu_1\right) \sum_{t=1}^{\infty} \left(\lambda\phi\right)^{t-1} H_{t-1}\left(\hat{\mathbf{x}}_{t-1}\right)\left(\left[1-\phi F\left(\hat{x}_t\right)\right] z\left(\hat{x}_t\right) - \phi F\left(\hat{x}_t\right) u\left(0\right) - Ex\right).$$

Before setting up the first order condition, it is helpful to notice that $\frac{\partial}{\partial \hat{x}_1} H_t(\hat{\mathbf{x}}_t) = \frac{f(\hat{x}_1)}{F(\hat{x}_1)} H_t(\hat{\mathbf{x}}_t)$ for $t \geq 1$, and $\frac{\partial}{\partial \hat{x}_1}[1 - \phi F(\hat{x}_1)] z(\hat{x}_1) = -\phi \hat{x}_1 f(\hat{x}_1)$. Finally, the first-order condition with respect to $\hat{x}_1$ is,

$$\begin{aligned}\frac{\partial}{\partial \hat{x}_1} V(\hat{\mathbf{x}}_\infty) &= \frac{f(\hat{x}_1)}{F(\hat{x}_1)} \sum_{t=1}^{\infty} \lambda^{t-1} H_t(\hat{\mathbf{x}}_t) \left(\left[\mu_1 + (1-\mu_1)\phi^t \right] u_t - \mu_1 u(1) \right) \\ &+ \frac{f(\hat{x}_1)}{F(\hat{x}_1)} (1-\mu_1) \sum_{t=2}^{\infty} (\lambda\phi)^{t-1} H_{t-1}(\hat{\mathbf{x}}_{t-1}) \left([1 - \phi F(\hat{x}_t)] z(\hat{x}_t) - \phi F(\hat{x}_t) u(0) - Ex \right) \\ &- \phi f(\hat{x}_1)(1-\mu_1)(\hat{x}_1 + u(0)) = 0.\end{aligned} \tag{A.11}$$

Denote $H_{t,-1}(\hat{\mathbf{x}}_{t,-1}) = \prod_{s=2}^{t} F(\hat{x}_s)$, where $\hat{\mathbf{x}}_{t,-1} = (\hat{x}_2, \ldots, \hat{x}_t)$ and $t \geq 2$, and define $H_{1,-1} \equiv 1$. After dividing by $f(\hat{x}_1)$ the first-order condition can then be expressed as

$$\begin{aligned}\frac{\partial}{\partial \hat{x}_1} V(\hat{\mathbf{x}}_\infty) &= \sum_{t=1}^{\infty} \lambda^{t-1} H_{t,-1}(\hat{\mathbf{x}}_{t,-1}) \left(\left[\mu_1 + (1-\mu_1)\phi^t \right] u_t - \mu_1 u(1) \right) \\ &+ (1-\mu_1) \sum_{t=2}^{\infty} (\lambda\phi)^{t-1} H_{t-1,-1}(\hat{\mathbf{x}}_{t-1,-1}) \left([1 - \phi F(\hat{x}_t)] z(\hat{x}_t) - \phi F(\hat{x}_t) u(0) - Ex \right) \\ &- \phi(1-\mu_1)(\hat{x}_1 + u(0)) = 0.\end{aligned} \tag{A.12}$$

The optimal cutoff in period 1 must satisfy

$$\begin{aligned}x_1^* &= \frac{1}{\phi(1-\mu_1)} \sum_{t=1}^{\infty} \lambda^{t-1} H_{t,-1}(\mathbf{x}^*_{t,-1}) \left(\left[\mu_1 + (1-\mu_1)\phi^t \right] u_t - \mu_1 u(1) \right) \\ &+ \frac{1}{\phi} \sum_{t=2}^{\infty} (\lambda\phi)^{t-1} H_{t-1,-1}(\mathbf{x}^*_{t-1,-1}) \left([1 - \phi F(x_t^*)] z(x_t^*) - \phi F(x_t^*) u(0) - Ex \right) - u(0).\end{aligned} \tag{A.13}$$

Proof of proposition 2: Policy monotonicity *Monotonicity:* Notice that optimal cutoffs are independent of past cutoffs, because how beliefs μ_t were arrived at is payoff-irrelevant for the future. The problem is stationary, in the sense that if the individual remains uncertain after a future period t, then one just moves forward all time indices so that μ_1 is replaced by μ_t, and so on. Thus, we need to show only that x_1^* is increasing in μ_1. Two steps are important before tackling the third, and final step. First, recognizing that all effects of μ_1 through $\mathbf{x}^*_{t,-1}$ can be ignored thanks to the Envelope theorem. (See online appendix for more detailed intermediate steps.) Second, note that only the first summation in (A.13) depends on μ_1; therefore, we need only be concerned with partial differentiation of the first line in (A.13). The third and final step is to obtain this partial

derivative and prove it is positive. Differentiating the first line (A.13) we get,

$$\frac{dx_1^*}{d\mu_1} = \frac{1}{\phi(1-\mu_1)^2}\sum_{t=1}^{\infty}\lambda^{t-1}H_{t,-1}\left(\mathbf{x}_{t,-1}^*\right)\left(\left[\mu_1+(1-\mu_1)\phi^t\right]u_t-\mu_1 u(1)\right) \tag{A.14}$$

$$+\frac{1}{\phi(1-\mu_1)}\sum_{t=1}^{\infty}\lambda^{t-1}H_{t,-1}\left(\mathbf{x}_{t,-1}^*\right)\left[\left(1-\phi^t\right)u_t-u(1)+\left[\mu_1+(1-\mu_1)\phi^t\right]\frac{du_t}{d\mu_1}\right]$$

$$= \frac{1}{\phi(1-\mu_1)^2}\sum_{t=1}^{\infty}\lambda^{t-1}H_{t,-1}\left(\mathbf{x}_{t,-1}^*\right)\left[u_t-u(1)+(1-\mu_1)\left[\mu_1+(1-\mu_1)\phi^t\right]\frac{du_t}{d\mu_1}\right] \tag{A.15}$$

It is then sufficient to show the positivity of the bracketed term inside the summation for every $t \geq 1$. Start by noting that

$$\frac{du_t}{d\mu_1} = u'\left(\mu_{t+1}\right)\frac{d\mu_{t+1}}{d\mu_1} = \phi^t\left(\frac{\mu_{t+1}}{\mu_1}\right)^2\mu_{t+1}^{-\rho}, \tag{A.16}$$

$$\text{where} \qquad \frac{d\mu_{t+1}}{d\mu_1} = \phi^t\left(\frac{\mu_{t+1}}{\mu_1}\right)^2 \text{ was used.} \tag{A.17}$$

Applying (11) and (A.16) inside the bracketed term of (A.15), we see that we need to show that

$$\frac{\mu_{t+1}^{1-\rho}-1}{1-\rho}+(1-\mu_1)\left[\mu_1+(1-\mu_1)\phi^t\right]\phi^t\left(\frac{\mu_{t+1}}{\mu_1}\right)^2\mu_{t+1}^{-\rho} \geq 0. \tag{A.18}$$

After some rearranging and simplifying, this becomes

$$\mu_{t+1}^{-\rho}-1-\rho\left(\frac{1-\mu_1}{\mu_1}\right)\phi^t\mu_{t+1}^{1-\rho} \geq 0. \tag{A.19}$$

At $\rho = 0$ this holds as an equality for all $\mu_1 \in (0,1)$. For $\mu_1 = 1$ this holds as an equality for all $\rho \in [0,1)$. It remains to show that the left side of (A.19) is decreasing in μ_1 at all $\rho \in (0,1)$. Differentiating it with respect to μ_1 gives

$$-\rho\mu_{t+1}^{-\rho-1}\frac{d\mu_{t+1}}{d\mu_1}-\rho(1-\rho)\left(\frac{1-\mu_1}{\mu_1}\right)\phi^t\mu_{t+1}^{-\rho}\frac{d\mu_{t+1}}{d\mu_1}+\frac{\rho}{\mu_1^2}\phi^t\mu_{t+1}^{1-\rho}. \tag{A.20}$$

Using (A.17), this becomes

$$-\rho\frac{\mu_{t+1}^{1-\rho}}{\mu_1^2}\phi^t-\rho(1-\rho)\left(\frac{1-\mu_1}{\mu_1^3}\right)\phi^{2t}\mu_{t+1}^{2-\rho}+\frac{\rho}{\mu_1^2}\phi^t\mu_{t+1}^{1-\rho} \tag{A.21}$$

$$= -\rho(1-\rho)\left(\frac{1-\mu_1}{\mu_1^3}\right)\phi^{2t}\mu_{t+1}^{2-\rho} \tag{A.22}$$

which is indeed negative for all $\mu_1 \in (0,1)$, $\rho \in (0,1)$, $t \geq 0$.

To understand what drives the result, let's look back at the third step in more detail. Denote the first line of the optimal cutoff (A.13) as $\sum_{t=1}^{\infty}\lambda^{t-1}H_{t,-1}\left(\mathbf{x}_{t,-1}^*\right)g_t$, where $g_t = \frac{\left[\mu_1+(1-\mu_1)\phi^t\right]u_t-\mu_1 u(1)}{\phi(1-\mu_1)}$. Whether x_1^* is increasing in μ_1 depends (by virtue of the envelope theorem) on whether the g_t terms are increasing in μ_1. They capture the same trade-off as described in (10), only now in direct as

opposed to recursive form. Still, as in (10), present and future terms are isomorphic, reflecting the benefits of reduced risk over future beliefs (the numerator) and the probability of override leading to forgoing a temptation (the denominator). Thus, for g_t to be increasing in μ_1, the denominator must decrease faster in μ_1 than the numerator, for all $\mu_1 \in [0,1]$. An indication that this should happen is that, abstracting from utility considerations, the pure reduction in variance over beliefs attained by override decreases in μ_1 at a rate easily computed to be $\frac{\phi\mu_1(1-\mu_1)}{\mu_1+(1-\mu_1)\phi}$, which is smaller than the rate ϕ at which the denominator decreases. Thus, the benefit/cost ratio governing the decision to override is everywhere increasing in μ_1.

Finite limit of policy function It is sufficient to show that, for every $\mu < 1$, a sufficiently large temptation makes it optimal to give up. The continuation value of an uncertain individual is bounded above by the present value of getting the highest possible expected period utility forever, which is finite since $\lambda < 1$, $u(1) < 1$, and $Ex < \infty$. Therefore $E_x V\left(b\left(\mu_t\right), x_{t+1}\right)$, with $V\left(.\right)$ given by (8), is bounded above. By contrast, the current period temptation on offer, x_t, enters linearly in $U(0, x_t, \mu_t)$. Thus, a sufficiently high temptation at hand in the current period will make giving up ($a_t = 0$) optimal.

Proof of Proposition 3: comparative statics To show that $x^*(\mu)$ is lower at every $\mu \in (0,1]$ when the distribution of temptations is higher, consider a small increase in the distribution, such that the new distribution is $F_\varepsilon\left(x\right) = F\left(x - \varepsilon\right)$ for some $\varepsilon > 0$. Higher ε implies a higher distribution in the sense of first-order stochastic dominance. (This implies that there is then zero probability of $x < \varepsilon$). Before proceeding with differentiation, let's rearrange (A.13) for convenience. Using equivalence

$$\left[1 - \phi F\left(\hat{x}_t\right)\right] z\left(\hat{x}_t\right) - \phi F\left(\hat{x}_t\right) u\left(0\right) - Ex = -\phi F\left(\hat{x}_t\right) E[x + u\left(0\right) | x < \hat{x}_t] \tag{A.23}$$

and rearranging we can write

$$\begin{aligned} x_1^* &= \frac{\left[\mu_1 + \left(1-\mu_1\right)\phi\right] u_1 - \left[\mu_1 u\left(1\right) + \phi\left(1-\mu_1\right) u(0)\right]}{\phi\left(1-\mu_1\right)} \\ &\quad + \frac{1}{\phi\left(1-\mu_1\right)} \sum_{t=2}^{\infty} \lambda^{t-1} H_{t,-1}\left(\mathbf{x}_{t,-1}^*\right) \left[\begin{array}{c} \left[\mu_1 + \left(1-\mu_1\right)\phi^t\right] u_t - \mu_1 u\left(1\right) \\ -\left(1-\mu_1\right)\phi^t E[x + u\left(0\right) | x < x_t^*] \end{array} \right]. \end{aligned} \tag{A.24}$$

Only the second line of (A.24) depends on F, so the optimal policy, with distribution F_ε, is

$$x_1^* = \text{constant} \tag{A.25}$$

$$+ \frac{1}{\phi\left(1-\mu_1\right)} \sum_{t=2}^{\infty} \lambda^{t-1} \prod_{s=2}^{t} F\left(x_s^* - \varepsilon\right) \left[\begin{array}{c} \left[\mu_1 + \left(1-\mu_1\right)\phi^t\right] u_t - \mu_1 u\left(1\right) \\ -\left(1-\mu_1\right)\phi^t E[x + \varepsilon + u\left(0\right) | x < x_t^*] \end{array} \right] \tag{A.26}$$

Now consider the derivative of x_1^* with respect to ε. The bracketed term, when evaluated at $\varepsilon = 0$, is the expected period t utility gain in trying to resist temptations below x, it is positive by the

optimality of x_t^*. Thanks to the envelope theorem, the effects through $x_2^*, x_3^*, \ldots$ are zero. By inspection, all partial derivatives involving $F(x_s^* - \varepsilon)$ and $-(1-\mu_1)\phi^t E[x + \varepsilon + u(0) | x < x_t^*]$ are negative. Hence x_1^* is decreasing in ε.

The result for λ is clear by inspection of (A.13); all partial derivatives are either positive or vanish by the envelope theorem.

References

Akerlof, George and Rachel Kranton (2000), Economics and Identity, Quarterly Journal of Economics 115 (August), 715-53.

Ali, S. Nageeb (2011), Learning Self-Control, Quarterly Journal of Economics 126(2), 857-893.

Aristotle (1998), Nichomachean Ethics, Dover.

Bargh, John and Tanya Chartrand (1999), The Unbearable Automaticity of Being, American Psychologist 54(7), 462-79.

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

Becker, Gary and Kevin M. Murphy (1988), A Theory of Rational Addiction, Journal of Political Economy 96(4), 675-700.

Bénabou, Roland and Marek Pycia (2002), Dynamic Inconsistency and Self-Control: A Planner-Doer Interpretation, Economics Letters 77(3), 419-424.

Bénabou, Roland and Jean Tirole (2004), Willpower and Personal Rules, Journal of Political Economy 112, 848-886.

Bénabou, Roland and Jean Tirole (2006), Incentives and Prosocial Behavior, American Economic Review 96(5), 1652-1678.

Bénabou, Roland and Jean Tirole (2011), Identity, Morals and Taboos: Beliefs as Assets, Quarterly Journal of Economics 126, 805-855.

Bernheim, B. Douglas and Antonio Rangel (2004), Addiction and Cue-Triggered Decision Processes, American Economic Review 94(5), 1558-1590.

Berridge, Kent (2003), Irrational Pursuits: Hyper-Incentives From a Visceral Brain, in Brocas, I., and J. Carrillo (eds.) The psychology of economic decisions. Oxford University Press.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg (2003), An Economic Model of Moral Motivation, Journal of Public Economics 87, 1967-1983.

Camerer Colin, George Loewenstein, and Drazen Prelec (2005), Neuroeconomics: How Neuroscience Can Inform Economics, Journal of Economic Literature 43(1), 9-64.

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

Caselli, Francesco and Massimo Morelli (2004), Bad Politicians, Journal of Public Economics 88(3-4), 759-782.

Cervellati, Matteo, Joan Esteban and Laurence Kranich (2006), The Social Contract With Endogenous Sentiments, mimeo Institut d'Anàlisi Econòmica.

Clark, John, James Austin and Alan 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, Olivier and Andrew Postlewaite (2004), Confidence-Enhanced Performance, American Economic Review 94(5), 1536-1557.

Eliaz, Kfir and Ran Spiegler (2006), Can Anticipatory Feelings Explain Anomalous Choices of Information Sources?, Games and Economic Behavior 56(1), 87-104.

Dal Bó, Ernesto, Pedro Dal Bó, and Rafael Di Tella (2006), Plata o Plomo?: Bribe and Punishment in a Theory of Political Influence, American Political Science Review 100(1), 41-53.

Fischbacher, Urs and Franziska Heusi (2008), Lies in Disguise: An Experimental Study on Cheating, University of Konstanz Research Paper #40.

Fisman, Raymond and Edward Miguel (2007), Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets, Journal of Political Economy 115(6), 1020-1048.

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

Gino, Francesca. and Lamar Pierce (2008), The Abundance Effect: Unethical Behavior in the Presence of Wealth, Organizational Behavior and Human Decision Processes 109(2), 142-155.

Gottfredson, Michael and Travis Hirschi (1990), A General Theory of Crime, Stanford Univ Press.

Grace, Anthony, Stan Floresco, Yau Goto, and Daniel Lodge (2007), Regulation of firing of dopaminergic neurons and Control of Goal-Directed Behaviors, TRENDS in Neuroscience 30(5), 220-227.

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

Kamenica, Emir and Matthew Gentzkow (2011), Bayesian Persuasion, American Economic Review 101(6), 2590-2615.

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

Kreps, David and Robert Wilson (1982), Reputation and Imperfect Information, Journal of Economic Theory, 27, 253-79.

Miller, Earl and Jonathan Cohen (2001), An Integrative Theory of Prefrontal Cortex Function, Annual Review of Neuroscience 24, 167-202.

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

Nisbett Richard and Timothy Wilson (1977), Telling More Than We Can Know: Verbal Reports of Mental Processes, Psychological Review 84(3), 231-259.

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

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

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

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

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

Shefrin, H. M. and Richard H. Thaler (1981). "An Economic Theory of Self-Control." Journal of Political Economy, 89(2), 392–406.

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

Vallacher, Robin and Daniel Wegner (1987), What Do People Think They Are Doing?: Action Identification and Human Behavior, Psychological Review 94(1), 3-15.

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

Wegner, Daniel and Thalia Wheatley (1999), Apparent mental causation: Sources of the experience of will, American Psychologist 54(7), 480-492.

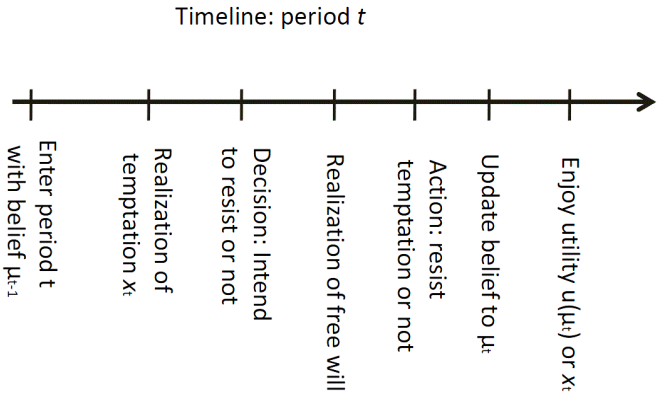

Figure 1. At the start of period t planner has prior μ_{t-1} and observes temptation x_t. He then decides whether to attempt override ($a_t = 1$), or not ($a_t = 0$). The externally observable action r_t is then determined. Under $a_t = 0$, the doer alone determines r_t, so $r_t = 1$ if $\theta = \theta_g$ and $r_t = 0$ if $\theta = \theta_b$. Under $a_t = 1$, $r_t = 1$ for sure if $\theta = \theta_g$ and with probability ϕ if $\theta = \theta_b$. At the end of period t, knowing a_t and r_t, the planner updates belief to μ_t.

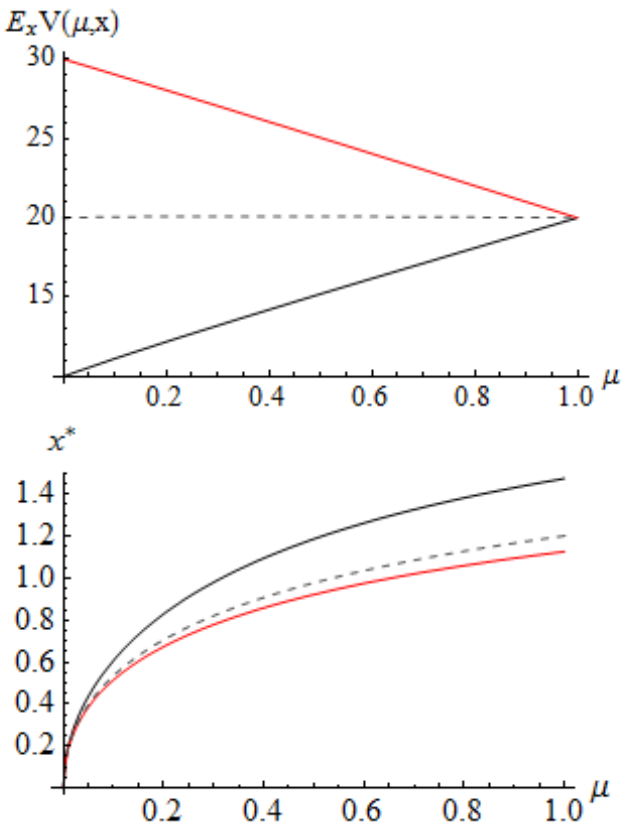

Figure 2. Expected value function $V(\mu) \equiv E_x V(x, \mu)$ and optimal policy $x^*(\mu)$ function under three different cases. The case with a benign environment, where full self-esteem offers significantly more utility than the average temptation, results when $u(1) - u(0) >> Ex$; it is depicted in black. The opposite case of a harsh environment is depicted in red, and the balanced case is depicted with dashed curves. (The relatively flat value function of the balanced case is also strictly concave, with a maximum at $\mu = 0.294$.) Parameters: $\rho = 0.5$, $\phi = 0.75$, $\lambda = 0.9$, and x is distributed exponentially with $Ex = 1$, 3, and 2 respectively.

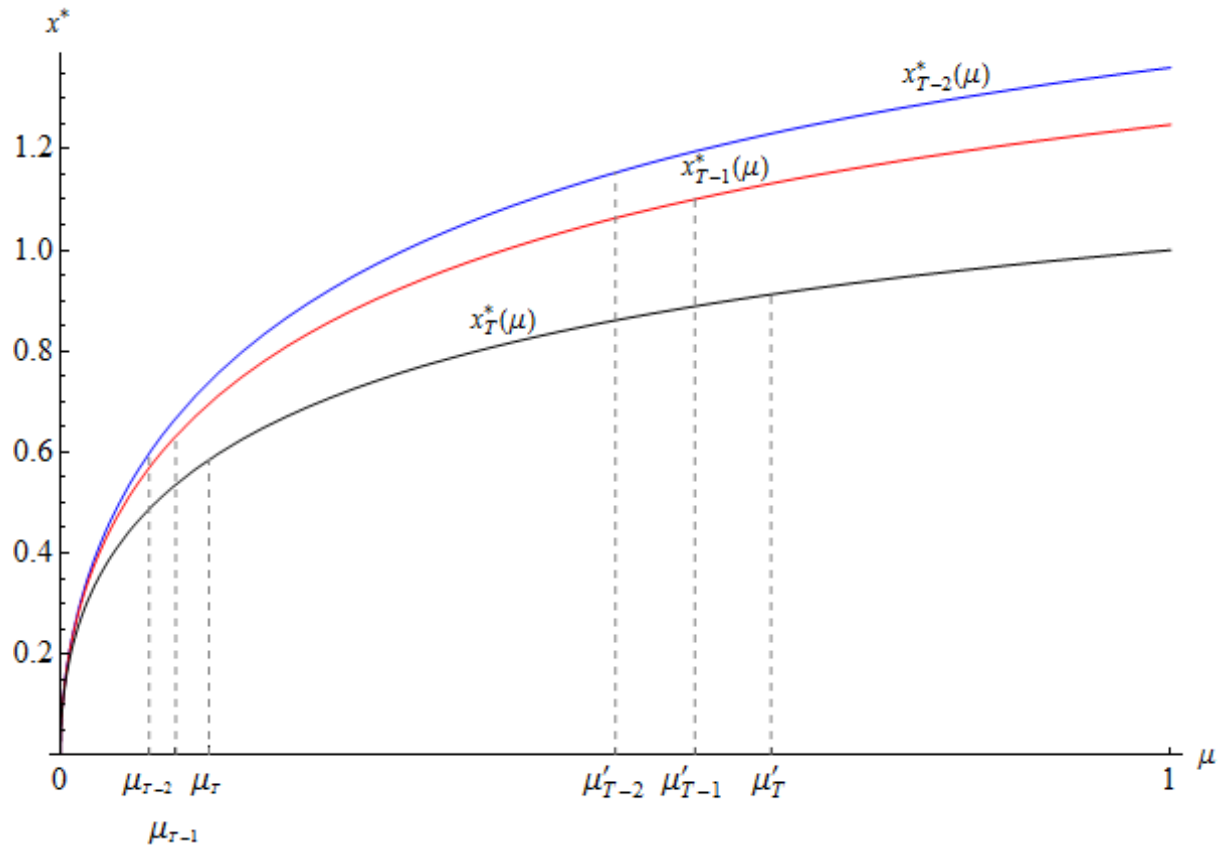

Figure 3. Optimal policy and histories of beliefs for the last three periods of a finite horizon (*T-2*, *T-1*, and T), conditional on successful override. Same parameters as in Figure 2,with $Ex = 3$. For the individual with low initial beliefs μ_{T-2}, moral standards increase in $T-1$ (due to "Aristotelian" moral growth) then decrease in the last period due to the finite horizon effect. For high initial beliefs μ'_{T-2}, the finite horizon effect dominates and moral standards decrease over time.

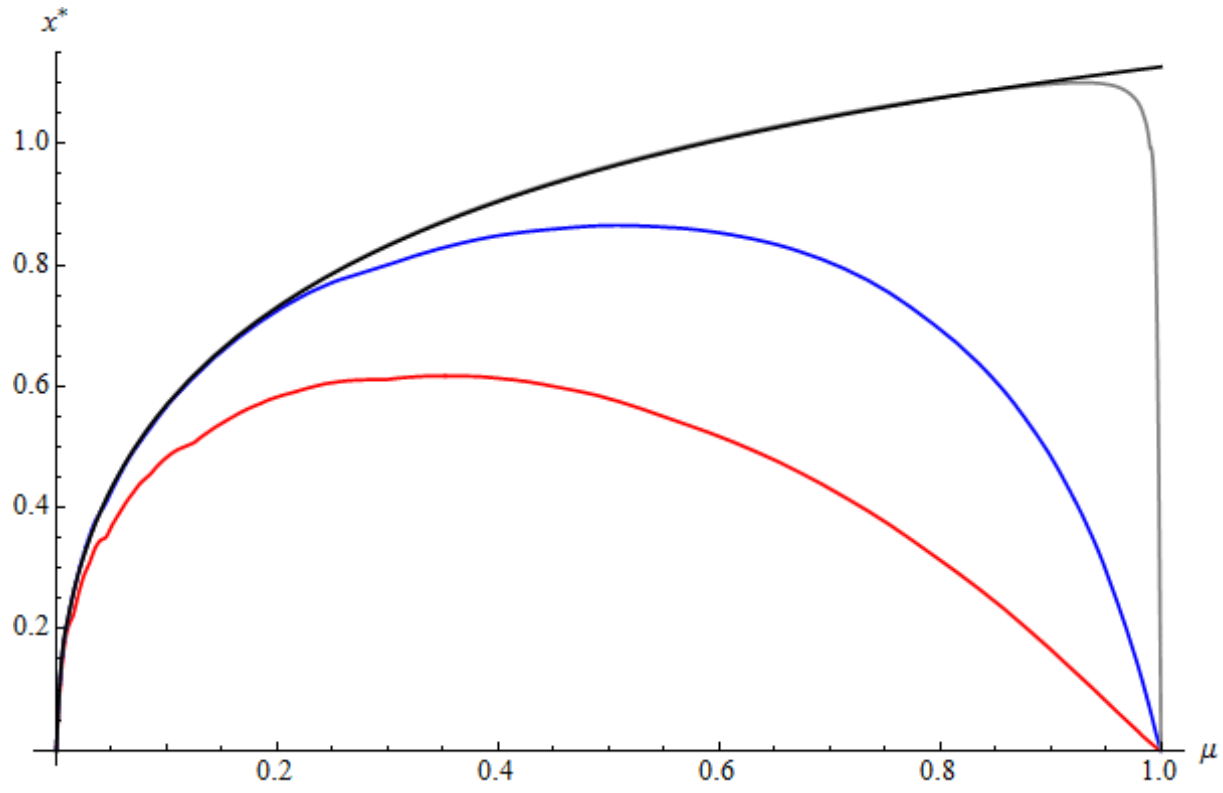

Figure 4. Optimal policy for fallible types, under different values of symmetric error rate, $\varepsilon = 1 - \gamma_g = \gamma_b$. The lowest of the curves (depicted in red) corresponds to the highest error rate $\varepsilon = 0.25$. The following two curves in have $\varepsilon = 0.1$ (blue), and $\varepsilon = 0.001$ (gray). The highest values of x^* obtain in the infallible case $\varepsilon = 0$ (black), which is the basic model. Other parameters are as in Figure 2, with $Ex = 3$.

Online Appendix for Self-Esteem, Moral Capital, and Wrongdoing

Proving the existence of the value function We assume, for technical convenience, that x has a finite but arbitrarily high upper bound, $\bar{X}$.[23] Let us define the mapping T as,

$$\begin{aligned} T\varphi(\mu,x) &= \mu V_1 + (1-\mu)V_0 + (1-\mu)(x-Ex) + \\ &\quad + \max\left\{ \begin{pmatrix} (\mu+\phi(1-\mu))\left[u(b(\mu)) + \lambda E\varphi(b(\mu),x')\right] - \mu V_1 - \\ -\phi(1-\mu)(V_0+x-Ex) \end{pmatrix}, \right\}, \quad \text{(B.1)} \\ &\qquad \left\{ 0 \right\} \end{aligned}$$

where the right hand side follows–after simple algebra–from that in (9).

Recall that u and b are continuous increasing functions, ϕ and λ are parameters in $(0,1)$, and V_0, V_1 and Ex are known constants. It is clear by inspection that T preserves continuity and boundedness in μ and x. Thus, T maps the space $\mathcal{C}$ of functions that are continuous and bounded in (μ,x), to itself. Endowing the space $\mathcal{C}$ with the sup norm yields a complete metric space. We are interested in a unique fixed point for T. Blackwell's sufficient conditions for T to be a contraction mapping–and for it to have a unique fixed point–are,

$$\text{(Discounting)} \quad T(\varphi+k) \leq T\varphi + k\beta \text{ for some } \beta\in[0,1) \text{ and any } k>0 \quad \text{(B.2)}$$

$$\text{(Monotonicity)} \quad \varphi \leq \vartheta \Longrightarrow T\varphi \leq T\vartheta, \quad \text{(B.3)}$$

for any φ, ϑ in $\mathcal{C}$.

It is immediate that for the purpose of showing satisfaction of the Blackwell conditions it is sufficient to restrict attention to following map M,

$$M\varphi(\mu,x) = \max\left\{ \begin{matrix} ((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x), \\ 0 \end{matrix} \right\}. \quad \text{(B.4)}$$

We start with the first discounting condition: $M(\varphi+k) \leq M\varphi + k\beta$ for some $\beta\in[0,1)$. Note $M(\varphi+k)(\mu,x) \leq M\varphi(\mu,x) + \beta k$ implies,

$$\begin{aligned} &(\mu+\phi(1-\mu))\lambda k + \max\left\{ \begin{matrix} ((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x), \\ -(\mu+\phi(1-\mu))\lambda k \end{matrix} \right\} \leq \\ &\qquad \leq \max\left\{ \begin{matrix} ((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x), \\ 0 \end{matrix} \right\} + \beta k. \end{aligned}$$

[23] We do not think that allowing for unbounded temptations is important for the substantive message of the model. However, it is convenient to use unbounded distributions for numerical examples, and indeed we use exponentially distributed x for the cases depicted in our graphs. The proof of existence can be extended to the case with unbounded temptations by relying on a weighted contraction mapping theorem.

Now take $\beta = (\mu + \phi(1-\mu))\lambda < 1$, and the last inequality becomes

$$\max\left\{\begin{array}{c}((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x),\\ -(\mu+\phi(1-\mu))\lambda k\end{array}\right\} \leq$$
$$\leq \max\left\{\begin{array}{c}((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x),\\ 0\end{array}\right\},$$

which true from the max function being increasing in both its arguments and $(\mu + \phi(1-\mu))\lambda k > 0$.

We finish with the monotonicity condition $\varphi \leq \vartheta \Longrightarrow M\varphi \leq M\vartheta$. Using the expression for M, the inequality $M\varphi \leq M\vartheta$ is,

$$\max\left\{\begin{array}{c}((\mu+\phi(1-\mu))\lambda E\varphi(b(\mu),x') - \phi(1-\mu)x),\\ 0\end{array}\right\} \leq$$
$$\leq \max\left\{\begin{array}{c}((\mu+\phi(1-\mu))\lambda E\vartheta(b(\mu),x') - \phi(1-\mu)x),\\ 0\end{array}\right\},$$

which is obviously true whenever $\vartheta \geq \varphi$ by virtue of $(\mu + \phi(1-\mu))\lambda > 0$ and the max operator being increasing in its two arguments.

Additional detail on applying the Envelope Theorem in the Proof of Proposition 2. This note explains in detail why the envelope theorem allows us to ignore the terms involving $\partial x_t^*/\partial\mu$ (for $t > 1$) in the comparative statics analysis of x_1^*.

The expected present value of utility $V(\mathbf{x}_\infty^*)$ can be decomposed into two additive components, one contributed by the current period and those future states of the world where the type is revealed during the current period, denoted S below; and second, expected present value contributed by future periods in the event that the type is not revealed in the current period. This is because current period policy $\hat{x}_1$ only affects the affects the probability P at which the doer's type is not revealed in the current period, but it is not directly payoff relevant once the future arrives, and so does not affect remaining present value next period in the event that the planner is still uncertain, $\tilde{V}$. Similarly, if the type is revealed then the decision problem is over, hence future decisions $\hat{x}_t$ do not enter S. Thus the present value of utility can be written in the following form.

$$V(\hat{\mathbf{x}}_\infty) = S(\hat{x}_1) + P(\hat{x}_1)\lambda\tilde{V}(\hat{\mathbf{x}}_{2,\ldots,\infty}) \tag{B.5}$$

Each of $S, P, \tilde{V}$ depend on μ_1 (the latter through Bayesian updates $\mu_2, \mu_3, \ldots$). The optimal value of x_1^* must satisfy the first order condition

$$\frac{\partial}{\partial\hat{x}_1}V(\mathbf{x}_\infty^*) = \frac{\partial}{\partial\hat{x}_1}S(x_1^*) + \left[\frac{\partial}{\partial\hat{x}_1}P(x_1^*)\right]\lambda\tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*) = 0. \tag{B.6}$$

To study the comparative statics of x_1^* with respect to μ, first differentiate the first-order condition with respect to to x_1^* and μ get

$$\begin{aligned}
&\frac{\partial^2}{\partial^2 x_1^*} S\left(x_1^*\right) \mathrm{d}x_1^* + \left[\frac{\partial^2}{\partial^2 x_1^*} P\left(x_1^*\right)\right] \mathrm{d}x_1^* \lambda \tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*) \\
&+ \frac{\partial^2}{\partial x_1^* \partial \mu} S\left(x_1^*\right) \mathrm{d}\mu + \left[\frac{\partial^2}{\partial x_1^* \partial \mu} P\left(x_1^*\right)\right] \mathrm{d}\mu \left[\lambda \tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*)\right] \\
+ \quad &\left[\frac{\partial}{\partial x_1^*} P\left(x_1^*\right)\right] \lambda \left[\sum_{t=2}^{\infty} \frac{\partial}{\partial x_t^*} \tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*) \frac{\partial x_t^*}{\partial \mu} + \frac{\partial}{\partial \mu} \tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*)\right] \mathrm{d}\mu = 0. \qquad \text{(B.7)}
\end{aligned}$$

Since future choices will again maximize present value in those states of the world, the "future FOC" terms $\frac{\partial}{\partial x_t^*} \tilde{V}(\mathbf{x}_{2,\ldots,\infty}^*)$ are zero. This means that all partial derivatives of future cutoffs $\partial x_t^* / \partial \mu$ vanish from this equation, because they only appear multiplied by the future FOCs. Finally, the comparative static result is obtained by solving for $\mathrm{d}x_1^*/\mathrm{d}\mu$, which is done explicitly in the proof. Knowing that all the terms multiplied by $\partial x_t^*/\partial \mu$ (for $t > 1$) can be set to zero after differentiating (A.13) with respect to μ simplifies the proof considerably.

Proof of Lemma 2. The uncertain person facing punishment N_f in the current period t faces the problem (using the recursive formulation in (9)),

$$V(\mu, x) = \max \left\{ \begin{array}{c} [\mu + \phi(1-\mu)] \left[u\left(\frac{\mu}{\mu+(1-\mu)\phi}\right) + \lambda EV\left(\frac{\mu}{\mu+(1-\mu)\phi}, x'\right)\right] \\ + (1-\mu)(1-\phi)\left[u(0) + x - N_f + \lambda EV(0, x')\right], \\ \mu\left[u(1) + \lambda EV(1, x')\right] + (1-\mu)\left[u(0) + x - N_f + \lambda EV(0, x')\right] \end{array} \right\}. \qquad \text{(B.8)}$$

which readily implies that the optimal cutoff under punishment $x_t^{*p} = x_t^{*p} + N_f$. Since punishment is one time only all subsequent cutoffs remain the same. Therefore, as the punishment period was labeled with 1, N_f achieves a reduction in wrongdoing equal to $\phi\left(F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right)$ in the first period. Current punishment affects future wrongdoing through its impact on the share of uncertain individuals who resist in period 1 and enter the future uncertain. Specifically, of those who are saved from temptation in period 1, $\phi\lambda F\left(x_2^*\right)$ are saved again in period 2, so $\phi^2\left(F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right)\lambda F\left(x_2^*\right)$ is the reduction of wrongdoing in period 2 as a result of punishment N_f having been present in period 1. Next, $(\phi\lambda)^2 F\left(x_2^*\right) F\left(x_3^*\right)$ are saved in period three, and so on. As a result, the one time punishment N_f leads to an expected wrongdoing reduction equal to $\phi\left(F\left(x_1^* + N_f\right) - F\left(x_1^*\right)\right)\left[1 + \phi\lambda F\left(x_2^*\right) + (\phi\lambda)^2 F\left(x_2^*\right) F\left(x_3^*\right) + ...\right]$. Because the planner discounts future reductions in crime according to the factor δ, we obtain the expression in the lemma.■

Proof of Proposition 4. The first-order conditions for N_r and N_f are,

$$\begin{aligned}
\phi f\left(N_r\right) - c'\left(N_r + N_f\right) &= 0, \\
\phi f\left(x_1^* + N_f\right) Z - c'\left(N_r + N_f\right) &= 0.
\end{aligned}$$

Solving for $c'(N_r + N_f)$ and combining yields $f(N_r) = f(x_1^* + N_f) Z$. Note from (14) that Z approaches 1 as λ or δ approach zero. Recall that $x_1^* > 0$. Therefore, in the neighborhood of $Z = 1$, $f(x_1^* + N_f)$ is arbitrarily close to $f(N_r)$, which yields $N_f \simeq N_r - x_1^*$ and hence $N_f < N_r$.

The second order conditions are standard and their satisfaction guaranteed by the convexity of costs and the assumption $f'(x) < 0$.

Proof of Proposition 5. Let us parametrize the distribution of temptations with a shift parameter σ, that also captures the mean temptation, such that $F(x|\sigma) < F(x|\sigma') \ \forall x, \sigma > \sigma'$. Denote the mean temptation in the two careers by $\sigma_H > \sigma_L > 0$. There only one period. Given belief μ, from (6), $x^* = \frac{\mu}{\phi(1-\mu)} \left(\frac{b(\mu)^{1-\rho}}{b(\mu)} - 1 \right)$. Notice that x^* is increasing in μ and ρ but independent of σ, and that $\lim_{\mu \to 1} x^*(\mu) = \rho$.

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

$$\begin{aligned} V(\mu, \sigma) &= F(x^*|\sigma) \left([\mu + (1-\mu)\phi] u(b(\mu)) + (1-\mu)(1-\phi) E[x|x < x^*, \sigma] \right) \\ &\quad + (1 - F(x^*|\sigma)) \left(\mu u(1) + (1-\mu) E[x|x \geq x^*, \sigma] \right) \\ &= F(x^*|\sigma) \left([\mu + (1-\mu)\phi] u(b(\mu)) - \mu \right) + \mu \\ &\quad + (1-\mu) \left(\sigma - \phi \int_0^{x^*} x f(x|\sigma) \, dx \right) \\ &= F(x^*|\sigma) \mu \left[b(\mu)^{-\rho} - 1 \right] + \mu + (1-\mu) \left(\sigma - \phi \int_0^{x^*} x f(x|\sigma) \, dx \right). \end{aligned} \tag{B.9}$$

The compensating differential for type μ for entering the low-temptation career is

$$\begin{aligned} V(\mu, \sigma_H) - V(\mu, \sigma_L) &= (F(x^*|\sigma_H) - F(x^*|\sigma_L)) \mu \left[b(\mu)^{-\rho} - 1 \right] + (1-\mu)(\sigma_H - \sigma_L) \\ &\quad - (1-\mu) \phi \int_0^{x^*} x \left[f(x|\sigma_H) - f(x|\sigma_L) \right] \mathrm{d}x. \end{aligned} \tag{B.10}$$

Now hold any $\sigma_L > 0$ as fixed and consider the difference $V(\mu, \sigma_H) - V(\mu, \sigma_L)$. Showing this difference is decreasing in μ proves the proposition: for any $\sigma_H > \sigma_L$, the compensating differential required to attract individuals into the low-temptation sector is decreasing in μ. Denote $M(\mu) \equiv \mu \left[b(\mu)^{-\rho} - 1 \right]$. All terms involving $x^{*\prime}(\mu)$ cancel out by virtue of the envelope theorem, so differentiation of (B.10) with respect to μ yields,

$$\begin{aligned} &V_\mu(\mu, \sigma_H) - V_\mu(\mu, \sigma_L) = \\ &(F(x^*|\sigma_H) - F(x^*|\sigma_L)) M'(\mu) - (\sigma_H - \sigma_L) + \phi \int_0^{x^*} x \left[f(x|\sigma_H) - f(x|\sigma_L) \right] \mathrm{d}x. \end{aligned} \tag{B.11}$$

Using integration by parts (i.e., $\int_0^{x^*} x f(x|\sigma) \mathrm{d}x = x^* F(x^*|\sigma) - \int_0^{x^*} F(x|\sigma) \mathrm{d}x$), (B.11) becomes

$$(F(x^*|\sigma_H) - F(x^*|\sigma_L)) \left(M'(\mu) + \phi x^* \right) - (\sigma_H - \sigma_L) - \phi \int_0^{x^*} \left[F(x|\sigma_H) - F(x|\sigma_L) \right] \mathrm{d}x. \tag{B.12}$$

The first term of (B.12) is negative if $M'(\mu)+\phi x^*$ is positive. And since $\partial b(\mu)/\partial\mu = \phi\left(b(\mu)/\mu\right)^2$ we can write

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

Thus

$$\begin{aligned} M'(\mu)+\phi x^* &= \left[b(\mu)^{-\rho}\left(1-\rho\phi\frac{b(\mu)}{\mu}\right)-1\right]+\phi\left[\frac{\mu}{(1-\mu)\phi}\left(b(\mu)^{-\rho}-1\right)\right] \\ &= \left(\frac{1}{1-\mu}\right)\left[b(\mu)^{-\rho}\left(\frac{\mu+(1-\rho)(1-\mu)\phi}{\mu+(1-\mu)\phi}\right)-1\right]. \end{aligned}$$

This is always positive if

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

which is implied by equation (A.19).

A simple model of conscious control over actions We present a simple setting for the determination of externally observable actions that tracks descriptions in psychology and neuroscience of how the prefrontal cortex (PFC) attempts to influence externally observable actions by biasing the signals available to circuits engaged in implementing those actions. The planner-doer setup of the main model, as well as the formulation with fallible types, are reduced forms of this setup. This framework follows that in Bernheim and Rangel (2004, henceforth BR) in various important respects, but introduces some key modifications. To the extent possible we maintain their notation to facilitate comparison between the two setups.

We follow BR in assuming that external actions are related to the level of susceptibility M to external cues, relative to some threshold M^T (this capital T stands for "threshold" and should not be confused with a terminal time, an object which does not play a role in this subsection). We assume that whenever the realized susceptibility M is above the threshold M^T, the person yields to temptation ($r_t = 0$) and the person resists ($r_t = 1$) otherwise.[24] The realized susceptibility M depends on three elements: a random external cue ω drawn from a known distribution G over the interval $[\underline{\omega}, \bar{\omega}]$, the override attempt by the planner a_t, and the doer's type θ. To make things simple, normalize the threshold M^T to zero, and suppose that susceptibility is determined by the additive relation $M = \theta - a + \omega$, with the additional assumptions that cues predispose towards

[24] Differently from BR we do not suppose that for $M < M^T$ the person selects a consumption decision according to his conscious desires (although it is possible to pursue a model in that direction), but rather that if $M < M^T$ the person behaves "well." Thus, the planner attempts to steer the externally observable action r by biasing the distribution of susceptibility levels below the threshold M^T.

temptation $\bar{\omega} > \underline{\omega} \geq 0$ and that all else equal the good type has lower susceptibility than the bad type: $\theta_g < \theta_b$. Thus, the probability of resisting a temptation is $\Pr(\theta - a + \omega < 0) = G(a - \theta)$, which takes four possible values depending on the values of a and θ. These probabilities satisfy the relations $G(a - \theta_g) \geq G(a - \theta_b), a = 0, 1$, and $G(1 - \theta) \geq G(-\theta), \theta = \theta_g, \theta_b$, which hold with strict inequality whenever the argument $a - \theta$ lies strictly within the interval $[\underline{\omega}, \bar{\omega}]$ on at least one side of each relation. These relations mean that, all else equal, externally observed resistance is more likely when the doer's type is good and when the planner attempts an override.

As in BR, susceptibility M is not directly welfare relevant; it is linked to the subconscious, and to the (possibly very distorted) workings of the reward prediction centers of the brain. There is an inherent discrepancy between the "objective," or welfare relevant, value x of the temptation, which is relevant for the planner's optimization problem, and the cues that trigger automatic response processes. To make the separation as clean as possible we assume no correlation between x and M—the value x will only affect externally observable actions of resistance through its effect on the planner's decision on whether to attempt an override. The planner makes a decision on override after observing the temptation x, knowing that a subconscious susceptibility level M will be drawn and will affect the externally observable action r. The planner knows that if the doer has a bad type, then the susceptibility level will be drawn from a distribution with a higher mean.

At the end of a period t, the planner knows the welfare relevant value x_t, whether or not he has attempted an override a_t, and the realized externally observable action r_t. Critical to the learning process, the susceptibility M_t and its determinants θ and ω_t remain unobserved, so the final action r_t contains information about the doer's type θ.[25]

The baseline model is obtained by imposing a number of restrictions: namely that θ_g satisfies $\theta_g - a + \bar{\omega} < M^T = 0$ for both $a \in \{0, 1\}$ so that good types behave regardless of state and override, and that $\left[\theta_b + \underline{\omega} > M^T = 0, \theta_b - 1 + \underline{\omega} < 0 = M^T\right]$ so that bad types are sure to misbehave if the planner selects $a = 0$, but have a probability $G(1 - \theta_b) \equiv \phi$ of behaving well if the planner selects $a = 1$. The model with fallible types in subsection 2.5 of the paper can also be obtained as a particular case of this more general framework by establishing appropriate relations between $G(a, \theta)$ and the parameters γ_g, γ_b, and ϕ.

[25] It is possible to allow the planner to observe M as well, in which case, under the MLRP property, a lower M would increase the posterior on $\theta = \theta_g$. However, externally observable actions would contain no additional information.