Willpower and Personal Rules

Roland Bénabou[1] and Jean Tirole[2]

First draft May 2000

This version: February 2004[3]

Forthcoming in the *Journal of Political Economy*

[1]Princeton University, NBER and CEPR (rbenabou@princeton.edu).

[2]IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT (tirole@cict.fr).

[3]We are grateful for helpful comments to George Ainslie, Marco Battaglini, Samuel Bowles, John Cochrane, George Loewenstein, Botond Köszegi, Drazen Prelec, Eric Rasmusen, Stefano Della Vigna and two anonymous referees, as well as to conference and seminar participants at many institutions. Sebastian Ludmer provided excellent research assistance. Bénabou gratefully acknowledges research support from the National Science Foundation and the John Simon Guggenheim Memorial Foundation, as well as the hospitality of the Institute for Advanced Study during the academic year 2002-2003.

Abstract

We develop a theory of internal commitments or "personal rules" based on self-reputation over one's willpower, which transforms lapses into precedents that undermine future self-restraint. The foundation for this mechanism is the imperfect recall of past motives and feelings, leading people to draw inferences from their past actions. The degree of self-control an individual can achieve is shown to rise with his self-confidence and decrease with prior external constraints. On the negative side, individuals may also adopt excessively rigid rules that result in compulsive behaviors such as miserliness, workaholism, or anorexia. While apparently displaying a "salience of the future," these phenomena are actually generated by (a concern over) present-oriented preferences. We also study the cognitive basis of self-regulation, showing how it is constrained by the extent to which self-monitoring is subject to opportunistic distortions of memory or attribution, and how rules for information-processing can themselves be maintained.

Keywords: self-control, willpower, motivation, memory, time inconsistency, psychology.

JEL Classification: A12, C70, D60, D91, E21, J22, J24.

1 Introduction

The problem of self-control –or, to use a more ancient term, willpower– is attracting renewed attention from economists. People's common tendency to succumb to short-run impulses at the expense of their long-run interests is generally seen as reflecting conflicting internal preferences, where the individual's current self overweighs the present relative to the future. In recent years such models have been applied to a wide range of economic issues, including consumption and savings decisions, asset pricing, addiction, procrastination, and fiscal policy.[1]

Most of this literature either takes as a given that agents are unable to commit to an optimal course of action, or else it emphasizes the *external* commitment devices that they use to substitute for their deficient willpower: avoiding sources of temptation, holding illiquid assets, signing binding contracts and other forms of tying oneself to the mast or "asking for controls".[2] The present paper, by contrast, focuses on the *internal* commitment mechanisms, or *personal rules,* that receive much greater emphasis in psychology. Examples include diets, monthly savings targets, resolutions to smoke only after meals, jog twice a week, write five pages a day, always finish what you started, conduct your life with dignity, and many similar "promises to oneself."[3]

Given that these rules are entirely self-imposed, the first question that arises is how they could actually constrain a person's behavior. We thus seek to understand and model genuine *self*-control –or, in the words of Adam Smith, self-command, defined as the deliberate and reasoned overriding of powerful impulses and appetites, at the time they occur. To the extent that personal rules can help individuals achieve such goals, the behavioral and economic distortions emphasized in models with either no commitment or costly external controls could well be overestimated.[4]

Building on the seminal work of Ainslie (1992, 2001), we develop a theory of personal rules based on self-reputation. The key idea is that because people have imperfect knowledge of their willpower they see their own choices as indicative of "what kind of a person" they are, implying that lapses can have a severe adverse impact on future behavior. "If I eat this tempting dessert, there goes my whole diet. If I cannot turn down this drink, I might as well admit that I am still a hopeless alcoholic." The fear of *creating precedents* and losing faith in oneself then creates an incentive that helps counter the bias towards instant gratification.

We also show, however, that personal rules may give rise to a very different kind of cost, which

[1] See, e.g., Strotz (1956), Thaler and Shefrin (1981), Ainslie (1992, 2001), Laibson (1997), or O'Donoghue and Rabin (1999).

[2] See, e.g., Ariely and Wertenbroch (2002) for a recent empirical demonstration of voluntarily chosen deadlines. People also attempt to reduce the divergence between their long- and short-run preferences through *cognitive* forms of precommitment, such as strategic ignorance or self-deception about the costs and benefits of perseverance and/or indulgence (Carrillo and Mariotti (2000), Bénabou and Tirole (2002)).

[3] Personal rules have also been discussed by a few economists, most notably Adam Smith (1759) and Schelling (1984). Reduced-form formalizations of some of Smith's ideas can be found in Meardon and Ortmann (1996) and Palacios-Huerta (2002).

[4] See, e.g. Mulligan (1996) for a criticism of hyperbolic discounting models.

until now has received surprisingly little attention in economics. These are the "compulsive" or "obsessive" behaviors of people who feel compelled to work or accumulate constantly without ever properly enjoying leisure or consumption (workaholism, avariciousness), fail to dissave in old age, or even engage in dangerous self-deprivations such as anorexia. Our model can thus account for both underregulation and overregulation, and makes clear that these are often two sides of the same coin.[5] In particular, it shows that agents with hyperbolic discounting can actually behave *as if* they overweighed the future, rather than the present.

There are two key ingredients in our model. The first one is of course imperfect willpower, or salience of the present, in the form of time-inconsistent preferences. The second one –whose essential role the paper brings to light– is *imperfect recall.* Indeed, whenever people look back to their own past actions to infer what they are likely to do in the future, it must be that the motives that led to these actions being chosen at the time are no longer accessible with complete accuracy or reliability. More generally, the paper's primary objective of providing a rigorous account of rule-based behavior is closely integrated with an important secondary aim, which is to analyze the cognitive foundations of self-regulation. To that effect, we first study how equilibrium behavior varies with the degree to which past actions and circumstances are accurately (as opposed to self-servingly) recalled. Ultimately, we examine how memory itself can be endogenously determined through the use of cognitive rules and resolutions.

After discussing related literature and evidence in the rest of this section, we present the model in Section 2. In Section 3 we show how personal rules can be sustained, and establish three main results. First, the degree of self-control achieved by an individual increases with his confidence in his own willpower (keeping actual preferences constant). Second, self-restraint is greater when the situation is repeated, and when lapses are more likely to be brought back to awareness. Third, an initial phase of externally enforced controls or incentives reduces the probability that, later on, the individual will trust himself enough to even put his will to the test. Thus, forced choices imposed by "controlling" parents or rigid social norms inhibit the development of self-confidence and autonomy.

In Section 4 we examine the flexibility or rigidity of personal rules, that is, the extent to which they allow for exceptions. There are again three main results. First, variability in situational factors leads to retrospective attribution (did I give in due to weakness of will or special circumstances?) that can give rise to multiple equilibria, sustained by alternative interpretations of one's own actions. Second, and most importantly, we obtain the potential for both beneficial "bright line" rules and harmful compulsive behaviors. The latter represent costly forms of self-signalling where the individual is so afraid of appearing weak to himself that every decision becomes a test of his willpower, even when the stakes are minor or self-restraint is not even desirable *ex-ante.* Third, we show that compulsive behavior is more likely when self-confidence is low and when the

[5] The terms under- and over-regulation are borrowed from Baumeister, Heatherton and Tice (1994), who provide a very good survey of recent psychological research on these topics.

veracity of excuses and ex-post rationalizations is difficult to ascertain.

Section 5 extends the study of personal rules to the cognitive realm. Better self-regulation can be achieved through strategies like keeping a record of one's behavior or systematically reflecting upon it, using concrete mnemonic aides to mental accounting, or adopting and rehearsing "bright line" resolutions that make lapses more salient. Such cognitive rules must also be self-enforcing, however; we show how, in equilibrium, they are jointly determined with the behavioral rule.

1.1 Related Literature

Personal rules. These behaviors are described by Ainslie (1992, p. 143) as *"the kind of impulse control... which allows a person to resist impulses while he is both attracted by them and able to pursue them".* This definition makes clear the difference not only with external commitments, but also with cognitive strategies that aim to reduce the intensity of temptation, by either manipulating ex ante one's future *perceptions* of the payoffs attached to alternative choices, or by distracting attention away from the source of temptation.[6] The idea is instead that the individual should come to see each decision as a possible precedent for future ones, so that giving in today raises the probability that he will do the same in the future. By thus tying together sequences of choices he raises the stakes on each one, and better aligns his short-term incentives with his long-run interests. The notion that "good" behavior is achieved by choosing according to universal principles rather than individual contingencies has a long history. Ainslie traces it from the writings of Aristotle on ethics and Galen on passions to Kant's categorical imperative that one act as if each choice *"should become a universal law";* and, later on, to Victorian psychologists who stressed that "the unity of the will" is built up by repeatedly framing choices in terms of classes of similar actions, and can be undone by even a little backsliding.

As to *why* misbehaving today should make it more likely that one will also misbehave tomorrow (setting a precedent), Ainslie, like his predecessors, is much more elusive, but what he writes suggests uncertainty and learning about the strength of one's own will.[7] Many other psychologists such as Baumeister et al. (1994) and Rachlin (2000) also emphasize the importance of self-monitoring (keeping track of one's actions) for successful self-regulation, as well as the often devastating effects on a subject's self-view and subsequent behavior of breaking a personal rule. These observations all point in the same direction, leading us to propose a model based on self-reputation over one's degree of time-inconsistency (or salience of the present).

[6] On strategic ignorance and self-deception, see footnote 2. On emotion control and attention control, see Ainslie (1992) and Mischel et al. (1972).

[7] "*But how does a person arrange to choose a series of rewards all at once?... In situations where temporary preferences are likely, he is apt to be genuinely ignorant of what his future choices will be. His best information is his knowledge of his past behavior under similar circumstances...Furthermore, if he has chosen the poorer reward often enough that he knows self-control will be an issue, but not so often as to give up hope that he may choose the richer rewards, his current choice is likely to be what will swing his expectation of future rewards one way or the other.*" (Ainslie, (1992), p. 150).

Self-signaling. The idea that people learn about themselves by observing their own actions, and conversely make choices in ways designed to achieve or preserve favorable self-conceptions, is quite prevalent in psychology, and well supported empirically.[8] Through this key feature our model is closely related to the work of Bodner and Prelec (1997, 2001), who examine the diagnostic value of actions in a dual-self model where the individual has "metapreferences" over his own, imperfectly known, tastes. Formally, they consider an intratemporal (one-shot) signaling game between two contemporaneous, asymmetrically informed subselves, such as ego and superego. For instance, the individual may derive hedonic value from thinking of himself as a generous person, and take actions to try and convince himself that he is of that type.[9]

Addiction. Our work is also related to theories of addiction. Thus, as in Becker and Murphy's (1988) model, consumption today increases the likelihood of consumption tomorrow, and the fear of addiction may generate conducts that resemble compulsive behavior.[10] There are also key differences, however. First, no matter how unhappy rational addicts may be along their optimally chosen consumption paths, they would be even more unhappy if prevented from consuming the addictive good, or lied to about its effects. By contrast, our impulsive and compulsive agents (those who are not able to sustain good behavioral rules) have strong demands for external commitments or information manipulations that would prevent them from behaving in this way. Second, the theory of rational addiction allows little scope for relapse following a sustained period of forced abstinence, as the stock of "addictive capital" depreciates over time. Yet relapses are very common, and even when they are avoided the cravings and temptations often persist.[11] Our model is consistent with this fact, and even shows that the temporary imposition of external restraints can actually *reduce* an individual's capacity to moderate his consumption later on.

1.2 Evidence on rules in economic decisions

We now turn to more standard decisions over consumption and labor supply, where one observes a broad class of rules that fall under the heading of *"mental accounting"* (Thaler (1980), (1985)). We briefly review here evidence from both consumer research and economic studies documenting such behaviors. We also point out that while substantial attention has been paid to the details

[8] See e.g., Bem (1972) on self-perception and self-persuasion, and Quattrone and Tversky (1976) on the confounding of causal and diagnostic contingencies. In a well-known experiment, the latter found that subjects who were led to believe that tolerance for a certain kind of pain (keeping one's hand in very cold water) was diagnostic of either a good or a bad heart condition reacted by, respectively, extending or shortening the amount of time they withstood that pain.

[9] Hirschleifer and Welch (2001) emphasize imperfect memory and retrospective inference, as we do, but focus on a different set of questions, relating to excessive inertia or volatility in individual and organizational decisions.

[10] On this last point, see Orphanides and Zervos' (1995) generalization of Becker and Murphy's model to allow for ex ante uncertainty over one's susceptibility to addiction. For a different view, based on a self-control problem, see Carrillo (2004).

[11] Thus *"patients with addictions and other impulsive disorders report intense, continuing urges to backslide even after years of continence,"* and *"minor disturbances in a person's regimen can produce episodes of renewed impulsiveness"* (Ainslie (1992, p.125). See also Loewenstein (1999), who emphasizes the role of external cues in triggering sudden and powerful relapses.

of *how* these mental accounts are structured, the question of *why* they work –what makes them self-enforcing– needs to be examined more closely.

Ethnographic accounts document the prevalence, especially in poor households, of "mental budgeting", the hallmark of which is a violation of the fungibility of money (e.g., Zelizer (1997)). Individuals or families thus earmark certain sources of incomes to specific uses (primary wage earner's salary for necessities, secondary earnings for savings, windfalls and capital income for luxuries, etc.), or keep in separate envelopes or "tin cans" the monies reserved for food, rent, school supplies, "fun," and the like. In experiments, Heath and Soll (1996) similarly find that college students pre-set for themselves binding budgets for specific categories of goods, particularly "hedonic" ones. Thus, their stated willingness to pay for entertainment (say) was significantly lower if they had already made a purchase in the same category (e.g., gone to a restaurant) at the start of the reference period.[12] Ameriks, Caplin and Leahy (2003) find that 37% of their TIAA-CREF survey respondents set a detailed spending budget for themselves, and of these 45% declare that without it their "spending would rise a great deal". Using both experimental and supermarket scanner data, Wertenbroch (1998) finds evidence that consumers use self-rationing rules to limit their consumptions of "sinful" products (e.g., never buy more than n units at a time, or per week). Comparing pairs of matched "virtue" and "vice " products for a host of food, drink and tobacco categories (i.e., reduced fat, calorie, alcohol, or tar versions, versus regular ones), he finds that the latter are consistently characterized by smaller package sizes that sell at a premium, and lower price elasticities of purchases in response to quantity discounts.

A related set of personal rules against impulse spending are those leading to "debt aversion" (Prelec and Loewenstein (1998)). In experiments as well as a field study using actual spending records, Wertenbroch, Soman, and Nunes (2002) found that individuals with a stronger need for self-control, whether situational (purchasing a hedonic good, versus a utilitarian one) or personal (a high rating on Puri's (1996) impulsivity scale), had a greater propensity to pay by cash, check or debit rather than by credit card, particularly when the latter had an indefinite term rather than a 30-day limit. When they did consider financing (e.g., for a car purchase), subjects also expressed a preference for self-imposing shorter payment terms, even at a premium, if the item was framed as hedonic rather than utilitarian.

There is also evidence that personal rules against profligacy can be excessively binding –as in our compulsiveness results– so that ex-ante, people may want to "precommit to indulgence". Kvetz and Simonson (2002) had subjects choose which of two prizes they would receive if they won a subsequent lottery or sweepstakes (sometimes fictitious, sometimes actual). When given a choice between a "hedonic" good (massage, facial, fancy wine, gourmet meal) and a cash prize of equal or even higher value, a significant minority (about 25%) of subjects chose the good. When

[12] This effect was not present if the prior hypothesized expense was instead a monetary loss of the same amount, such as a parking ticket (thus ruling out global budget effects), nor if the prior consumption had been received as a gift (thus controlling for satiation).

the proposed prize was "utilitarian", by contrast (credit towards grocery bills), almost everyone chose cash (92%); and when given a direct choice between winning either type of good, 64% chose the utilitarian one. Furthermore, those who chose the self-indulgent good over cash explained it in terms of precommitting to enjoyment, and fear that the money would have gone to utilitarian uses such as paying the rent.

Finally, personal rules also apply to the uses of time as well as to those of money. Analyzing the earnings and hours of work of New York City cab drivers, Camerer, Babcock, Loewenstein and Thaler (1997) find evidence that many of them follow a daily-earnings target rule (implying a backward-bending labor supply), whereby they quit when, and only when, they have met their targeted amount.[13] Among the proposed explanations is that this may serve as a self-control device, helping drivers resist "the temptation to quit early today" what is often a tedious and exhausting job.

Mental accounts and other personal rules thus appear to be as common in economic decisions as in personal or health-related ones (e.g., dieting). Yet while their value from an ex-ante point of view seems quite intuitive, the mechanism by which they could actually constrain someone's impulses ex-post remains elusive. Indeed money *is* fungible, and no external constraints prevent a tempted individual from dipping into the wrong mental account or paper envelope, perhaps with the "hope" of making it up later.[14] Similarly, whereas leaving one's credit card at home to avoid impulse purchases of tempting goods is a clear external precommitment device, once actually making a hedonic purchase that decision is sunk, so choosing cash when a credit alternative is available (as in the experiments, and many instances in real life), or opting for a shorter, more costly loan, must be something that the individual finds "internally" optimal to do. The same is true for following through on a long-term financial plan (Ameriks et al. (2001)), which after all is essentially a resolution. And while the self-control value of a daily earnings target seems readily apparent, one must ask what compels the driver, alone and exhausted in his cab, to stick to the rule ex post and stay longer on the job on a bad day where customers are few and far between.

Our model provides an explicit answer, namely self-reputation, to the question of *what gives force* to personal rules. Furthermore, we document throughout the paper that both its premises and its results match many ideas and experimental findings of psychologists, which are thus brought together in a unified analytical framework. Is there also direct evidence that rule-based behavior is mediated by self-reputation? While no study has focused directly on this question, a clever experiment by Kirby and Guastello (2001) provides support for the closely related decision-linking effect, whereby people resist temptation because of fear that doing otherwise would mean that they will succumb to it again in the future. In a first stage, each subject's preferences for

[13] See, however, Farber (2003) for a dissenting study on this subject.

[14] The case to bear in mind is that of a household with a single decision maker, e.g. a student or single parent. When there are two parents with divergent preferences, envelopes and tin cans are also ways of monitoring the other, and deviations from the rule (e.g., drinking the school money) do have external consequences.

monetary payoffs at different horizons were elicited using computerized second-price auctions, until a pair of payoffs $y < Y$ exhibiting preference reversal was found –namely, such that he or she prefers y today over Y in 6 days, but Y in 16 days to y in 10 days. In a second stage subjects faced the same choice of y immediately versus Y six days later, but were (truthfully) advised that it would be offered to them again in 10, 20, 30 and 40 days, with complete freedom to make either the same or a different selection on each occasion. A statistically significant 33% of participants then switched to the patient behavior, merely from considering a series of five repeated but *entirely independent* choices. Explicitly suggesting a precedent or diagnostic effect further raised this proportion to 46%, although given the sample size the difference was not significant.[15] These results lend support to the idea that people see their current choices as predictors of future behavior, and that awareness of this (purely internal) linkage helps them overcome temptation.

2 The Model

2.1 Preferences and Decisions: State-Dependent Willpower

We consider an individual with a horizon of two periods, $t = 1, 2$ (e.g., a weekend), each of which is itself divided into two subperiods, $\tau =$ I, II (e.g., morning and afternoon). At the start of each subperiod I, this agent chooses between:

1) A "no willpower"(NW) option, which yields a known, immediate payoff a. This corresponds to indulging in immediate gratification by drinking, smoking, eating, spending or slacking off without even trying to resist the urge. The important point is that by pursuing this course of action the individual *avoids putting his will to the test.*

2) Undertaking a "willpower-dependent" (W) activity or project: attempting to exercise moderation or abstinence in drinking, smoking, spending, etc., working on a challenging task (homework, exercising, ambitious project, etc.), or participating in a social relationship.

If he attempts W the agent will face at the start of subperiod II a decision of whether to *persevere* until completion (P), or *give up* along the way (G). Holding fast entails a "craving" cost $c > 0$ but yields future payoffs (better health or career, higher consumption) whose present value, evaluated as of the end of period t, is B. Caving in ensures a painless subperiod II, but yields only a delayed payoff $b < B$.

The three possible outcomes –not even trying, trying but later quitting, or trying and persevering– are illustrated on Figure 1**.** In many cases *some* self-control is better than none, even if it does

[15] This "suggested linking" condition was identical to the previous "free linking" one, except that subjects were also told that "the choice you make now is the best indication of how you will choose every time", so that they "would probably" make the same choices again. In a third condition subjects chose once and for all between the two series of rewards. While such "imposed linking" is not informative about precedent effects, the fact that it led 84% of subjects to switch to the delayed rewards is further evidence of hyperbolic-like discounting.

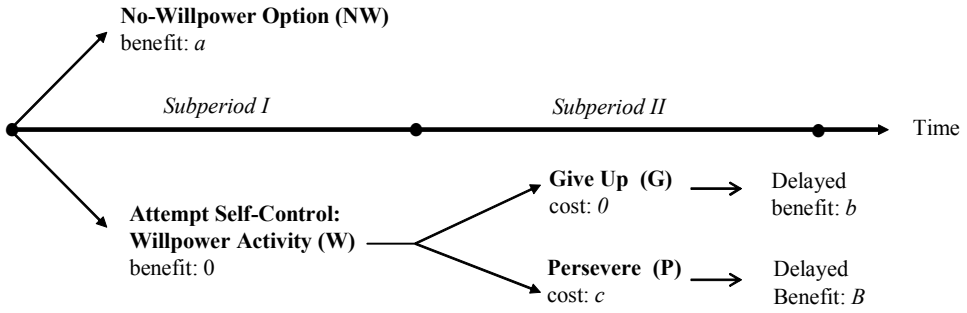

Figure 1: decisions and payoffs in any given period $t = 1, 2$.

not last forever ($b > a$). Thus, a person is better off if he refrains from smoking or does some work for a while than if he makes no effort at all. On the other hand it may be better not to start a difficult project that one is unlikely to complete, or a firm that eventually fails ($a > b$).

In addition to a standard discount rate δ between periods 1 and 2, our agent's preferences exhibit time-inconsistency. Specifically, we enrich the standard quasi-hyperbolic specification with two important new elements: the intensity of temptation, or "salience of the present", is *state-contingent,* and *imperfectly known* in advance. Thus:

1) When choosing between NW and W, the immediate payoff a to be received from the first option may be particularly salient or tempting, relative to the costs and benefits which the second one would bring about, starting next period. Accordingly, the agent discounts the latter at a rate $\gamma \leq 1$; equivalently, he values the immediate gratification from NW at a/γ instead of just a.

2) If he nonetheless decides to attempt W, he is again confronted with another (typically more intense) temptation during subperiod II: the cravings he experiences loom larger than the future benefits of holding fast, so in his decision-making the cost c gets magnified to c/β, with $\beta < 1$. Equivalently, all future payoffs are discounted by β.

In general β and γ need not be equal, as they reflect preferences in very different situations. For instance, $1/\gamma$ measures how much the individual craves a cigarette or a glass of alcohol in the morning, and $1/\beta$ how much he craves it in the afternoon, after hours of deprivation. Another key difference, discussed below, is that whereas γ is known at the initial stage when the agent decides whether to attempt W, β is *revealed only through the experience* of actually putting one's will to the test. A natural interpretation is that one's ability to resist impulses is a known γ in normal times, but may be either $\beta_H \leq \gamma$ or $\beta_L < \beta_H$ in times of stress –whether caused by abstinence, the proximity of temptation, or cues that intensify "visceral" cravings (Loewenstein (1996), (1999)).[16]

[16] One would expect β and γ to be correlated across individuals. This is allowed by our assumptions (e.g., let

The fact that $\beta < 1$ creates a conflict of interest between the individual's successive temporal selves. At the start of each period t he would like the W activity, if he attempts it, to be pursued until completion unless $c > B - b$. Ex post, however, he will give up whenever $c/\beta > B - b$ (in the absence of reputation concerns). We shall thus refer to β as the individual's *strength of will,* that is, his intrinsic ability to withstand discomfort and delay gratification in situations of intense temptation or stress.

Similarly, when $\gamma < 1$ the agent in period 1 would like W to be undertaken in period 2 as long as it yields an expected benefit of at least a. But, ex post, he will instead take the "path of least resistance" and select NW whenever that same expected benefit is less than a/γ.[17] This distortion in the future self's preferences is what creates an incentive to hide from him the fact that one's will may be weak, or make sure that he learns that it is strong: if tomorrow I recall that today I caved in, I will not even attempt self-control –"what is the point?"– but default to NW. As in Seligman's (1975) theory of learned helplessness, *"quitting is mediated by thinking that one will fail".*

2.2 Information: Experience and Memory

As discussed earlier, people frequently look back to their own past behavior to infer what they are likely to do in similar situations. Conversely, their choices often reflect important concerns for maintaining "self-respect" or other valued "identities". The starting point of any theory that aims to shed light on these phenomena *must* be an imperfect recall of one's earlier preferences: for past actions to be informative, the motives that led to their being chosen in the first place must no longer be accessible with complete accuracy or reliability.

The key assumption in our model is thus that people have limited knowledge of their strength or weakness of will: it can be truly known only through direct experience, and later on its value, no longer salient, is hard to remember through pure introspection.

Assumption 1 *The individual's strength of will (or degree of time consistency) when confronted with the decision to persevere or give up in the W activity is fixed over time, and equal to either β_L or β_H, with $\beta_L < \beta_H \leq 1$. He initially does not know β, but has priors ρ_1 and $1 - \rho_1$ on β_H and β_L. Furthermore, if in period 1 he momentarily discovers the true value of β through the experience of craving, later on he cannot directly recall this value.*

There are several reasons, and supporting bodies of evidence, why β cannot just be directly remembered from previous experience. First, it appears difficult for people to accurately recall from

$\beta_H = \gamma$), as long as there remains uncertainty over β. A variant of the model where $\beta = \gamma$ is fixed but types differ by their craving costs c leads to similar results (see Battaglini, Bénabou and Tirole (2002)).

[17] Note that NW need not yield a flow payoff only in subperiod I. The key assumption, in line with the examples given earlier, is that the course of action whose final payoff is least sensitive to the degree of willpower in subperiod II (action NW) is also the one that yields more instant gratification in subperiod I. In particular, NW could also lead to the P/G decision node, but with a lower probability than W, without changing any of the results.

"cold" introspection the intensity of stress, temptation or other short-run feelings corresponding to "hot" (visceral, emotional, not easily quantifiable) internal states experienced in the past. Kahneman, Wakker and Sarin document a systematic divergence between subjects' moment-by-moment reports during painful medical procedures or unpleasant laboratory experiments, and their later retrospective evaluation of the experience a whole. Loewenstein and Schkade (1999) report on many other experiments and field studies indicating that similar "hot-cold empathy gaps" occur in recollections and predictions about feelings (and the behaviors they trigger) such as hunger, exhaustion, drug or alcohol craving, or sexual arousal.[18]

The second reason why this economist-like *revealed preference* approach to predicting one's *own* behavior is warranted is that an individual will often have, ex post, a strong incentive to "forget" that he was weak-willed ($\beta = \beta_L$) and "remember" instead that he was strong-willed ($\beta = \beta_H$). Subjective memories of past feelings and motives represent very "soft" information whose veracity is much more difficult to verify than for the "harder" data of one's past deeds, which often leave a material record. Indeed, a lot of research has confirmed the common observation that recollections are often *self-serving:* people tend to remember their successes more than their failures, reframe their actions so as to see themselves as instrumental for good outcomes (or at least, motivated by honorable intentions) and absolve themselves of bad ones by attributing responsibility to others.[19]

Finally, accurate self-knowledge is often made more difficult by the need to disentangle true character from situational factors: did I give in to temptation because my willpower β is low, or because of an unusually high cost c? The next assumption will allow us to analyze the role of exceptions and excuse-making in rule-based behavior.

Assumption 2 *The disutility of deprivation or craving, c, is an i.i.d. random variable that takes values c_L with probability π and $c_H \geq c_L$ with probability $1 - \pi$.*

In interpreting his past behavior the agent thus faces a problem of signal-extraction, or *attribution.*[20] Of course, he will often have strong incentives to try and forget his failings, or, if he cannot, to try and "rationalize" them ex post by placing the blame on temporary or external causes. We therefore allow for imperfect recall, or imperfect retrospective verifiability, with respect to both past actions and past circumstances.

[18] As Loewenstein (1996, p.284) writes: *"It seems that the human brain is not well equipped for storing information about pain, emotions, or other types of visceral influences, in the same way that visual, verbal, and semantic information is stored".*

[19] For references to the psychology literature and experimental evidence, see Bénabou and Tirole (2002).

[20] As stated by Baumeister et al. (1994, p.19): *"There are three main reasons that someone would have inadequate strength for successful self-regulation: one chronic, one temporary, the other external. The person may...[be] a weak person who would probably never be able to override that same impulse. Alternatively, the person may be tired or exhausted... Lastly, the impulse may be so strong that even someone with well–developed self-regulatory skills would be unable to conquer it".* The first case corresponds to $\beta = \beta_L$, the second and third to temporary decreases in β and increases in c respectively. Since all that matters is c/β we merge them into fluctuations in c only. Alternatively, one could introduce temporary fluctuations in the long-run benefits from perseverance, $B - b$.

Assumption 3 *Suppose the agent attempts self-control* (W) *in period* 1. *If he perseveres, no lapse will be recalled at date* 2. *If he gives in to temptation he will remain aware of this lapse only with probability* λ. *With probability* $1-\lambda$ *he will have "forgotten" (become unaware of) it, and thus no longer be able to distinguish this state of the world from one where he really held fast.*[21]

Assumption 4 *If the cost of effort or craving at time* 1 *is high,* $c_1 = c_H$, *it will never be recalled at date* 2 *that perseverance was easy. If* $c_1 = c_L$, *on the other hand, the agent will recall it only with probability* ν; *with probability* $1-\nu$ *he will no longer be able to distinguish this state of the world from one where* c_H *occurred.*

The assumption that ego-favorable events are more likely to be remembered than unfavorable ones is not essential for our main results, but is supported by considerable empirical evidence.[22] Most importantly, permitting λ and ν to vary in $[0,1]$ (first parametrically, then endogenously) will shed light on the cognitive underpinnings of self-regulation, by linking the types of behaviors that a person can sustain to the reliability of his memory and inference processes.

3 Personal Rules and Self-Reputation

We are interested in personal rules that are self-enforcing, without reliance on external commitment devices.[23] Furthermore, a key aim of our theory is to demonstrate the central role played by the *uncertainty* which people face concerning their own preferences (here, willpower). This leads us to model the problem as a game of imperfect information between the individual's "incarnations" at each subperiod $(t,\tau) \in \{1,2\} \times \{I, II\}$, and to define a sustainable *behavioral rule* as a Perfect Bayesian Equilibrium (PBE) of that dynamic game.[24] This implies in particular that while the individual may engage in self-deception, at each point in time he processes all currently available information according to rational inference, taking into account any tendency he might have to selectively forget certain types of events.

[21] Formally, let us denote by $\sigma \in \{G,P\}$ the action chosen by Self 1 and by $\hat{\sigma} \in \{G,P\}$ the corresponding signal (or "message") that is encoded into memory and eventually retrieved by Self 2 –in other words, the individual's later recollection of σ. Then $\Pr(\widehat{\sigma} = P \mid \sigma = P) = 1$, but $\Pr(\widehat{\sigma} = G \mid \sigma = G) = \lambda \leq 1$. With similar notation, Assumption 4 means that $\Pr(\widehat{c} = c_H \mid c_1 = c_H) = 1$ but $\Pr(\widehat{c} = c_L \mid c_1 = c_L) = \nu \leq 1$.

[22] Selective recall also emerges endogenously when the individual has some measure of control –through rehearsal, cue or attention management, etc.– over his memory or awareness (see Bénabou and Tirole (2002)).

[23] *"Personal rules are promises to cooperate with the individual's own subsequent motivational states.... They are self-enforcing insofar as the expected value of cooperation exceeds that of defection at the time choices are made... It is this stake that gives the will his force."* (Ainslie (1992), p. 161-162).

[24] An another route would be to assume that the agent knows β, and analyze the infinitely repeated game (Laibson (1994)). This leads, however, to the usual problems of indeterminacy (infinity of equilibria) and lack of robustness to (intrapersonal) renegotiation of the underlying trigger strategies. In a learning model, by contrast, there is an actual state variable, namely one's self-image, that is irrevocably changed when one commits a lapse or successfully resists temptation –as amply documented in the psychology literature. For an alternative approach to the "unity of the self," see Caillaud et al. (1999).

We first study here how impulses for immediate gratification can be held in check by the fear of "losing faith in oneself," leading to a further collapse of self-control. On the cognitive side, we highlight the role played by the recall of past lapses (λ) in the maintenance of personal rules. To abstract from the additional problem of assessing excuses, we let $c_H = c_L = c$, with:

$$\frac{c}{\beta_H} < B - b < \frac{c}{\beta_L}. \tag{1}$$

Thus, absent reputational concerns, the β_L type always caves in and the β_H type never does. Such is the case in the second, last period. Knowing this, the agent will attempt W at the start of period 2 only when his self-confidence ρ_2 is above the threshold ρ_2^{W}defined by:

$$\rho_2^{\mathsf{W}}(B - c) + (1 - \rho_2^{\mathsf{W}})b \equiv \frac{a}{\gamma}, \tag{2}$$

which is between 0 and 1 as long as the following assumption is satisfied:

Assumption 5 $B - c > a/\gamma > b$.

Note that in (2) the payoffs on the left-hand-side are evaluated ex ante, since the agent's preferences between P and G are not yet subject to the distortion $1/\beta$ that he will experience later on, when actually confronted with the craving c. By contrast, his valuation of the NW option reflects his current temptation to pursue the immediate gratification of a/γ afforded by taking "the easy route" of avoiding any exercise of willpower.

3.1 Lapses as Precedents

We focus here on the case where some self-control is always better than none: $b > a$. Thus, ex ante, the individual would prefer that W be undertaken, even if he was certain to eventually give up.[25] Ex post, however, he is too tempted to not even try (NW) : when $\gamma < 1$, the threshold ρ_2^{W}in (2) is suboptimally high from Self 1's point of view. Because confidence in his own willpower (a higher ρ_2) helps shore up motivation it represents a valuable asset, worthy of protection in period 1. Formally, let $p_2(\rho)$ be the probability that W is selected at the start of period 2, and denote the resulting ex-ante (temptation-free) value functions for each type as $V_2^H(\rho) \equiv p_2(\rho)(B - c) + (1 - p_2(\rho))\, a$ and $V_2^L(\rho) \equiv p_2(\rho)b + (1 - p_2(\rho))\, a$. Since $p_2(\rho)$ equals 0 for

[25] By contrast, when $b \leq a$ he would never want to convince himself (i.e., his futures selves) that he is strong-willed when he is in fact weak-willed. This dichotomy is only a result of our simplifying assumptions, however, not a limitation of the theory. If the weak type's willpower is allowed to fluctuate across periods, so that $\Pr[\beta' = \beta_H \mid \beta = \beta_L] = \alpha < \rho_2^*$, his incentive to pool can remain even with $b < a$. Indeed if $a/\gamma > (1-\alpha)b + \alpha(B - C) > a$ all the results in this section go through, except that in Proposition 1 ρ_2^+ becomes

$$\rho_2^+ \equiv \frac{\rho_1 + \alpha(1-\rho_1)(q_1 + (1-q_1)(1-\lambda))}{\rho_1 + (1-\rho_1)(q_1 + (1-q_1)(1-\lambda))},$$

while $\tilde{\rho}_1(\lambda)$ is obtained by setting $\rho_2^+ = \rho_2^*$ with $q_1 = 0$ in this expression. Note also that all the "compulsiveness" results in Section 4, which involve separation by the strong type, already apply whether $b \gtrless a$.

$\rho < \rho_2^*$, 1 for $\rho > \rho_2^*$, and is unconstrained in-between, V_2^H is also an increasing step-function of ρ, and so is V_2^L when $b > a$.

We now consider behavior in period 1. We shall focus attention on equilibria satisfying the natural assumption of *monotonicity in beliefs,* that is, such that not recalling any lapse always raises (weakly) the probability that the individual is a strong-willpower type, while recalling a lapse always lowers it (weakly). Formally, $\rho_2^+ \geq \rho_1 \geq \rho_2^-$ for all ρ_1, where ρ_2^+ and ρ_2^- respectively denote the posteriors in each of these events.[26] Because $B-b > c/\beta_H$ and $V_2^H(\rho)$ is non-decreasing in ρ, monotonicity of beliefs implies that perseverance (P) is a dominant strategy for type β_H. As to type β_L, he will exert self-control if:

$$\frac{c}{\beta_L} - (B-b) \leq \delta\lambda \left[V_2^L(\rho_2^+) - V_2^L(\rho_2^-)\right]. \tag{3}$$

The left-hand side is the disutility of resisting temptation, while the right-hand side represents *the value of self-reputation* that will be foregone if one does, and this lapse is recalled next period. Since this reputational stake is at most $\delta\lambda(b-a)$ we shall assume that

$$C(\lambda) \equiv B - b + \delta\lambda(b-a) > \frac{c}{\beta_L}, \tag{4}$$

otherwise the unique equilibrium is one where the weak type always gives in to his impulses. Finally, we define, for each λ, the following threshold:

$$\tilde{\rho}_1(\lambda) \equiv \frac{(1-\lambda)\rho_2^*}{1-\lambda\rho_2^*}. \tag{5}$$

As we shall see, this is the minimal level of self-confidence required to sustain any self-discipline.

Proposition 1 *When $c/\beta_L < C(\lambda)$ there is a unique equilibrium. If the agent's initial reputation ρ_1 is below a threshold $\rho_1^* < \rho_2^*$ he does not put his willpower to the test, and chooses the NW option. For $\rho_1 > \rho_1^*$ he chooses W, in which case:*

i) the strong-willed type always perseveres;

ii) the weak-willed type perseveres with probability 1 for $\rho_1 \geq \rho_2^*$, with probability q_1 such that

$$\rho_2^+ \equiv \frac{\rho_1}{\rho_1 + (1-\rho_1)\left(q_1 + (1-\lambda)(1-q_1)\right)} = \rho_2^* \tag{6}$$

for $\tilde{\rho}_1(\lambda) < \rho_1 < \rho_2^*$, and with probability 0 for $\rho_1 < \tilde{\rho}_1(\lambda)$. Thus, q_1 rises with ρ_1 and λ.

In period 2, if a lapse is recalled the NW option is chosen. If none is recalled the W activity is chosen with probability 1 if $\rho_1 > \rho_2^*$ and with probability $p_2^* = \left(c/\beta_L - B + b\right)/\left[\delta\lambda(b-a)\right]$ if $\rho_1 \in [\tilde{\rho}_1(\lambda), \rho_2^*]$

[26] This restriction on out-of-equilibrium beliefs is relevant only when $\lambda = 1$, in which case it eliminates the unnatural equilibrium where both types choose G because beliefs would be very pessimistic following P. This kind of equilibrium is also not robust –for instance, it disappears if costs are random with a wide enough support.

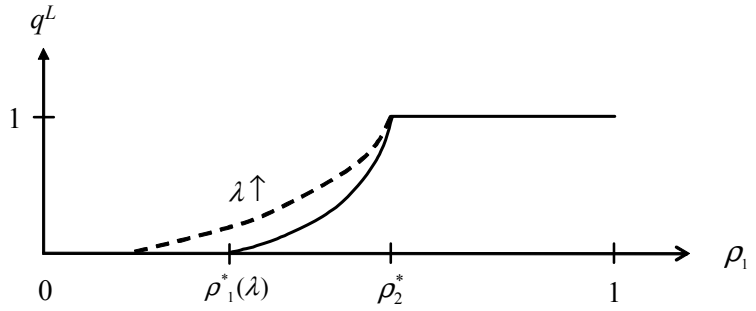

Figure 2: self-control by the weak-willed.

3.2 Main Implications

Proposition 1 has a number of important implications, illustrated on Figure 2.

1. The value of self-confidence. Self-control is easier to achieve for someone who is more confident in his strength of will, as this reputational capital can be staked upon proper behavior in the current period. Conversely, as ρ_1 declines below ρ_2^* self-control becomes increasingly difficult to achieve: mimicking the strong type by playing W with a high probability would leave priors largely unchanged, and thus fail to prevent NW from being chosen in at $t = 2$. Instead, q_1 must be low enough to make recalling no lapses from period 1 sufficiently "good news" about his type that the individual in period 2 becomes (just) willing to again put his will to the test. Finally, ex-ante welfare is easily shown to also be increasing in ρ_1.

2. External controls and autonomy. Suppose that during the first period (e.g., childhood), the individual's behavior is subject to external constraints, imposed for instance by strict parents or a society with rigid norms: when confronted with temptation, he is forced or strongly incentivized to select P. By contrast, date-2 behavior (e.g., in adulthood) remains subject to free will. Proposition 1 shows that such controls always (weakly) reduce the likelihood that the agent puts his will to the test in period 2. Indeed nothing is learned from period 1's actions, so ρ_2 remains equal to ρ_1. For $\rho_1 \geq \rho_2^*$ this has no impact, as he would have chosen P and then W anyway. For $\rho_1 < \rho_2^*$, however, it prevents him from acquiring the self-confidence needed to make the right choices on his own later on. Thus at $t = 2$ he will not even attempt self-restraint, whereas he would otherwise have chosen W with probability $p_2^* > 0$ (for $\rho_1 \geq \widetilde{\rho}_1(\lambda)$). It also follows that outside constraints will now be eagerly sought out by the individual in period 2 : a *dependence* on external rewards and punishments has been created.

External constraints thus have both benefits –better behavior initially– and costs: inferior reputation building and loss of autonomy.[27] Figure 2 shows that, on net, they are beneficial for

[27] We assume here that the individual is aware of having been controlled in period 1. The case where he is unaware of it is closely related to that of "false excuses" (recalling $\hat{c} = c_H$ even though c_1 was really equal to c_L), studied in

low self-confidence, costly for intermediate self-confidence, and irrelevant for high self-confidence. Finally, it is also the case that the anticipation of *future* constraints undermines *current* self-discipline, since it makes reputational capital less valuable. Thus a child or employee who is usually subject to tight external controls and sanctions has little intrinsic motivation to behave on occasions where he can "get away with it".[28]

3. The role of memory. Figure 2 shows that q_1 is increasing in λ : quite intuitively, the individual is more likely to persevere, the less forgettable are his lapses. It is also easily seen that ex-ante welfare is always higher with $\lambda = 1$ than with $\lambda = 0$.[29] As Ainslie (1992, p. 154) notes: *"Behavior therapists regularly observe that when patients systematically record either impulsive behaviors or avoidances of such behaviors, the occurrence of such behaviors decreases; a practice called self-monitoring".* Similarly, the use of envelopes to implement mental accounts, or a rule to always use checks when paying for self-indulgent goods –neither of which restricts the actual choice set– can now be understood as mnemonic or self-monitoring devices meant to increase the recall of lapses, λ.[30] We also note, however, that an individual would often want to cheat on his self-monitoring ex post, and forget his lapses so as to avoid the reputational damage. We shall come back to this issue when examining the sustainability of cognitive rules.

4 Rules, Exceptions, and Excuses

When really ill or on important family occasions, even the strongest-willed person would (or should) postpone work to another day. During extreme weather the jogging-every-day rule should be broken, and when a host prepared a special dessert it would be more impolite than heroic to refuse. In assessing one's strength of will, attention must thus be paid to the circumstances under which past actions took place. To analyze this attribution problem, let actions now always be remembered ($\lambda = 1$), but let the cost of resisting impulses take values c_H and c_L such that, in a static context, the weak-willed type always gives in, while the strong-willed type only does so when $c = c_H$.

---

the next section. Another way that intrinsic motivation may be damaged is if the "principal" imposing the external constraints or incentives is privately informed about the agent's ability or willpower, or about the nature of the task. See Bénabou and Tirole (2003) for such a *social* theory of extrinsic versus intrinsic motivation.

[28] These theoretical predictions are all in line with the evidence in psychology (and daily life) that "*The use of external constraints where personal rules might have served may undermine the maintenance of personal rules... Children who are largely restrained by parental pressure grow up to use fewer internal controls."* (Ainslie (1992, p. 174-175)). For further discussions and evidence, see also Lepper and Greene (1978).

[29] It is not obvious, however, whether it always rises monotonically with λ. Such is the case when $\rho_1 \geq \rho_2^*$, as a higher λ just makes it more likely that $c/\beta_L < C(\lambda)$. For $\rho_1 < \rho_2^*$, however, note that p_2^* is decreasing in λ.

[30] This memory-management view receives empirical support from Soman (2000), who tested (experimentally and in the field) consumers' recall of their recent expenditures (items and amounts of purchased in the previous month, or even hour). He found it to be much higher for those who used means of payments requiring them to write down the amount (checks and debit checks), or which implied an immediate rather than a delayed depletion of assets (debit card), than for credit and charge cards. Rehearsal and immediacy also significantly lowered subsequent purchase intentions.

Assumption 6 *Let $c/\beta > B-b$ for all $\beta \in \{\beta_H, \beta_L\}$ and $c \in \{c_L, c_H\}$, except for $c_L/\beta_H < B-b$.*

Furthermore, as laid out in Assumption 4, the occurrence of c_L is correctly recalled or interpreted at $t = 2$ only with probability $\nu \leq 1$, while with probability $1 - \nu$ the individual can no longer reliably tell whether c_L or c_H occurred. The first case (recall state $\hat{c} = c_L$) describes situations where no special circumstances can credibly be invoked to excuse a lapse, or make perseverance more heroic; the second (recall state $\hat{c} = c_H$) to those where such a claim can plausibly be made –it may or may not be true, but cannot be disproved.

4.1 Feasible Behavioral Rules

We first describe the four basic patterns of behavior that can occur in equilibrium, each of which highlights an interesting aspect of the problem, then turn to mixed-strategy combinations.

A. Pure strategies. We shall impose parameter restrictions such that P is always dominant for (β_H, c_L) and G dominant for (β_L, c_H). This leaves four possible types of pure strategy configurations to be played in each period:

$R_0 \equiv$

	β_H	β_L
c_L	P	G
c_H	G	G

, $R_1 \equiv$

	β_H	β_L
c_L	P	P
c_H	P	G

, $R_2 \equiv$

	β_H	β_L
c_L	P	P
c_H	G	G

, $R_3 \equiv$

	β_H	β_L
c_L	P	G
c_H	P	G

.

1. *Impulsive behavior.* The first rule, R_0, is the familiar impulsive one ("carpe diem"), where each type acts myopically. It will describe in particular the endgame played in period 2.

2. *Flexible or contingent rule.* The most desirable rule –if it can be sustained– is often R_2, which requires persevering except under very adverse circumstances. A high cost ($c = c_H$) thus constitute an allowable excuse for giving in, and it is invoked by both types when they do so.

3. *Bright line or legalistic rule.* The rule R_1, in contrast, admits no excuses: giving in is always (correctly) interpreted as a sign of weakness. This "bright line" feature has a clear benefit, in that the weak type is forced to exercise self-discipline when $c_1 = c_L$. On the other hand there is a potential cost: even in situations where a valid argument might be made for yielding "just this time" (meaning, when $c_1 = c_H$), the strong type will persevere for fear of appearing weak. The individual is "too hard on himself", but fears that by behaving otherwise he will lose self-control.

4. *Compulsive rule.* While R_1 illustrates the general tradeoff between the benefits (self-discipline) and the costs (excessive legalism) of rule-based behavior, R_3 shows how a "zero-tolerance" rule can also lead to the same costs from rigidity but without any benefits. In this case,

which corresponds well to compulsive or obsessional behavior, the strong type is bound by the fear of appearing weak, while the weak type exercises no self-restraint whatsoever.

B. Mixed strategies. As in the one-cost case, the equilibrium may also take the form of a mixture of the "pure" rules described above. That is, either type β_H randomizes between P and G in state $c_1 = c_H$, or type β_L randomizes in state $c_1 = c_L$. Denoting the mixture of R_i and R_j as R_{ij} and randomization by P/G, the four possible cases are:[31]

$R_{02} \equiv$

	β_H	β_L
c_L	P	P/G
c_H	G	G

, $R_{03} \equiv$

	β_H	β_L
c_L	P	G
c_H	P/G	G

, $R_{12} \equiv$

	β_H	β_L
c_L	P	P
c_H	P/G	G

, $R_{13} \equiv$

	β_H	β_L
c_L	P	P/G
c_H	P	G

.

4.2 Decisions and Reputational Dynamics

In the last period there are no reputational concerns, so the unique equilibrium is again R_0. At the start of period 2 the agent therefore knows that if he decides to put his will to the test and turns out to be a weak type he will get a payoff of b for sure, while if he is a strong type he will persevere in the event that $c = c_L$, resulting in an expected payoff of

$$\phi \equiv \pi(B - c_L) + (1 - \pi)b. \tag{7}$$

He again chooses W only if his self-reputation ρ_2 is above a threshold ρ_2^*, now defined by

$$\rho_2^* \; \phi + (1 - \rho_2^*) \; b \equiv \frac{a}{\gamma}. \tag{8}$$

Assumption 7 $\phi \equiv \pi(B - c_L) + (1 - \pi)b > a/\gamma > b$, *meaning that* $0 < \rho_2^* < 1$.

The probability $p_2(\rho)$ that willpower is put to the test (W is selected) in period 2 is thus still a step function at $\rho = \rho_2^*$, but with ρ_2^* now given by (8). The ex-ante value function for the high-willpower type, $V_2^H(\rho) \equiv p_2(\rho)\phi + (1 - p_2(\rho))\, a$, thus inherits the same property, and when $b > a$ so does that for the low-willpower type, $V_2^L(\rho) \equiv p_2(\rho)b + (1 - p_2(\rho))\, a$.

We now examine the self-control decision in period 1. Consider first the case where $c_1 = c_H$, which always leads to the recall state $\hat{c} = c_H$ next period (a hard task is never recalled as easy). An individual of type β_i, $i \in \{H, L\}$, therefore exerts self-control if and only if

$$\frac{c_H}{\beta_i} \leq B - b + \delta \left[V_2^i(\hat{\rho}_2^+) - V_2^i(\hat{\rho}_2^-) \right], \tag{9}$$

[31] The case $R_{01} = R_{23}$ would correspond to situations where both types use mixed strategies. This does not occur, however, because both cannot be made indifferent by the next period's single task-selection strategy.

where $\hat{\rho}_2^+$ and $\hat{\rho}_2^-$ denote posterior beliefs following the events P and G respectively, given that extenuating circumstances can plausibly be invoked ($\hat{c} = c_H$).[32]

The case where $c_1 = c_L$ leads to the recall state $\hat{c} = c_L$ (no possible excuse) with probability ν, and to $\hat{c} = c_H$ with probability $1 - \nu$. The individual therefore exerts self-control if

$$\frac{c_L}{\beta_i} \leq B - b + \delta\nu \left[V_2^i(\rho_2^+) - V_2^i(\rho_2^-)\right] + \delta(1-\nu)\left[V_2^i(\hat{\rho}_2^+)) - V_2^i(\hat{\rho}_2^-)\right], \tag{10}$$

where ρ_2^+ and ρ_2^- are the posteriors following P and G respectively, given $\hat{c} = c_L$.[33] The first two terms on the right-hand side are similar to (3) in the one-cost case, with the probability λ that a lapse is recalled now replaced by the probability ν that it cannot be rationalized away. The final term is new, and represents the more moderate loss in reputation that occurs when a case of hardship can be plausibly invoked ($\hat{c} = c_H$), but knowing that this excuse could also result from an accidental misinterpretation or a deliberate fabrication.

We shall again focus attention on equilibria satisfying monotonicity in beliefs, meaning that $\hat{\rho}_2^+ \geq \rho_1 \geq \hat{\rho}_2^-$ and $\rho_2^+ \geq \rho_1 \geq \rho_2^-$ for all ρ_1, both on and off the equilibrium path. Together with the above characterizations of $V_2^H(\rho)$ and $V_2^L(\rho)$ it implies that when $c_1 = c_L$ playing P is a dominant strategy for type β_H, and conversely when $c_1 = c_H$ playing G is a dominant strategy for type β_L, as long as Assumption 6 is complemented by:

Assumption 8 $c_H/\beta_L > B - b + \delta(b - a)$.[34]

The only possible equilibrium strategies in period 1 when confronted with the willpower activity are thus indeed the *four pure behavioral rules* described in Section 4.1, together with their four mixtures. Since all differ only through the prescribed actions in the states (β_H, c_H) and (β_L, c_L), we need only solve for the perseverance probabilities q^H for a strong-willed individual facing a high cost, and q^L for a weak-willed one facing a low cost.

[32] These are given by Bayes' rule. Let $q^i(\rho_1, c)$ be the probability with which type β_i plays P when facing a cost $c \in \{c_H, c_L\}$, given ρ_1. We impose below conditions ensuring that $q^H(\rho, c_L) = 1$ and $q^L(\rho_1, c_H) = 0$ are dominant strategies; thus:

$$\begin{aligned}
\frac{\hat{\rho}_2^+}{1-\hat{\rho}_2^+} &= \frac{\rho_1}{1-\rho_1} \cdot \frac{(1-\pi)q^H(\rho_1, c_H) + \pi(1-\nu)}{\pi(1-\nu)q^L(\rho_1, c_L)}, \\
\frac{\hat{\rho}_2^-}{1-\hat{\rho}_2^-} &= \frac{\rho_1}{1-\rho_1} \cdot \frac{(1-\pi)\left(1-q^H(\rho_1, c_H)\right) + \pi(1-\nu)}{\pi(1-\nu)\left(1-q^L(\rho_1, c_L)\right)}.
\end{aligned}$$

Both are always well-defined except in the boundary case $\nu = 1$, when $q^H(\rho_1, c_H) = 0$ (rules R_0, R_{02} and R_2). Equilibrium refinements will then have to be considered for beliefs following the zero-probability event ($\hat{\sigma} = P, \hat{c} = c_H$); see below.

[33] Taking again account of dominant strategies, we have $\rho_2^+ / \left(1 - \rho_2^+\right) = \rho/[(1-\rho)\, q^L(\rho_1, c_L)]$ and $\rho_2^- = 0$.

[34] This rules out compulsion by the weak type (β_L choosing P even when $c_1 = c_H$ due to reputational concerns). Since the insights would be exactly the same as for compulsion by the strong type, we impose the assumption to cut down on the number of cases.

4.3 Self-Discipline, Harmful Compulsiveness, and Excuses

As can be seen from equations (9)–(10), the maximum net benefits which types β_L and β_H respectively can ever expect from persevering rather than giving up in period 1 are

$$C_L \equiv B - b + \delta \max\{b - a, 0\}, \tag{11}$$

$$C_H \equiv B - b + \delta(\phi - a), \tag{12}$$

where the terms in δ correspond to the loss in reputation that playing G will bring about, if it induces a *sure* switch from W to NW next period. It is clear that:

– if $c_L/\beta_L > C_L$, caving in even when the cost is c_L is a dominant strategy for type β_L;

– if $c_H/\beta_H > C_H$, caving in when the cost is c_H is a dominant strategy for type β_H.

Let us therefore divide the $(c_L/\beta_L, c_H/\beta_H)$ plane into four regions, delimited by these two cutoffs; see Figure 3. In Region II, where (11) and (12) both hold, the *impulsive behavior* (rule R_0) is the unique equilibrium. The phenomena of interest will thus occur in the other ones, particularly Regions III and IV. For certain values of ρ there will be multiple equilibria (three at most), sustained by different, self-confirming *predictions* and retrospective *interpretations* of one's own behavior. Being agnostic about people's ability to coordinate their present and future selves on particular outcomes, we shall characterize the entire equilibrium set but also systematically identify its most efficient element, in the following sense.

Definition 1 *A behavioral rule R is (ex post) Pareto-superior to a rule R' if, when confronted with the P/G decision in period* 1, *both the strong-willed type β_H and the weak-willed type β_L are better off if R is played rather than R' (each with its continuation value in period* 2).

Although the intrapersonal signaling game now involves *four types* (two β's $\times$ two c types), we are able to solve for the equilibrium strategies and beliefs for all $\nu \in [0, 1]$. Due to to space constraints these proofs are provided in a separate Technical Appendix, while we focus here on expositing the polar cases $\nu = 1$ and $\nu = 0$, which convey the key intuitions.[35] In particular, contrasting the behaviors and welfare implications that emerge in these two cases will bring into sharper focus the cognitive foundations of personal rules.

4.3.1 Perfect Attribution ($\nu = 1$)

We first consider the case where, at date 2, the individual is always able to distinguish which cost realization occurred at date 1 –or, equivalently, to discriminate between legitimate exceptions and opportunistic rationalizations.

[35] We also focus attention on the subgame where the first-period decision node between P and G has been reached, as this is where the interesting issues arise. This test of willpower could have come about through an initial choice by the agent (requiring that ρ_1 not be too low), accidental circumstances (e.g., no alcohol or cigarettes were on hand that morning), or a constraint imposed by someone else; see the second part of footnote 17. The technical appendix is available at http://www.princeton.edu/~rbenabou/ and http://www.idei.asso.fr/.

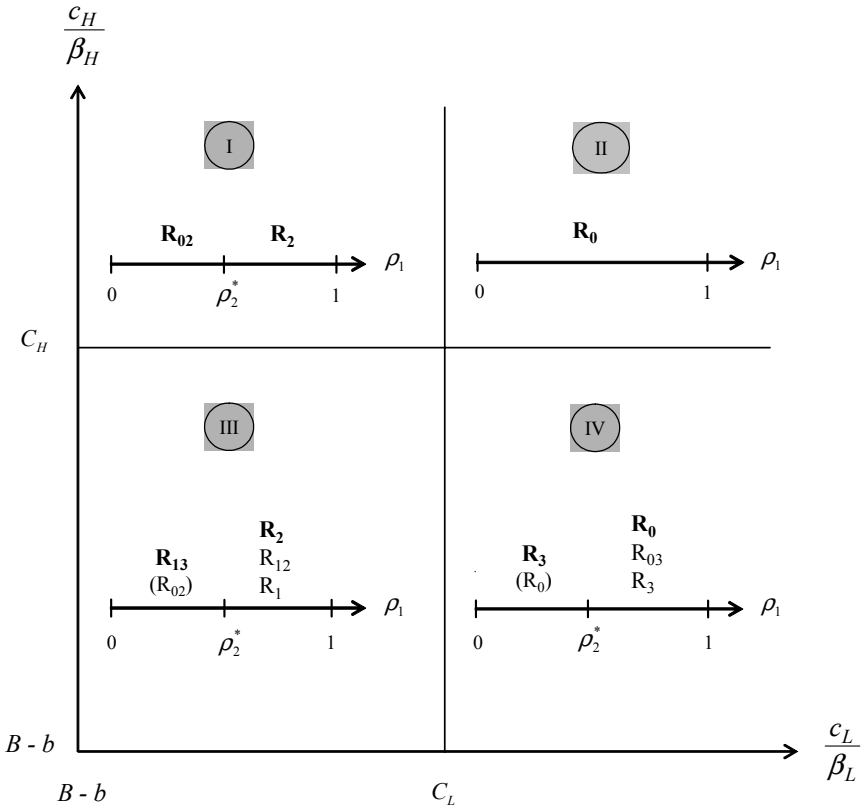

Figure 3: perfect attribution ($\nu = 1$). The equilibrium in bold is Pareto-dominant (for $b \geq a$); those in parentheses fail to satisfy the Cho-Kreps (1987) criterion. Notation: $R_0 \equiv (q^H = q^L = 0)$, "impulsive"; $R_1 \equiv (q^H = q^L = 1)$, "bright-line"; $R_2 \equiv (q^H = 0,\ q^L = 1)$, "flexible"; $R_3 \equiv (q^H = 1,\ q^L = 0)$, "compulsive"; $R_{ij} \equiv$ mixing between R_i and R_j.

Proposition 2 *For $\nu = 1$, the set of equilibria in period 1 is described by Figure 3. Multiple equilibria may arise, in which case the Pareto-dominant one is indicated in bold and those that are not robust to Cho and Kreps' (1987) intuitive criterion in parentheses.*

In all that follows we shall abstract from the non-robust equilibria.[36] The above results demonstrate five important phenomena.

1. Self-control. For $c_H/\beta_H > C_H$, both types give in when $c_1 = c_H$. Since this state can be perfectly distinguished from that where $c_1 = c_L$, the problem in Regions I and II reduces to the one-cost case, $c = c_L$. Thus the strong type β_H always perseveres, while in Region I the weak type β_L will *pool* with the strong, with the same probability $q^L \leq 1$ as in Proposition 1 (for $\lambda = 1$).[37]

[36] In addition to being ruled out by the intuitive criterion, they are largely artefacts of the boundary case $\nu = 1$ (there are no zero-probability events for any $\nu < 1$; see footnote 32). Thus, as shown in the Technical Appendix, the one in Region IV is never part of limiting equilibrium set obtained as $\nu \to 1^-$, and the one in Region III is only part of that set under additional parameter restrictions (which correspond to Subregion III$^+$ on Figure IV).

[37] As explained earlier, pooling is valuable for a weak type (Regions I and III are non-empty) only when $b > a$. See footnote 25 for a simple way to relax this assumption, however.

Combining behavior across the two cost states, the individual follows the flexible rule R_2 or R_{02}, depending on whether $\rho_1 \gtrless \rho_2^*$.

2. Compulsiveness. Region IV displays a novel type of behavior. Here the weak-willed individual always caves in, so the issue is whether the strong-willed one will persevere even in the high-cost state c_H in order to *separate* from the weak. Depending on his choice the equilibrium will be R_3, R_0 or their mixture R_{03}. Specifically, for low values of ρ_1 the individual does not give himself the "benefit of the doubt" even when $\hat{c} = c_H$, so the β_H type can never afford to ease up ($q^H = 1$). For $\rho_1 > \rho_2^*$, on the other hand, R_0 is an alternative, more "forgiving" equilibrium ($q^H = 0$), as explained below in more detail.

3. Flexibility versus rigidity. Behavior in Region III entails a tradeoff between the costs and benefits of self-regulation. For an individual with high enough confidence in his power of will the flexible, excuse-contingent rule R_2 can be sustained (as in Region II). For one who is initially doubtful of his fortitude, on the other hand, compulsiveness will be the price to pay for securing some of the benefits of self-restraint (rule R_{13}).

4. When are excuses admissible? For someone with low initial self-confidence, the only (robust) equilibrium when $c_H < C_H$ is one that requires him to engage in costly self-reassurance (R_3 in Region IV, R_{13} in Region III). For a sufficiently confident person, by contrast, compulsiveness is only *one of* several (robust) equilibria, and may thus be avoided. Indeed when Self 2 "takes no excuses", interpreting even a "minor" lapse (G when $\hat{c} = c_H$) as revealing that $\beta = \beta_L$, the strong type has no choice but to obey this zero-tolerance rule. With the weak type a lapse will occur, followed by a complete collapse of self-restraint (e.g., binging) next period.[38] Suppose, however, that Self 2 takes a more forgiving interpretation of lapses for which excuses exist ($\hat{c} = c_H$), in that he expects the strong type to give in whenever $c_1 = c_H$ (rule R_0). A minor lapse is then not as bad news, so if the initial reputation ρ_1 was high enough the high-willpower type will indeed be able to "relax" when faced with $c_1 = c_H$, without jeopardizing future self-restraint.[39]

5. The cost of compulsion. From an ex-post point of view, the β_H type with $\rho_1 > \rho_2^*$ is always better off when given the benefit of the doubt (rule R_0) than when challenged to prove himself (rule R_3). The weak type β_L is also better off provided $b > a$, so in that case (and for $\rho_1 > \rho_2^*$) R_0 Pareto-dominates R_3 and R_{03} in Region IV; otherwise, there is no dominant equilibrium. Most interestingly, persevering in spite of high costs can even be *suboptimal ex ante.* This case, which occurs when $c_H > B - b$ (notice that β_H does not appear) corresponds best to compulsive or "obsessive" behaviors such as those of the miser, the workaholic or the anorexic: the individual is so afraid of appearing weak to himself that he chooses a degree of self-restraint that is out of proportion with the benefits, and results in lower welfare.[40] To an outside observer, he may *appear*

[38] Baumeister at al. (1994) point out that zero-tolerance beliefs such as those underlying "Just Say No" campaigns (against drugs, premarital sex, etc.) often severely backfire, as even minor transgressions of the rule can lead to large collapses of self-esteem and self-regulation. This is referred to as "lapse-activated snowballing".

[39] The W decision is always the one he prefers, since $\phi > a$ (whether $b \gtrless a$).

[40] From here on, when speaking of "compulsiveness" we shall be referring to the case $c_H > B - b$.

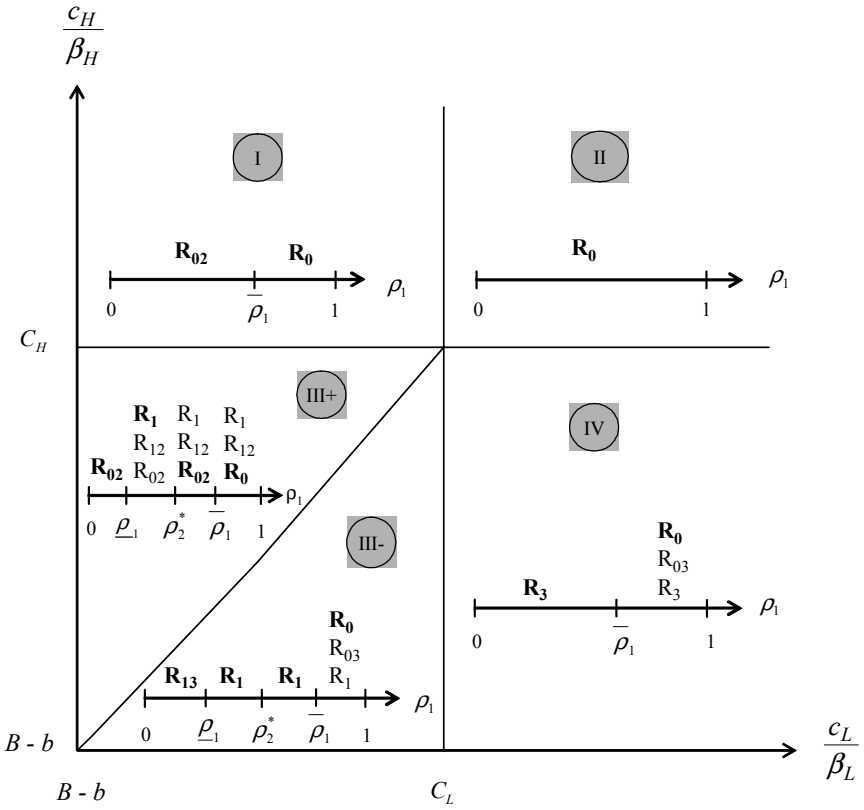

Figure 4: unreliable attribution ($\nu = 0$). The equilibrium in bold is Pareto-dominant (for $b \geq a$). Notation: $R_0 \equiv (q^H = q^L = 0)$, "impulsive"; $R_1 \equiv (q^H = q^L = 1)$, "bright-line"; $R_2 \equiv (q^H = 0,\ q^L = 1)$, "flexible"; $R_3 \equiv (q^H = 1,\ q^L = 0)$, "compulsive"; $R_{ij} \equiv$ mixing between R_i and R_j.

to be acting *as if* his β was greater than 1; in reality, it is precisely the fear that β might be too low that gives rise to these compulsions. As Baumeister et al. (1994, p. 85-86) write, "*Obsessions and compulsions are attempts to compensate for some self-regulatory deficit... The quest for such structure (boundaries, limits, time markers, and the like) and the excessive adherence to such structure, which have been commonly observed among these individuals, may be a response to the inner sense that they cannot control themselves without those externals aids.*"

4.3.2 Unreliable Attribution ($\nu = 0$)

We now turn to the case where, at date 2, the individual is unable to distinguish which cost realization occurred at date 1 –or, equivalently, where he is always able to come up with excuses or ex-post rationalizations ("plausible deniability"):

Proposition 3 *For $\nu = 0$, the set of equilibria in period 1 is described by Figure 4. Multiple equilibria may arise for $\rho_1 > \underline{\rho}_1$, in which case the Pareto-dominant equilibrium is indicated in bold. All satisfy Cho and Kreps' (1987) intuitive criterion.*

The two new thresholds which appear on Figure 4 are defined as follows:

$$\bar{\rho}_1 \equiv \frac{\rho_2^*}{\rho_2^* + (1-\pi)(1-\rho_2^*)} > \rho_2^* > \frac{\rho_2^*}{\rho_2^* + (1-\rho_2^*)/\pi} \equiv \underline{\rho}_1. \tag{13}$$

The higher one is such that for $\rho_1 > \bar{\rho}_1$, recalling a lapse with $\hat{c} = c_H$ will never bring reputation below ρ_2^*, if it is expected that the strong-willed type always gives in when $c_1 = c_H$. The lower one is such that for $\rho_1 < \underline{\rho}_1$ recalling perseverance even with $\hat{c} = c_H$ will never bring reputation above ρ_2^*, if it is expected that the weak-willed type always perseveres when $c_1 = c_L$.

Comparing Figure 4 ($\nu = 0$) with Figure 3 ($\nu = 1$) yields a number of interesting insights on the cognitive underpinnings of self-restraint and compulsiveness.

1. Loss of flexibility. One of the main results is that the flexible rule R_2 is no longer an equilibrium where it used to be self-enforcing ($\rho_1 > \rho_2^*$ in Regions I and III). The only rules that can be sustained in its place entail either a *loss of self control,* partial or total, in the low-cost state $c_L : R_{02}, R_0$; or *compulsiveness,* partial or total, in the high-cost case $c_H : R_1, R_{12}$; or *both* at the same time (R_{03}, R_{13}).

2. Differential responses to ambiguity. Region I shows that for a weak-willed type, the fall in ν leads to a general *loss* of self-control. As seen on Figure 5a it is more drastic the higher is ρ_1, so that q^L eventually decreases toward zero as initial reputation improves. Intuitively, when the self-monitoring technology is imperfect the "principal" (Self 2) is more receptive to claims of hardship (leaving the weak type more leeway for misrepresentation) when they come from a more trusted "agent" (Self 1). For a strong-willed type, by contrast, Region IV and Figure 5b show that less reliable attribution leads to *increased compulsiveness.* Indeed, a low ν makes recalling G and $\hat{c} = c_H$ worse news about β than when $\nu = 1$, because there is now a chance that the true c_1 was actually c_L. Distrusting excuses, the individual is then tougher on himself, demanding more proof that willpower is high. The prediction that a greater availability of "easy" excuses tends to undermine self-control for weak-willed individuals but reinforce compulsiveness for strong-willed ones is novel, and empirically testable.

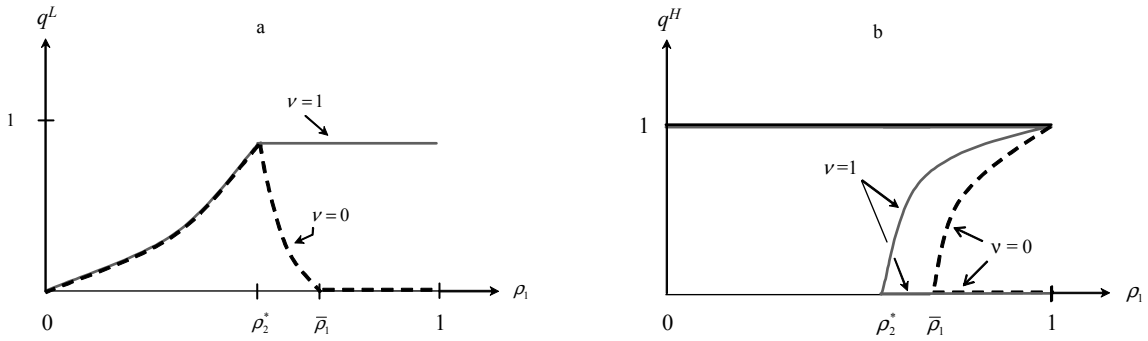

Figure 5: cognition and self-regulation. Panel (a): self-control by the weak-willed . ($c_1 = c_L$, Region I). Panel (b): compulsion by the strong-willed ($c_1 = c_H$, Region IV).

3. The benefits of forgetting. Another interesting result is that the individual may be better off if he could *forget his lapses* altogether, especially if he knows himself to be too adept at ex-post rationalizations. Indeed, with $\lambda = 0$ the only equilibrium is always R_0 (impulsiveness) which (for $c_H > B - b$ and $b > a$) is always better than R_3 (compulsiveness), both ex ante and ex post. Thus, a person suffering from anorexia would typically be better off if only she could "forget" about controlling her weight, and accept being heavier than is optimal.

5 Cognitive Rules and Resolutions

A person cannot just select a good behavioral rule: for given cognitive parameters (λ, ν), only some are self-enforcing. Most personal rules in practice therefore involve the simultaneous choice of *behavioral* and *information-processing* principles. Simple examples of the latter include keeping a daily record of one's successes and failures in dealing with impulsive tendencies, weighing oneself every day, and using separate envelopes or similar devices to supplement mental accounting. A more subtle case is that of resolutions, which consist in setting for oneself ex-ante desirable mandates in ways that affect the self-monitoring process, such as taking a solemn oath on a memorable occasion.

5.1 Endogenous Self-Monitoring

As noted earlier, psychologists observe that people who fail at self-regulation are often those with poor self-monitoring skills. Conversely, much of behavioral therapy involves endowing patients with a more reliable, less manipulable cognitive "technology". Self-help groups and church confessions are also, in large part, self-monitoring devices.[41] While it is clear from our results how improving the recall of actions (λ) or their attribution (ν) can be beneficial ex-ante, the harder question is why a lapsed individual would ever provide confirmation of his weakness ex post. Yet a cognitive rule such as keeping a journal or using separate accounts for luxuries and necessities can, like any other rule, be effective only to the extent that it is self-enforcing.

We analyze this issue with a simple extension of the model in Section 3, where recall is endogenized. At the end of period 1 ("in the evening"), the agent can record in a journal an account of his day claiming that P occurred, or that G occurred; he can also not enter anything. Entering a truthful account involves a cost ϵ^T, while coming up with a believable falsification costs ϵ^F; leaving a blank or uninformative page costs 0.[42] At the start of period 2 (morning of the second day), if he had caved in to temptation there is again a probability λ that he will receive

[41] Neither his peer group nor the priest can (or even try) to force the alcoholic or sinner to refrain from his vice –in contrast to, say, a rehabilitation clinic. Whatever else they accomplish, the direct and inevitable effect of such sessions is that the individual is *forced to think about his own behavior,* even if he chooses to misrepresent it.

[42] These costs depend only on the individual's action and narration of it, and not on his degree of willpower. Presumably $\epsilon^F \geq \epsilon^T$, since a lying individual would have to elaborate a credible scenario, check it for inconsistencies, etc., as opposed to simply relating what happened. Our results will only require a weaker condition, however.

a piece of "hard information" (salient memory, feedback from others, physical evidence) showing that he did choose G; if he chose P no such signal will be received. In addition, the agent may access his journal and read (or simply recall) its contents; this occurs with a probability μ which, for simplicity, we take to be exogenous.

We shall now derive conditions ensuring that the following *cognitive and behavioral* strategies jointly constitute a PBE of the enriched game:

i) an agent who has chosen P makes a record of it, whether his type is β_H or β_L; one who has chosen G leaves a blank page, or puts no real effort into fabricating a misleading account;

ii) in period 2, if no evidence that G was chosen is received –either directly or in the form of a self-incriminating (G) or missing journal entry– the agent randomizes between W and NW with a probability $p_2^* \in [0,1]$. He chooses W with probability 1 when faced with an entry of P, and with probability 0 when faced with a record of G or a missing entry.

iii) in period 1, the choice between P and G is the same as in Proposition 1, except that λ is replaced by $\lambda + (1-\lambda)\mu$.

Indeed, suppose that (i) and (ii) above are believed to be true. A blank page is then a sure sign that one has caved in, while an entry of P is fully credible. A weak-willed agent who persevered will then prefer to document his performance, provided that:

$$\frac{\epsilon^T}{\beta_L} < \delta\mu(b-a). \tag{14}$$

We shall assume that (14) holds; a strong type who persevered will then also make a record it, as $\epsilon^T/\beta_H < \delta\mu(B-c-a)$. Let us now verify that a weak-willed type who chose G prefers not to write anything. If he wrote down G he would incur the cost ϵ^T and perfectly reveal himself, so there is no point in doing so. Recording nothing is also better than misleadingly recording P, provided that

$$\frac{\epsilon^F}{\beta_L} > \delta(1-\lambda)\mu(b-a). \tag{15}$$

The term $(1-\lambda)\mu$ reflects the fact that falsification pays off only in the joint event where no outside hard evidence is received and the journal's contents are accessed. For the two incentive-compatibility conditions to be mutually consistent, ensuring the optimality of (i), requires only that $\epsilon^T < \epsilon^F/(1-\lambda)$.

We next turn to the ways in which the behavioral rule (self-control) is modified by the presence of the cognitive rule (record-keeping), and verify (ii) and (iii). Since perseverance at $t=1$ leads to en entry of P while a lapse leads to an incriminating blank page, the individual's posterior ρ_2 at $t=2$ is equal to 1 in the first case, and to the same function of the self-control probability q_1 as in (6) in the second, except that the probability that a lapse goes undetected is now $(1-\lambda)(1-\mu)$ instead of $1-\lambda$. Therefore the equilibrium value of q_1, which equates ρ_2 to ρ_2^*, is higher. The only remaining equilibrium requirement is that p_2^* make the weak-willed type indifferent between

persevering and giving in to temptation. Thus

$$\frac{c+\epsilon^T}{\beta_L} = B - b + \delta\left[\mu + (1-\mu)\lambda p_2^*\right](b-a), \tag{16}$$

which uniquely defines p_2^* for appropriate parameter values. When $\epsilon^T = \mu = 0$ this reduces to the earlier condition; otherwise, the probability that perseverance today will be rewarded by a willpower decision tomorrow (namely, $\mu + (1-\mu)\lambda p_2^*$) is simply adjusted upward to compensate for the higher effective cost of perseverance, $c + \epsilon^T$.

Proposition 4 *Let $c_H = c_L = c$ and $b > a$. Suppose also that $\lambda < 1$, but the agent has access to a record-keeping technology satisfying (14)–(16). The cognitive and behavioral strategies described in (i)–(iii) above then constitute a PBE. The agent makes a record when, and only when, he has persevered. Moreover, he exercises self-control with a higher probability than in the absence of record-keeping, both in the first period (P) and in the second (W).*

These results show how the joint adoption of a cognitive and a behavioral rule can be beneficial, and how the two are mutually sustaining. Finally, stepping back to the initial stage, the individual will invest in a journal, separate accounts, etc., if the required setup cost k is small enough relative to the welfare gains from self-restraint described in Proposition 4.[43]

5.2 Why Do Resolutions Matter?

A resolution (for instance a financial plan) consists in the simultaneous adoption of an explicit behavioral rule ("from now on, I will save at least $500 each month") and a perhaps more implicit cognitive rule that helps sustain it ("I will check my savings account statement and credit card balances every month, and compute the difference"). Resolutions, vows, oaths and the like may be very specific ("I never eat dessert," "I finish every paper that I start"), or define a broader mandate ("I am on a diet", "I stick to my choices"); some, like religious principles, even have a universal character. All, however, ultimately rest on a commitment to regularly scrutinize one's behavior, and can thus be understood using a formal analogy with journal-keeping.

Suppose thus that every evening, the individual can choose to pause and reflect on whether he followed his resolution that day, or just put the thought out of his mind. If he does examine his behavior he may either acknowledge things the way they happened (at a cost of ϵ^T), or make efforts at denial and self-justification (at a cost of ϵ^F). With probability μ his resolution will come back to awareness the next day, triggering the recall of his previous days' introspection, or

[43] Proposition 4 focusses on the case where the individual is always concerned about maintaining a sufficiently high degree of optimism into the future ($b > a$). This is indeed when it is hardest to enforce the cognitive rule ex-post, because he would then rather forget his lapses. When $b < a$, on the contrary, he is more concerned about avoiding overconfidence (embarking on projects doomed to failure), and therefore quite willing to record lapses and other ego-damaging signals, as a safeguard against his imperfect or selective memory.

lack thereof. In addition, with probability λ he will be reminded of previous lapses through other channels. The implications of this model are clearly similar to those derived above. The cognitive rule (reflecting upon one's behavior every evening) is similar to journal-keeping in that it creates another, "redundant" channel of activation of lapse-related memories.[44]

In addition to affecting cognitive decisions ex-post, resolutions can also operate ex-ante, by making certain types of actions particularly salient. This is why they are often adopted after a memorable event (close call with tragedy, birthday, new year), or in formats and environments that increase the likelihood (μ) that they will come back to awareness later on (solemn oath, religious ceremony). This is analogous to investing in a better, longer-lasting journal, with a tradeoff between a higher μ and a higher initial cost $k(\mu)$: more extreme initial event, more solemn oath or demanding ritual, etc.

The same results showing how cognitive strategies can help a person more effectively monitor his actions (raise λ) clearly also apply to the understanding of their underlying motives (determination of ν), through resolutions that embody an "examen of conscience" requirement to regularly scrutinize the true nature of one's behavior and desires (did I face c_H or c_L?). When the conditions in Proposition 4 do not hold, on the other hand, the cognitive part of a resolution can be a prescription to engage in *less* information-processing. Thus, if by searching for plausible (non-falsifiable) excuses it is relatively easy to come up with one even in the absence of real hardship, a rule to "never look for excuses" can be both ex-ante optimal and self-enforcing ex-post, as a pooling equilibrium where recalling that one spent any time thinking about justifications would be salient, and interpreted as a sign that there really was none.[45]

A final observation stems from the fact that a lapse in one activity can easily spill over to other ones, because of its impact on the individual's belief about his strength or weakness of will, which is relevant for a host of decisions.[46] This linkage operates like an increase in the general concern for self-reputation (a higher δ), and is therefore conducive to both self-control and compulsion. For instance, a resolution to "always stick to one's choices" will make any lack of perseverance more memorable. The self-control benefits argue in favor of universality, while the compulsion costs call for viewing activities as independent, and self-indulgence as confined to narrow "lapse districts"[47] where the individual confesses to being "hopeless" (e.g., eating, exercising) in order to avoid the propagation of self-doubt to other dimensions of his life (e.g., working).

[44] See Anderson (2000, p. 212) for a discussion of redundancy in memory processes.

[45] For instance, let natural memory be imperfect but unbiased: in the absence of particular effort or rehearsal, $\nu_H \equiv \Pr[\hat{c} = c_H \mid c_1 = c_H] = \bar{\nu} = \Pr[\hat{c} = c_L \mid c_1 = c_L] \equiv \nu_L$, where $1/2 < \bar{\nu} < 1$. By spending time poring over the situation, someone who faces $c_1 = c_H$ can increase ν_H above $\bar{\nu}$; but by engaging in opportunistic rationalization, someone for whom $c_1 = c_L$ may be able to decrease ν_L even more.

[46] Baumeister et al. (1994, p. 11) state that *"self-regulatory capacity is a central, powerful, stable, and beneficial aspect of personality,"* then describe abundant research which has shown that children's capacity to delay gratification is significantly correlated with many important outcomes and attitudes later in life, such as resourcefulness, cooperativeness, ability to deal with stress, etc.

[47] The term is borrowed from Ainslie (1999), who uses it by analogy with the "vice districts" tolerated in most cities.

6 Conclusion

This paper has shown how people may achieve self-control through the adoption of personal rules, and identified the costs and benefits of such self-regulation. Our theory is based on the idea of self-reputation over one's willpower as the mechanism that transforms lapses in a personal rule into precedents that undermine future self-restraint. The foundation of any self-reputational mechanism, in turn, was shown to be the imperfect recall of past motives and feelings (intensity of temptation, cravings) which leads people to monitor, and infer revealed preferences from, their own past actions.

These ideas also offer several interesting directions for further research, on both individual behavior and social interactions. The first one stems from our extension of personal rules to the cognitive realm, where we studied certain prescriptions bearing on the processing of information such as record-keeping or simple forms of mental budgeting. It should be possible to extend this framework to formalize the workings of a broader set of mental accounts (Thaler (1980), (1985)) and other self-imposed mandates on thought processes. Indeed all raise the same issues of internal enforcement, self-monitoring and excuse-making as do more explicitly behavioral resolutions relating to, say, smoking or saving.

A second set of questions arises from the observation that self-control decisions are often influenced by the examples set by others. These informational spillovers can be beneficial, as with self-help groups like Alcoholic Anonymous, or detrimental, as with peer interactions among youths that often aggravate impulsive tendencies towards the procrastination of effort, drinking and other forms of substance abuse. The issue of peer effects in self-control is pursued in Battaglini et al. (2001), by embedding the present model in a context of social learning by agents with correlated costs (or benefits) of self-restraint.

Finally, the central idea that people have only imperfect access to their own preferences and motives, and must therefore infer them from their past decisions, can provide the foundation for a theory of personal, professional or sociocultural identity as a cognitive investment.

Appendix

Proof of Proposition 1. Consider first the weak type's probability of perseverance at date 1.

Pooling: $q_1 = 1$. Then $\rho_2^+ = \rho_1$, while ρ_2^- can be any $\rho' \leq \rho$. Optimality in (3) then requires $\rho_1 \geq \rho_2^* > \rho'$, otherwise the right-hand side would be zero. Let therefore $\rho_1 > \rho_2^*$ (leaving aside the measure-zero case where $\rho_1 = \rho_2^*$). Given that $c/\beta_L < C(\lambda)$, this is indeed an equilibrium.

Semi-separation: $q_1 \in (0,1)$. This implies $\rho_2^+ \in (\rho_1, 1)$ and $\rho_2^- = 0$. Furthermore, (3) must now hold with equality, $c/\beta_L = B - b + \delta\lambda \left[V_2^L(\rho_2^+) - a\right]$. This can only occur if

$$\rho_2^+ \equiv \frac{\rho_1}{\rho_1 + (1-\rho_1)(q_1 + (1-q_1)(1-\lambda))} = \rho_2^*, \tag{A.1}$$

requiring $\tilde{\rho}_1(\lambda) < \rho_1 < \rho_2^*$, and if the mixing probability $p_2^* \equiv p_2(\rho_2^*)$ in period 2 satisfies $c/\beta_L = B - b + \delta\lambda p_2^*(b-a)$. This condition and the one above uniquely determine q_1 and p_2^* in $[0,1]$, as given in Proposition 1.

Separation: $q_1 = 0$. This implies again that $\rho_2^- = 0$, and thus one must have $c/\beta_L \geq B - c + \delta\left[V_2^L(\rho_2^+) - a\right] = V_2^L(\rho_2^+) - a$. With $c/\beta_L < C(\lambda)$ this can only happen for $\rho_2^+ < \rho_2^*$, which means that $\rho_1 < \tilde{\rho}_1(\lambda)$.

Finally, we turn to the individual's task selection in period 1. For $\rho_1 \geq \rho_2^*$ both types choose P with probability 1, so it is optimal to select W. Indeed, this yields $B - c$ in period 1 and $\delta\left[\rho_1(B-c) + (1-\rho_1)b\right]$ in period 2, against a/γ in period 1 and the same expected payoff in period 2 if NW is chosen instead (there is then no new information, so $\rho_2 = \rho_1$ and W is chosen in period 2). Consider now the case where $\tilde{\rho}_1(\lambda) < \rho_1 < \rho_2^*$. Choosing W rather than NW then leads to expected net gains of Δ_1 in period 1 and Δ_2 in period 2, where:

$$\Delta_1 \equiv \rho_1 \left(B - c - a/\gamma\right) + (1-\rho_1)\left[q_1\left(B-c\right) + (1-q_1)\, b - a/\gamma\right] \tag{A.2}$$

is increasing in ρ_1, both directly and through q_1, and the same is true for

$$\begin{aligned}
\Delta_2/\delta \;&\equiv\; \rho_1\left[p_2^*(B-c) + (1-p_2^*)a\right] + \\
&\quad (1-\rho_1)\left\{\left[q_1 + (1-q_1)(1-\lambda)\right)\right]\left[p_2^* b + (1-p_2^*)a\right] + (1-q_1)\lambda a\right\} - a \\
&=\; p_2^*\left\{\rho_1(B-c-a) + (1-\rho_1)\left[q_1 + (1-q_1)(1-\lambda)\right](b-a)\right\}.
\end{aligned} \tag{A.3}$$

By continuity, the total gain $\Delta_1 + \Delta_2$ is positive just below $\rho_1 = \rho_2^*$. Therefore, the choice between W and NW in period 1 is indeed governed by a cutoff $\rho_1^* < \rho_2^*$. It is ambiguous, on the other hand, whether ρ_1^* is greater or smaller than the threshold $\rho_1 = \tilde{\rho}_1(\lambda)$ where $q_1 = 0$. ■

Proof of Propositions 2 and 3. See the paper's separate Technical Appendix, posted at http://www.princeton.edu/~rbenabou/ and http://www.idei.asso.fr/.

References

Ainslie, George. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person* (Studies in Rationality and Social Change). Cambridge, England: Cambridge University Press, 1992.

—— *Breakdown of Will.* Cambridge, England: Cambridge University Press, 2001.

Ameriks, John, Caplin, Andrew; and Leahy, John J. "Wealth Accumulation and the Propensity to Plan." *Quarterly Journal of Economics* 118 (June 2003): 1007–1048.

Anderson, John R. *Cognitive Psychology and its Implications.* 5th edition. New York: Freeman and Co., 2000.

Ariely, Dan, and Wertenbroch, Klaus. "Procrastination, Deadlines and Performance: Self-Control by Precommitment." *Psychological Science* 13 (May 2002): 219–224.

Baumeister, Roy; Heatherton, Todd; and Tice, Dianne M. *Losing Control: How and Why People Fail at Self-Regulation.* San Diego, CA: Academic Press, 1994.

Battaglini, Marco; Bénabou, Roland; and Tirole, Jean. "Self-Control in Peer Groups." CEPR Discussion Paper 3149, January 2001.

Becker, Gary, and Murphy, Kevin. "A Theory of Rational Addiction." *Journal of Political Economy* 96 (August 1988): 675–700.

Bem, Daryl J. "Self-Perception Theory," in L. Berkowitz ed., *Advances in Experimental Social Psychology.* New York, NY: Academic Press, 1972.

Bénabou, Roland, and Tirole, Jean. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (August 2002): 871–915.

—— "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (July 2003): 489–520.

Bodner, Ronit, and Prelec, Drazen. "The Diagnostic Value of Actions in a Self-Signaling Model." MIT mimeo, 1997.

—— "A Neo-Calvinist Model of Conscience," in I. Brocas and J. Carrillo eds. *Collected Essays in Psychology and Economics.* Oxford University Press, forthcoming.

Caillaud, Bernard; Cohen, Daniel; and Jullien, Bruno. "Toward a Theory of Self-Restraint." CERAS mimeo, June 1999.

Carrillo, Juan "To Be Consumed With Moderation." (2004) *European Economic Review,* in press.

Carrillo, Juan, and Mariotti, Thomas. "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies* 67 (July 2000): 529–544.

Cho, In-Koo, and Kreps, David (1987) "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics,* 102: 179–221.

Farber, Henry. "Is Tomorrow Another Day? The Labor Supply of New York City Cab Drivers." NBER Working Paper No. 9706, May 2003.

Heath, Chip, and Soll, Jack B. "Mental Budgeting and Consumer Decisions." *Journal of Consumer Research* 23 (June 1996): 40–52.

Hirschleifer, David, and Welch, Ivo. "An Economic Approach to the Psychology of Change: Amnesia, Inertia and Impulsiveness." *Journal of Economics and Management Strategy* 11 (Fall 2002): 379–421.

Kahneman, Daniel, and Wakker, Peter P. "Back to Bentham? Explorations of Experienced Utility." *Quarterly Journal of Economics* 112 (May 1997): 375–407.

Kirby, Kris N., and Guastello, Barbarose. "Making Choices in Anticipation of Similar Future Choices Can Increase Self-Control." *Journal of Experimental Psychology: Applied* 7 (June 2001): 154–164.

Kivetz, Ran, and Simonson, Itamar. "Self-Control for the Righteous: Toward a Theory of Precommitment to Indulgence." *Journal of Consumer Research* 29 (September 2002): 199–217.

Laibson, David. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (May 1997): 443–478.

Laibson, David. "*Essays in Hyperbolic Discounting.*" MIT Ph.D. dissertation, 1994.

Lepper, Mark R., and Greene, David. "Overjustification Research and Beyond: Toward a Means-Ends Analysis of Intrinsic and Extrinsic Motivation" in *The Hidden Cost of Reward: New Perspectives on the Psychology of Human Motivation.* New York: John Wiley, 1978.

Loewenstein, George. "Out of Control: Visceral Influences on Behaviors." *Organizational Behavior and Human Decision Processes* 65 (May 1996): 272–292.

—— "A Visceral Account of Addiction" in J. Elster and O.J. Skog, eds. *Getting Hooked: Rationality and Addiction.* Cambridge: Cambridge University Press, 1999.

Loewenstein, George, and Schkade, David. "Wouldn't It Be Nice? Predicting Future Feelings" in D. Kahneman, E. Diener and N. Schwartz, eds. *Well-Being: Foundations of Hedonic Psychology.* New York, NY: Russel Sage Foundation, 1999.

Meardon, Stephen J., and Ortmann, Andreas. "Self-Command in Adam Smith's *Theory of Moral Sentiments:* A Game-Theoretic Reinterpretation." *Rationality and Society* 8 (February 1996): 57–80.

Mischel, Walter; Ebbesen, Ebbe; and Zeiss, A. Raskoff. "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Psychology* 21 (1972): 204–218.

Mulligan, Casey. "A Logical Economist's Argument Against Hyperbolic Discounting." University of Chicago mimeo, February 1996.

O'Donoghue, Ted, and Rabin, Matthew. "Doing it Now or Later." *American Economic Review* 89 (March 1999): 103–124.

Orphanides, Athanasios, and Zervos, David. "Rational Addiction with Learning and Regret." *Journal of Political Economy*, 103 (August 1995): 739–758.

Palacios-Huerta, Ignacio. "Time Inconsistent Discounting in Adam Smith and David Hume." *History of Political Economy* 35 (Summer 2003), 241-268.

Prelec, Drazen, and Loewenstein, George. "The Red and the Black: Mental Accounting of Savings and Debt." *Marketing Science* 17 (1998): 4–28.

Puri, Radhika. "Measuring and Modifying Consumer Impulsiveness: A Cost-Benefit Accessibility Framework." *Journal of Consumer Psychology* 5 (1996): 87–113.

Quattrone George A., and Tversky, Amos. "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion." *Journal of Personality and Social Psychology* 46 (1984): 237–248.

Rabin, Matthew. "Moral Preferences, Moral Constraints, and Self-Serving Biases." U.C. Berkeley Working Paper in Economics 95/241, (August 1995).

Rachlin, Howard. *The Science of Self-Control.* Cambridge, MA: Harvard University Press, 2000.

Schelling, Thomas. "Self-Command in Practice, in Policy and in a Theory of Rational Choice." *American Economic Review, Papers and Proceedings* 74 (1984): 1–11.

Seligman, Martin E. P. *"Helplessness: On Depression, Development, and Death."* San Francisco, CA: Freeman and Co., 1975.

Smith, Adam. *The Theory of Moral Sentiments.* Edinburgh, Scotland: J. Hay and Co. (1813), 1759.

Soman, Dilip. "Effects of Payment Mechanisms on Spending Behavior: The Role of Rehearsal and Immediacy of Payments." *Journal of Consumer Research* 27 (March 2001): 460–474.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies* 23 (1956): 165–180.

Thaler, Richard. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behavior and Organization* 1 (1980): 39–60.

—— "Mental Accounting and Consumer Choice." *Marketing Science* 4 (1985): 199–241.

Thaler, Richard, and Shefrin, Hersch M. "An Economic Theory of Self Control." *Journal of Political Economy* 89 (April 1981): 392–406.

Wertenbroch, Klaus. "Consumption Self-Control by Rationing Purchase Quantities of Virtues and Vice." *Marketing Science* 17 (1998): 317–337.

Wertenbroch, Klaus; Soman, Dilip; and Nunes, Joe. "Debt Aversion and Self-Control: Consumer Self-Management of Liquidity Constraints." INSEAD mimeo, 2002.

Zelizer, Vivian. *The Social Meaning of Money.* Princeton, NJ: Princeton University Press, 1997.