Willpower and Personal Rules

Roland Bénabou\textsuperscript{1} and Jean Tirole\textsuperscript{2}

First draft May 2000

This version: October 2001\textsuperscript{3}

(in progress)

\textsuperscript{1}Princeton University, CEPR, NBER, and IRP (rbenabou@princeton.edu).

\textsuperscript{2}IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT (tirole@cict.fr).

\textsuperscript{3}We are grateful for helpful comments to George Ainslie, Marco Battaglini, George Loewenstein, Danny Kahneman, Botond Kőszegi, Drazen Prelec, Eric Rasmusen, Stefano Della Vigna, as well as participants at the ECARES–CEPR conference on Psychology and Economics (Brussels, June 2000), the NBER Conference on Macroeconomics and Individual Decision Making (Cambridge, November 2000), and seminars at Chicago, Columbia, DELTA-Paris, Harvard, MIT, Princeton, Rochester, and Toulouse. Bénabou gratefully acknowledges financial support from the National Science Foundation (SES-0096431).
Abstract

Much of the literature on time inconsistency has studied the external commitment devices that individuals use to address their self-control problems: tying oneself to the mast, staying away from temptation, holding illiquid assets, or “asking for controls” from others. This paper, by contrast, focuses on internal commitment mechanisms or “personal rules” (diets, exercise regimens, resolutions, moral or religious precepts, etc.) through which people attempt to achieve self-discipline.

The basic idea, which builds on Ainslie (1992), is that rules cause lapses to be interpreted as precedents, resulting in a loss of self-reputation which has an adverse impact on future self-control. We thus model the behavior of individuals who are unsure of their willpower (ability to delay gratification) in certain states of the world, and characterize rules as (Markovian) self-reputational equilibria where impulses for immediate gratification are held in check by the fear of “losing faith in oneself” –which would lead to a further collapse of self-discipline. We then examine how equilibrium conduct reflects the extent to which the individual’s self-monitoring is subject to opportunistic distortions of memory or inference, such as finding excuses for one’s past behavior.

We also show that people will sometimes adopt excessively rigid or “legalistic” rules that result in compulsive behavior such as miserliness, workaholism, or anorexia. These can be understood as costly forms of self-signalling where, in equilibrium, 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 when self-restraint is not even desirable ex-ante. Such common instances where individuals appear to display a “salience of the future” are thus not only consistent, but actually generated by (a concern over) present-oriented preferences.

Keywords: psychology and economics, behavioral economics, willpower, motivation, time inconsistency, self control.

JEL Classification: A12, C70, D60, D91, E21, J22, J24.
“The unexamined life is not worth living.”

Socrates (related in Plato’s Apology).

1 Introduction

The problem of self-control –or, to use a more ancient terminology, willpower– has recently attracted renewed attention from economists. People’s common tendency to succumb to short-run impulses to seek pleasure or avoid discomfort at the expense of their long-run interests is often attributed to time-inconsistent preferences (for example of the quasi-hyperbolic kind), where the individual’s current self “overweighs” the present relative to the future. In recent years this type of model has been applied to a wide range of economic issues, such as consumption and savings decisions, asset pricing, addiction, procrastination, fiscal policy, and aggregate growth.

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. These include tying oneself to the mast, staying away from sources of temptation, holding illiquid assets, signing binding contracts or similar forms of “asking for controls” (e.g., committing to writing a paper for a conference).

This paper, by contrast, focuses on the internal commitment mechanisms or “personal rules” that receive much greater emphasis in psychology. Examples include diets, resolutions to smoke only after meals, jog four times a week, write five pages a day, always finish what you started, conduct your life with dignity, and many similar “promises to oneself.”

Given that these rules are entirely self-imposed, the first question that arises is of course how they could actually constrain the individual’s behavior. In other words, we seek to model genuine self-control, 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.

We also show, however, that personal rules may give rise to a very different kind of cost, which until now has received little attention in economics. It correspond to the “compulsive” or “obsessive” behaviors of people who feel compelled to work or save constantly without ever properly enjoying leisure or consumption (workaholism, avariciousness), 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 just two sides of the same coin.¹ In particular, it shows that agents with hyperbolic discounting can actually behave as if they overweighed

¹ The terms under- and over-regulation are borrowed from Baumeister et al. (1994), which provides a good survey of recent psychological research on these topics.
the future, rather than the present.

Building on Ainslie’s (1992) seminal work, we develop a theory of personal rules based on self-reputation. The idea is that because individuals have imperfect knowledge of their willpower, lapses are interpreted as precedents, and may thus have an 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 traditional bias towards instant gratification.

There are two key ingredients in our model. The first one is of course imperfect willpower, or salience of the present. The second one – whose essential role the paper brings to light – is imperfect recall. Indeed, whenever individuals 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.

Elaborating further, personal rules have, in our view, two facets:

1) Behavioral rules. Personal rules first refer to the reputation-dependent strategies that people employ to pursue self-control. When his will is put to the test, the individual takes into account not only the direct current benefits and delayed costs of choosing instant gratification in this particular instance, but also the indirect reputational impact of the precedent that this would set. We formalize this self-reputation game and derive a rich set of behavioral predictions, summarized below. We also examine how the equilibrium outcomes depend on the extent to which the individual’s self-monitoring is subject to opportunistic distortions of memory or inference.

2) Cognitive rules. An individual may influence the set of behavioral rules that he can sustain by using certain cognitive strategies that help increase the probability that lapses, as well as the circumstances under which they occurred (was there a legitimate excuse?) will be remembered and incorporated into self-reputation. This may be achieved by keeping a journal of one’s lapses and successes, or by attending a self-help group or therapy sessions which will force the individual to at least reflect on his past behavior; through the choice of a religion, philosophy or moral code, and the regular rehearsal of their precepts; or through the adoption of behavioral principles based on “bright lines”. In a nutshell, the effectiveness of such strategies arise from the fact that lapses become more salient when the individual not only caves in but, in doing so also breaks his personal rules.

Let us now briefly sketch our model. Each period has two subperiods (e.g., morning and afternoon). In the first subperiod, the individual either tries or does not try to exercise self-control. Trying means attempting a willpower-dependent activity (W), where his ability to exercise self-restraint will be put to the test in the second subperiod. Not trying means choosing a no-willpower option (NW) where no such “trial” will occur. When attempting activity W the individual will
thus be faced with a choice between persevering \((P)\) and giving up \((G)\). Persevering involves short–run costs and yields delayed gratification. As a result, the payoff to embarking on \(W\) will be highly dependent on the agent’s ability to resist short–run impulses to seek pleasure or avoid discomfort –that is, on his degree of willpower. We further assume that the individual is generally unsure of whether he will be able to resist the temptation to quit in the second subperiod; he can, however, make inferences about it from past experience. This learning process is why lapses may act as precedents, so that there can be no violating a rule “just this once”. The following examples illustrate our basic choice structure:

a) An alcoholic or smoker must decide every morning whether to try and abstain that day (activity \(W\)), or just start drinking or smoking right away (option \(NW\)). The first choice is potentially more profitable, but its success hinges on sustained self-restraint \((P)\), which may or may not be achieved. If he was confident of his ability to resist his cravings throughout the afternoon and evening, when temptation and stress will reach their peak, the agent might be willing to make the initial effort. If, on the other hand, he expects to cave in and “binge” before the day’s end anyway \((G)\), he is likely to ask himself “what is the point?” and just indulge himself from the start.

b) For a student, writer or researcher, activity \(W\) could take the form of an ambitious project that will pay off only if he perseveres when hardship is encountered along the way. Activity \(NW\), by contrast, would be either slacking off or devoting one’s time to an easier project which one knows “can be done” (so that willpower will not be seriously put to the test).

c) Social interactions often pay off only if one consistently cooperates, thereby foregoing short-term gains for long-term benefits. A person’s degree of impulsiveness is therefore a key determinant of his or her expected payoff from engaging in social relationships. With this in mind, activity \(NW\) refers to isolation (bachelors’ life, staying home to watch television, working alone, etc.), while activity \(W\) consists in attempting long-term relationships (friendship, marriage, partnerships etc.) This investment is worthwhile only if the individual has enough willpower to do what it takes maintain such relationships through trying times.

This paper will study the simplest framework in which self-reputation can operate, namely the two-period version of the general problem described above, with both periods linked to one another through imperfect memory and retrospective inference. This will allow us to identify and formalize the key ingredients of a theory of rule–based behavior, and obtain number of interesting results:

1) Self-restraint is greater when the situation is repeated, and when lapses are more likely to be brought back to awareness.
2) The extent of self–control achieved by an individual varies with his self–confidence, which acts as a “reputational capital”. Thus for a given actual willpower (or salience of the present), the
individual is often more likely to exert self-control if he has sufficient confidence in his fortitude—for instance because he is unaware of previous lapses. We also identify, however, situations where restraint declines with self-confidence. Under either scenario, self-control will be highly path-dependent.

3) An initial phase of external controls or incentives reduces the probability that, later on, the individual will trust himself enough to even put his will to the test. Forced choices, imposed for instance by “controlling” parents or a society with rigid norms, thus inhibit the development of self-confidence and autonomy.

4) Variability in the costs and benefits of resisting impulses creates an important role for retrospective attribution: “did I cave in because of low willpower, or because of special circumstances?” The availability of such excuses, in turn, tends to undermine self-control for the most weak-willed individuals, and reinforce compulsiveness for those who are more stronger-willed.

5) Compulsive behavior is more likely when the individual’s initial self-reputation is low, and when the veracity of self-excuses and ex-post rationalizations is difficult to ascertain.

6) An individual’s choice of cooperating with people he will never meet again can be viewed as a form of compulsion, reflecting a concern on his part about his self-reputation for being able to sustain satisfactory long-term relationships.

7) Rehearsed cognitive rules (resolutions, religious, moral or philosophical precepts, goals, etc.) increase the probability that lapses are brought back to awareness in the future. They are therefore conducive to both self-restraint and compulsiveness.

The remainder of the paper is organized as follows. Section 2 reviews some related literature. Section 3 presents the formal framework and discusses our modelling choices. Section 4 and 5 analyze how sustainability of different behavioral rules is affected by, respectively, the recall of lapses and the inferences made about the causes of these lapses. Section 6 investigates the nature and role of cognitive rules, and Section 7 concludes with suggestions for future research. All proofs are gathered in the appendix.

2 Related Literature

2.1 Self-Control

A rapidly growing literature views agents as subject to a form of intrapersonal conflict which can generate preference reversals, and creates a demand for commitment. This conflict may occur between successive intertemporal selves whose preferences differ due to non-exponential discounting (e.g., Strotz (1956), Ainslie (1992), Laibson (1994), O’Donoghue and Rabin (1999)), or between simultaneously coexisting subselves seen as representing competing interests or networks
As mentioned earlier, much of this work focuses on agents’ recourse to external commitment devices as a partial remedy for their intrapersonal conflicts. A more recent strand of work focuses on time-inconsistent individual’s incentives to manipulate their future perceptions of the payoffs attached to alternative courses of action. Through strategic ignorance (Carrillo and Mariotti (2000)), selective memory (Bénabou and Tirole (2000a)) and the like, the individual aims to minimize the intensity of future temptation, that is, to reduce the divergence between his ex-ante (or long-run) and his ex-post (or temporary) preferences.

Personal rules, by contrast, are defined by Ainslie (1992, 1999) as “the kind of impulse control which we call willpower, which allows a person to resist impulses while he is both attracted by them and able to pursue them”. The basic idea is that the individual should come to see each of his decisions as a possible precedent for future ones, so that giving in to temptation today raises the probability that one will do the same in the future. By thus tying together sequences of future choices between rewards, he “raises the stakes” on his current decision, and consequently will be more likely to behave according to his long-run interests. Ainslie traces this idea back to Victorian psychologists studying the question of The Will, and even to Antiquity with the work of Galen. As to why misbehaving today should make it more likely that one will also behave misbehave tomorrow (setting a precedent), Ainslie is much more elusive, but what he writes strongly suggests uncertainty and learning about the strength of one’s own will—leading us to propose a model based on of self-reputation over one’s degree of time-inconsistency (or salience of the present):

“But how does a person arrange to choose a series of rewards all at once?... Assuming that he is familiar with the expectable physical outcomes of his possible choices, the main element of uncertainty will be what he himself will actually choose. 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.”

This discussion also points out several important features of personal rules—and lapses—which our model will capture:

(a) Rules involve making attributions about the causes of one’s own behavior, and in particular about one’s general propensity to succumb or resist to temptations: “A person should consider, for each daily task, whether it is better to be the slave of passion or to go by reason. That is, he should classify every behavior into one of two accounts: passionate or reasonable” (Galen).

(b) The effects of past actions are cumulative: “A man who has for a long time habitually
fallen into error finds it difficult to remove the defilement of the passions from his soul”.

(c) A little backsliding can be disproportionately dangerous: “He must not relax his vigilance for a single hour” (Galen). “Every gain on the wrong side undoes the effects of many contests on the right” (Bain 1986, cited by Ainslie).

Numerous other psychologists (e.g., Baumeister et al. (1994)) have also emphasized the importance of self-monitoring (keeping track of one’s actions) for successful self-regulation, and pointed out the often devastating to a subject’s self-view and subsequent behavior breaking a personal rule. These observations all point to the building up and running down of self-reputation over one’s willpower, defined as the inverse of the bias towards immediate gratification.

2.2 Self-signaling

Most closely related to our paper is the work of Bodner and Prelec (2000), who examine the informational value of actions in a split self (e.g., ego-superego) model where the individual has “metapreferences” over his own, imperfectly known, tastes. They thus consider an intratemporal signaling game in which two contemporaneous selves are asymmetrically informed. The sender chooses a costly action that may signal a characteristic unknown to, but valued by, the receiver. For instance, the individual may derive hedonic value (or meta-utility) from thinking of himself as a generous person, and take actions to try and convince himself that he is of that type. Bodner and Prelec’s model thus allows for a broader set of self-signaling behaviors than ours, which focuses more specifically on willpower. On the other hand, it requires that the individual derive utility from pure beliefs. As to the feature that the individual should both know and not know his real type, it could be resolved intertemporally by introducing imperfect recall, as in our model.

2.3 Addiction

Our work is also related to theories of addiction (see Herrnstein and Prelec 1992 for a survey). For instance, some of its empirical predictions are similar to those Becker and Murphy’s (1988) model of rational addiction. First, consumption (instant gratification) today increases the likelihood of a similar behavior tomorrow. Second, repetition and low discounting are conducive to more restraint (avoidance of instant gratification). Third, the fear of addiction may generate behaviors that resemble our compulsive behavior.2 There also significant difference between the two models, however. First, no matter how unhappy rational addicts may be along their optimally chosen consumption paths, they would be even more unhappy if they were prevented from consuming

---

2 Orphanides and Zervos (1995) generalize the Becker and Murphy model to account for uncertainty about one’s own preferences. In such a world, even a rational individual may rigidly shun from trying when he does not know that the good is actually not very addictive. For a different view, based on self-control problems, see Carrillo (1998).
the addictive good. 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. Another key difference is the occurrence of relapse following a period during which the agent is prevented from consuming the addictive substance. There is little scope for relapse in the theory of rational addiction, as the stock of “addictive capital” depreciates over time. Yet relapses are notoriously common and, even when they are avoided, the cravings and temptations often remain. 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), p125). Our model will be consistent with this kind of account, and even show that the temporary imposition of external restraints can actually undermine the individual’s capacity to achieve self-control later on, resulting in more substance abuse rather than less.

3 The Model

3.1 Decisions

We consider an individual with a two-period horizon, \( t = 1, 2 \). Furthermore, each period \( t \) is divided into two subperiods, I and II (e.g., morning and afternoon). At the start of each subperiod I, the individual chooses between:

1) A “no willpower” (NW) activity, which yields a known payoff \( a \) in subperiod I. This option corresponds to indulging in “immediate gratification” by drinking, smoking, eating, shopping, 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) project or investment: attempting to exercise moderation or abstinence in drinking, smoking, eating, or buying; or taking on a challenging task (homework, exercising, ambitious project), etc. Depending on his strength of will, which is described below, the individual may then, at the beginning of subperiod II, opt to persevere until completion \( (P) \), or give up along the way \( (G) \). Holding fast entails a short-run cost \( c > 0 \) (e.g., craving) during subperiod II, but yields delayed gratification in the form of future payoffs (better health, higher consumption etc.) whose present value, evaluated as of the end of period \( t \), is \( B \). Caving in, on the other hand, results in a painless subperiod II, but yields only a delayed payoff \( b < B \).

To summarize, there are three possible outcomes, illustrated on Figure 1: (a) not even trying; (b) trying but then giving in; (c) trying and persevering. The resulting flow payoffs during \((t, I)\)
and \((t, II)\), and the associated delayed benefits or costs as of the end of period \(t\) are respectively \((a, 0; 0), (0, 0; b)\) and \((0, -c; B)\). An important issue in some of our results will be whether some self-restraint (trying and then giving up) is better or worse than none at all—that is, whether \(b\) exceeds \(a\). The examples offered in the introduction show that either situation may arise. An individual is probably better off if he refrains from smoking or does some work for a while than if he makes no effort at all \((b > a)\). On the other hand, it may be better not to start on a difficult project that one will not complete, or to remain single than to get married and ultimately divorced \((a > b)\).

3.2 Preferences: State–Dependent Willpower

In addition to a standard discount rate \(\delta\) between periods 1 and 2, we assume that the individual’s preferences exhibit time–inconsistency, represented by the usual quasi–hyperbolic specification. We shall, however, enrich this now familiar setup with two important and realistic new elements: the intensity of temptation, or “salience of the present”, is state–contingent, and imperfectly known in advance. Thus:

1) When deciding whether to choose the default activity \(NW\) or attempt \(W\), 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 option would bring about, starting in the next sub-period. Accordingly, the agent discounts the latter at a rate $\gamma \leq 1$; equivalently, he values the immediate gratification from $NW$ at $a/\gamma$ instead of just $a$.

2) If he nonetheless decides to attempt the $W$ activity, he is again confronted with another (typically more intense) temptation at the beginning of subperiod II: should he hold fast in his resolution, incurring the craving $c$ for the sake of higher future benefits $B - b$, or should he give in? Once again, immediate payoffs loom larger than future ones, so in the agent’s decision–making the cost $c$ gets magnified to $c/\beta$, with $\beta < 1$. Equivalently, future payoffs are discounted by $\beta$.

There is no reason a priori why $\beta$ and $\gamma$ should be equal, as they reflect the individual’s temporal preferences in very different situations. For instance, $1/\gamma$ measures how much he craves a cigarette or a glass of alcohol in the morning; $1/\beta$ measures how much he craves it in the afternoon, after a whole morning of deprivation. Another key difference between $\gamma$ and $\beta$, discussed below in more detail, is that the former is known at the initial stage when the agent decides whether or not to attempt the $W$ activity, whereas the latter 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 equal to a known $\gamma$ in “normal” times, but can take values $\beta = \beta_H$ or $\beta = \beta_L$ in times of stress—whether caused by abstinence, by the proximity of temptation, or by cues that intensify “visceral” cravings (Loewenstein (1996), (1999)).

The fact that $\beta < 1$ creates, as usual, a conflict of interest between the individual’s successive temporal selves. At the beginning 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). Thus $1/\beta$ measures the strength of the temptation to “give in”. Conversely, we shall refer to $\beta$ as the individual’s strength of will, that is, his ability to withstand discomfort and delay gratification in situations of intense temptation or stress.

Similarly, when $\gamma < 1$, the individual in period 1 would like task $W$ to be undertaken in period 2 as long as it yields an expected benefit of at least $a$. But when making the decision at the start of period 2 he will actually take the “path of least resistance” and select $NW$ whenever this same expected benefit is less than $a/\gamma$. 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 absolutely sure that he learns that it is strong: if he sees that today one caved in, he will not even make a partial attempt at self-discipline in the next period (will not choose $W$), but will give up the fight without trying. Thus, as in Seligman’s (1975) theory of learned helplessness, “quitting is mediated by thinking that

---

3 One could thus have $\gamma = \beta_H$. More generally, the idea that craving or pain becomes increasingly difficult to withstand, the longer it lasts, suggests that $\gamma \geq \beta_H$. This condition is not needed for any of our results, however.
Because even partial self-control is (ex ante) generally better than no self-control, the individual will try to avoid such a costly “demoralization”.4

3.3 Information: Experience and Memory

As noted by psychologists, people frequently look back to their own past behavior to infer what they are likely to do in the 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 individuals have only limited knowledge of their strength or weakness of will, and learn about it only by putting it to the test (an extreme example is a person’s ability not to speak under torture). Furthermore, because $\beta$ can only be truly known through direct experience, 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 $\beta = \beta_L$ or $\beta = \beta_H$, with $\beta_L < \beta_H \leq 1$. The individual initially does not know $\beta$, but has priors $\rho_1$ and $1 - \rho_1$ on $\beta_H$ and $\beta_L$. Furthermore, if in period 1 he discovers the true value of $\beta$ through the experience of craving, he later on cannot directly recall this value.

As a result of this imperfect recall, the individual in period 2 will have to try and infer his $\beta$ from the record of what he actually did when confronted with temptation, plus possibly his reconstruction of the circumstances under which he made his choice. There are several reasons, and supporting bodies of evidence, why $\beta$ cannot just be directly remembered from previous experience.

First, it does appears difficult for individuals to accurately recall from “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. Thus Kahneman et al. (1997) document a systematic divergence between subjects’ moment-by-moment reports during a painful medical procedure or unpleasant laboratory experiment, and their later retrospective evaluation of the experience a whole. Loewenstein and Schkade (1999) report on many other

---

4Note 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.
experiments and field studies indicating that similar “hot–cold empathy gaps” occur in recollections and predictions about feelings such as hunger, exhaustion, drug or alcohol craving, or sexual arousal.\textsuperscript{5} Just as people cannot accurately answer retrospective questions such as “how much did it hurt?” or “how much did you dislike that?”, we think that asking oneself “how much did I crave that cigarette, that glass of wine, or that new jacket yesterday?” or “how cold and stiff was I when I went jogging last week?” is unlikely be very informative, compared to asking what one actually chose to do.

The second reason why this economist–like “revealed preference” approach to predicting one’s own behavior is warranted is that the 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 extremely “soft” information, whose veracity is much more difficult to verify than for the somewhat “harder” data of one’s past deeds, which often leave a material record. Indeed, a lot of research has confirmed the common observation that people’s recollections are often self–serving: they tend to remember their successes more than their failures, reframe their actions so as to see themselves as instrumental for good but not bad outcomes, and find ways of absolving themselves by attributing responsibility to others.\textsuperscript{6}

Finally, both the recollection of past motives and their reconstruction on the basis of past behavior are often complicated by the need to discriminate between the role of one’s true character and that of temporary (situational) factors: did I crave so badly and ultimately give in to temptation because my \(\beta\) is low, or because of an unusually high \(c\)? In other words, can I find extenuating circumstances for myself? The following assumption will allow us to address this important aspect of rule–based behavior.

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

The key point is that the (direct) return perceived by the individual when deciding whether or not to “stick with it”, \(-c + \beta (B - b)\), reflects both a fixed trait of his personality and some idiosyncratic (external or internal) circumstances.\textsuperscript{7} In interpreting past lapses or fortitude, he

\textsuperscript{5}As Loewenstein argues: “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” (1996, p284). These limitations may have benefits in terms of evolutionary fitness, as with the inability of women deciding whether or not to request an epidural anesthesia to accurately recall the pain of previous deliveries.

\textsuperscript{6}On self-serving memory, see Korner (1950), Crary (1996), Mischel et al. (1976), Kunda and Sanitioso (1989), or Murray and Holmes (1994). On self-serving attributions, see Zuckerman (1979) and Snyder et al. (1983).

\textsuperscript{7}The general problem is that of trying to filter the permanent from the transitory component of \(c/\beta\). 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
therefore faces a problem of signal—extraction, or attribution. Of course, he will often be tempted, ex post, to “rationalize” his failings, if not to forget (repress the memory of) his lapses altogether. We shall therefore allow for imperfect recall (or imperfect retrospective verifiability) with respect to both past actions and past circumstances.

**Assumption 3** Suppose the individual undertakes the $W$ activity 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 $\lambda$. 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.

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 $\sigma$. Then $\Pr(\hat{\sigma} = P \mid \sigma = P) = 1$, but $\Pr(\hat{\sigma} = G \mid \sigma = G) = \lambda \leq 1$. Similarly, there are two memory (or awareness) states concerning the period–1 environment (cost $c_1$) in which task $W$ was attempted: one in which no plausible excuse can be invoked, $\hat{c} \equiv c_L$, and one in which there exists a plausible excuse (it may or may not be true, but cannot be disproved), $\hat{c} = c_H$, with $\Pr(\hat{c} = c_H \mid c_1 = c_H) = 1$ but $\Pr(\hat{c} = c_L \mid c_1 = c_L) = \nu \leq 1$.

**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 individual will recall it only with probability $\nu$; with probability $1 - \nu$ he will no longer be able to distinguish whether $c_L$ or $c_H$ occurred.

The assumption that ego–favorable events are more likely to be remembered than unfavorable ones is not essential for our main results, but it is in accordance with a lot of empirical evidence, and will allow for interesting comparative statics that highlight the cognitive underpinnings of self–regulation. It also emerges endogenously when one recognizes that the individual has some measure of control (through rehearsal, cue or attention management, etc.) over his memory or awareness. Indeed, we noted earlier that he will generally have strong incentives, ex post, to try and forget his lapses (having succumbed to temptation); or, if a lapse is remembered, to try and find admissible “excuses” (a high $\hat{c}$) that allow him to attribute his behavior to temporary, external factors, rather than to enduring, personal ones.

---

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”. In the model the first case corresponds to $\beta = \beta_L$, the second and third to temporary decreases in $\beta$ and increases in $c$ respectively. Since all that matters is $c/\beta$ we merge them into fluctuations in $c$ only.

8Of course, the individual need not literally erase past actions and situations from his memory. The key point is that he can (within limits and at some cost) affect the probability of still being consciously aware of them in the future. See Bénabou and Tirole (2000a) for a model of endogenously selective memory or awareness.
Ex ante, on the other hand, the individual will try to find ways—which we define later on as “cognitive rules”—to prevent himself from driving $\lambda$ and $\nu$ toward zero. In most of the paper we shall take these two parameters as given, and analyze how the set of behavioral rules which the individual can credibly use to achieve self-control depends on the reliability of his memory and inference processes. In Section 7 we shall then discuss how $\lambda$ and $\nu$ can be endogenously determined through the use of cognitive strategies.

### 3.4 Behavioral Rules as Reputational Equilibria

“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. The difference in value can also be regarded as the stake of a private side bet that the person “makes” to precommit his future behavior. It is this stake that gives the will his force.” Ainslie (1992)

We are interested in rules that are self-sustaining, without reliance on external commitment devices. Accordingly, we shall define a “behavioral rule” as a Perfect Bayesian Equilibrium (PBE) between the successive selves at their different information states; that is, as a self-enforcing set of prescriptions for actions in different states $(\beta, c)$, together with consistent inferences across information sets. In particular:

(a) At the beginning of each period, $t = 1, 2$, the individual’s choice between the NW and the W options reflects his current beliefs about his own willpower, $\rho_t$.\(^9\)

(b) When deciding whether to persevere in the W activity or to give up, his decision reflects the strength of the current temptation he perceives, measured by $c_t/\beta$, his “private” (with respect to future selves) information about his true type, $\beta$, and the coarser “public” information $\rho_t$ (current reputation).

The key point in our theory of personal rules is thus to highlight and explicitly model the role played by the uncertainty which people face concerning their own willpower. This emphasis is what leads us to focus the analysis on a game with a finite horizon and imperfect information, where non-myopic rules can only be sustained only through their impact on self-reputation.

An alternative approach could have been to assume that the individual knows his intrinsic ability to resist impulses ($\beta$), and analyze the infinitely repeated version of the game.\(^10\) As shown by Laibson (1994), this supergame among the selves admits an infinite number of equilibria; in

\(^9\)In principle, both the NW/W and the P/G choices in period 2 could also be conditioned on other past events (such as which cost was observed in period 1), but in practice this will not happen. In other words, the (payoff-relevant) state variables described above turn out to be sufficient statistics.

\(^10\)We could also study reputation-based Markovian perfect equilibria of an infinite horizon game. That route has the advantage that rules are repeatedly applied, which is more realistic. On the other hand it is more complicated and makes it less clearly apparent that some of these equilibria are not of a “bootstrap nature”.

13
some the individual succeeds in resisting his impulses, whereas in others he systematically caves in. One could call the agent strong-willed in the first type of equilibrium and weak-willed in the second, but the supergame approach offers no clue as to when he is likely to be in one or the other.\textsuperscript{11} Moreover, the underlying trigger strategies typically lead to Pareto-dominated outcomes—that is, they are not renegotiation-proof among the selves. 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.

Besides methodological issues of indeterminacy and renegotiation, we feel that the case for a theory of willpower explicitly based on imperfect self-knowledge and ongoing self-evaluation is quite strong. As discussed earlier, both psychologists’ writings and everyday observation make clear that informational issues are central to the problem of regulating one’s behavior. These cognitive aspects include, as in our model, the fact that history matters through a person’s beliefs about his ability to withstand temptation (self-confidence), the importance of imperfect recall (self-monitoring, salience of lapses), and the tricky attribution problem of distinguishing weakness of will from adverse external factors (assessing excuses). For instance, the standard strategy in psychotherapy of helping the individual rebuild his credibility and self-confidence presupposes a learning process. Similarly, our cognitive model will make clear predictions as to how an initial phase of external constraints (e.g., “controlling” parents) will affect future self-control; a supergame approach, by contrast, would be completely silent on this question.

\section{Lapses as Precedents: Self-Regard and Motivation}

In this section we study how impulses for immediate gratification can be held in check by the fear of “losing faith in oneself,” which would lead to a further collapse of self-control. On the cognitive side, we highlight the role played by the recall of past lapses, parametrized by $\lambda$, in the maintenance of personal rules. To abstract from the additional inference problem of attributing such lapses to personal weakness or external circumstances, we assume that the cost of persevering in the $W$ activity takes only one value, $c_H = c_L = c$, with:

$$\frac{c}{\beta_H} < B - b < \frac{c}{\beta_L}. \quad (1)$$

\textsuperscript{11}See also Meardon and Ortman (1996). Caillaud et al. (1999) criticize this interpretation and argue that, if the individual knows his ability $\beta$ to resist impulses, he should be able to behave in a coherent way, consistent with the fundamental “unity of the self”. Going even further in that direction, when multiple PBE’s exist in our model we shall generally emphasize the Pareto-dominating one(s). This is not to say that multiplicity is uninteresting, especially when it involves different possible interpretations of the same action (different degrees of pooling or separating among types in the PBE). But it makes clear that the papers’ main insights do no hinge on such multiplicity.
Thus, in a static context, the $\beta_L$ type will always cave in and the $\beta_H$ type will never yield. Such is the case in the last period of our basic two-period model, as there are no more reputational concerns. Knowing this, the individual will opt for the $W$ activity at the beginning of period 2 only when his reputation is above the threshold $\rho_2^*$ defined by:

$$\rho_2^*(B - c) + (1 - \rho_2^*)b \equiv \frac{a}{\gamma},$$

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

**Assumption 5** $B - c > \frac{a}{\gamma} > b$.

### 4.1 Self-Reputation as a Mental Asset

We shall be primarily interested here in the case where even a partial effort at self-control is better than none: $b > a$. This means that, ex ante, the individual would prefer that $W$ be undertaken, even he was certain to eventually give up along the way.\(^{12}\) Ex post, however, he is too tempted to take the path of least resistance $NW$, and not even attempt self-restraint: when $\gamma < 1$, the threshold $\rho_2^*$ in (2) is suboptimally high. Because confidence in his own willpower (a higher $\rho_2$) helps shore up his motivation it represents a valuable asset, worthy of protection (self-restraint) in period 1.

Let us denote the probability that task willpower is put to the test ($W$ is selected) at the beginning of period 2 as:

$$p_2(\rho) : \begin{cases} 
1 & \text{if } \rho > \rho_2^* \\
[0, 1] & \text{if } \rho = \rho_2^* \\
0 & \text{if } \rho < \rho_2^*
\end{cases}$$

and the resulting ex-ante value functions for each type as $V_H^2(\rho) \equiv p_2(\rho)(B - c) + (1 - p_2(\rho))a$ and $V_L^2(\rho) \equiv p_2(\rho)b + (1 - p_2(\rho))a$ respectively.\(^{13}\) Note that $V_H^2$ is non-decreasing in $\rho$, and so is $V_L^2$ when $b > a$.

Let us now consider behavior in period 1. We shall focus our 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, this means that $\rho_2^+ \geq \rho_1 \geq \rho_2^-$ for all $\rho_1$, where $\rho_2^+$ and

---

12 By contrast, when $b \leq a$ the individual would never want to convince himself (i.e., his futures selves) that he is strong-willed when he is in fact weak-willed.

13 By ex-ante, we mean value functions where the evaluation of payoffs is free from the “salience of the present”.

---

15
respectively denote the posteriors in each of these events.\footnote{The specific expressions obtained from Bayes’ rule when $\hat{\sigma} = P$ and $\hat{\sigma} = G$ are given by setting $c_L = c$ in (A.4)--(A.5) in the appendix. The monotonicity assumption 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.}

Because $B - b < c/\beta_H$ and $V^H_2(\rho)$ is non-decreasing in $\rho$, the monotonicity of beliefs implies that perseverance ($P$) is a dominant strategy for type $\beta_H$. Consider now type $\beta_L$; the threat of revealing himself as weak-willed will deter him from yielding to temptation if and only if:

$$\frac{c}{\beta_L} - (B - b) \leq \delta \lambda \left[ V^L_2(\rho^+_2) - V^L_2(\rho^-_2) \right].$$  \hspace{1cm} (4)

The left-hand side is the disutility of not giving in, while the right-hand side represents the value of self-reputation that will be foregone if one does, and this lapse is recalled in the next period. From the results derived for period 2, this reputational stake is at most $\delta \lambda (b - a)$. As a result, it is clear that if $c/\beta_L > C(\lambda)$, where

$$C(\lambda) \equiv B - b + \delta \lambda (b - a),$$ \hspace{1cm} (5)

then playing $G$ is a dominant strategy, and the unique equilibrium is one where the weak type always gives in to his impulses. We shall assume from here on that $c/\beta_L < C(\lambda)$; note that given (1), this requires $b > a$. It will be useful to define, for each $\lambda$, the following threshold:

$$\tilde{\rho}_1(\lambda) \equiv \frac{(1 - \lambda)\rho^*_2}{1 - \lambda \rho^*_2}.$$ \hspace{1cm} (6)

As $\lambda$ increases from 0 to 1, $\tilde{\rho}_1(\lambda)$ decreases from $\rho^*_2$ to 0. As we shall see, it represents 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 individual’s initial reputation $\rho_1$ 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;

i) the weak–willed type perseveres with probability 1 when $\rho \geq \rho^*_2$, with probability $q_1$ such that

$$\rho^+_2 \equiv \frac{\rho_1}{\rho_1 + (1 - \rho_1)(q_1 + (1 - \lambda)(1 - q_1))} = \rho_2^*$$ \hspace{1cm} (7)

when $\tilde{\rho}_1(\lambda) < \rho_1 < \rho^*_2$, and with probability 0 when $\rho_1 < \tilde{\rho}_1(\lambda)$. Thus, $q_1$ rises with $\rho$ and $\lambda$.

In period 2, if a lapse is recalled the NW option is chosen. If the individual’s prior was $\rho_1 > \rho^*_2$ and no lapse is recalled, his posterior remains equal to $\rho_1$, and he chooses the $W$ activity. If his
prior was $\rho_1 \in [\bar{\rho}_1(\lambda), \rho_2^*]$ and no lapse is recalled, his posterior is equal to $\rho_2^*$ and the W activity is chosen with probability

$$p_2^* = \frac{c/\beta_L - B + b}{\delta\lambda(b - a)}.$$

4.2 Main Implications

Proposition 1 has a number of interesting implications.

1. The value of self-confidence. As illustrated on Figure 2, the degree of self-restraint $q_1$ increases with initial reputation $\rho_1$, rising from 0 for $\rho_1 \leq \bar{\rho}_1(\lambda)$ to 1 at $\rho_1 = \rho_2^*$. Self-control is easier to achieve for someone who is more confident of his strength of will, because this reputational capital is an asset that can be “staked” upon proper behavior in the current period.

Conversely, if the individual does not “trust himself” enough ($\rho < \rho_2^*$), complete self-control is not attainable in period 1: mimicking the strong type (playing $q_1 = 1$) would leave priors unchanged, and thus fail to prevent NW from being chosen in period 2. Instead, with $q_1 < 1$, and provided that $\rho_1$ is not too low, recalling no lapses from period 1 represents sufficiently “good news” about his type that the individual in period 2 becomes willing to again put his will to the test (in fact, he his just at the point of indifference). Finally, the individual’s ex-ante welfare is easily shown to also be increasing in $\rho_1$.

2. The effect of external controls. Suppose that during the first period (e.g., childhood), the
individual’s behavior is subject to external constraints —imposed for instance by “controlling” parents, or a society with rigid norms: when confronted with activity $W$ he is constrained or strongly incentivized to select $P$. By contrast, date–2 behavior (e.g., in adulthood) remains subject to free will. Proposition 1 implies that such controls always (weakly) reduce the probability that the individual puts his will to the test at date 2. Because nothing is learned from period 1’s actions, $\rho_2$ remains equal to $\rho_1$. For $\rho_1 \geq \rho_2^*$ this has no impact, as the individual would have persevered anyway; by contrast, for $\rho_1 < \rho_2^*$ external control raises the probability of date–1 perseverance to 1, but prevent the individual from acquiring the self-confidence required to make the right choices on his own later on. Thus he will not even attempt self-restraint in period 2, whereas he would otherwise have chosen $W$ with probability $p_2^* > 0$ (for $\rho_1 \geq \tilde{\rho}_1 (\lambda)$). It also follows that outside constraints will now be eagerly sought out be the individual in period 2: a dependence on external rewards or punishments has been created.

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

These theoretical predictions are quite consistent 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... A manipulated child or other prisoner simply depends less on his will than does a free person... Thus, strong, externally regulated motives should reduce the strength of will” (Ainslie (1992, p. 174-175)).

3. The effect of memory. Figure 2 also shows that $q_1$ is increasing in $\lambda$: the individual is more likely to persevere, the less forgettable are his lapses; this is quite intuitive. It is also easy to see that ex–ante welfare is always higher with $\lambda = 1$ than with $\lambda = 0$.

Comparing the equilibrium for different values of $\lambda$ thus already provides the intuition for the importance of cognitive rules (e.g., keeping a journal, rehearsing moral principles, going to

15 Another way the intrinsic motivation may be damaged is if the “principal” imposing the external constraints or other incentives is privately informed about the individual’s ability or willpower, or about the nature of the task. See Bénabou and Tirole (2000b) for a theory of the tradeoff between extrinsic and intrinsic motivation in social interactions.

16 Similarly, the equilibrium exhibits a measure of self-control only when $c/\beta_L < C(\lambda)$ and $\rho_1 > \tilde{\rho}_1 (\lambda)$. Since $C(\lambda)$ increases and $\tilde{\rho}_1 (\lambda)$ decreases with $\lambda$, a higher degree of recall of one’s lapses increases the range of situations in which the individual can achieve self-control.

17 It is not obvious, however, whether it always rises monotonically with ASuch is the case when $\rho_1 \geq \rho_2^*$, as a higher $\lambda$ just makes it more likely that $c/\beta_L < C(\lambda)$. For $\rho_1 < \rho_2^*$, however, note that $p_2^*$ is decreasing in $\lambda$. 18
confession or therapy) as a critical complement to behavioral rules. Only if lapses are sufficiently memorable will reputational equilibria with self-control be sustainable. There arises therefore a strong tension between the ex-post desires of the individual, who would generally want to forget his failings (drive λ toward zero) in order to avoid the reputational costs, and his ex-ante interest: he can only escape “being the slave of passion” if he is able to prevent or limit such selective memory or attention. As Ainslie (1992), p. 154) notes: “Behavior therapists regularly observe that when patients systematically record either impulsive behaviors or avoidance of such behaviors, the occurrence of such behaviors decreases; a practice called self-monitoring”.

5 Rules, Exceptions, and Excuses

In sufficiently adverse circumstances, even the strongest-willed individual would, and perhaps should, give up rather than persevere. When really ill, or on special family occasions, even a perfectly time-consistent person will postpone work to another day. During a snow storm or heat wave, the jogging-every-day rule should be broken; when a host prepared a special dessert, it would be more impolite than heroic to refuse. In inferring one’s strength of will from one’s actions, attention must thus be paid to the circumstances under which they took place. To highlight this attribution problem, we now assume that actions are always remembered (λ = 1), but let the cost of resisting one’s impulses to take two values, \( c_H \) and \( c_L \).\(^{18}\) To allow for both beneficial self-control and harmful compulsiveness, we assume a cost-benefit configuration such that, in a static context, the weak-willed type always gives in, while the strong-willed type only does so when the cost is high—which provides a legitimate “excuse”.

Assumption 6 Let \( c/β > B-b \) for all \( β \in \{β_H, β_L\} \) and \( c \in \{c_L, c_H\} \), except for \( c_L/β_H < B-b \).

Finally, the individual’s inference problem may be compounded by a self-serving, or simply imperfect, or memory. As explained in Assumption 2, the occurrence of \( c_L \) is correctly recalled or interpreted only with probability \( ν \leq 1 \) (recall state \( ˆc = c_L \)). With probability \( 1-ν \), the individual can no longer distinguish whether \( c_L \) or \( c_H \) occurred (recall state \( ˆc = c_H \)).

5.1 Feasible Behavioral Rules

We shall first describe the four basic behavioral rules that can occur in equilibrium, each of which highlights an interesting aspect of the problem. We then turn to the mixed-strategy patterns that can also emerge as combinations of these four basic cases.

\(^{18}\)Alternatively, one could introduce temporary fluctuations in the long-run benefits from perseverance, \( B - b \).
A. Pure strategies. We shall impose parameter restrictions such that $P$ is always dominant for $(\beta_H, c_L)$ and $G$ dominant for $(\beta_L, c_H)$. This leaves four possible types of pure strategy configurations to be played in each period:

\[
R_0 \equiv \begin{bmatrix} c_L & P & G \\ c_H & G & G \end{bmatrix}, \quad R_1 \equiv \begin{bmatrix} c_L & P & P \\ c_H & P & G \end{bmatrix}, \quad R_2 \equiv \begin{bmatrix} c_L & P & P \\ c_H & G & G \end{bmatrix}, \quad R_3 \equiv \begin{bmatrix} c_L & P & G \\ c_H & P & G \end{bmatrix}.
\]

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 end-game played in period 2.

2. Flexible or contingent rule. A generally superior rule is the flexible one, $R_2$: do not give in, unless circumstances are really unfavorable. In that case $c = c_H$ represents an excuse or exception that justifies giving in, and it is invoked by both types when they do so. For $R_2$ to be sustainable, however, requires sufficient confidence that if one deviates to $G$ when $c = c_L$, this lapse will be detected and classified as such. In other words, $\nu$ must be high enough, as we shall establish.

3. Legalistic rule. The next rule, $R_1$, is one which we might call “stringent” or “legalistic”, as it admits no excuses: giving in is always interpreted as a sign of weakness, and this inference is always correct in equilibrium. It has a clear benefit, in that the weak type is forced to exercise self-discipline (when $c = c_L$). On the other hand there is a potential cost: even in situations where an a valid argument might be made for yielding “just this time” (meaning “just in the state of the world $c_H$”), the strong type will persevere for fear of appearing weak. In that case the individual is “being 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 one could even adhere to a rule with the same costs from rigidity but no actual benefits from self-control. 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. As we shall see, rigid behavior such as $R_3$ or $R_1$ is more likely to emerge when the veracity of excuses is more difficult to assess ex post ($\nu$ is low).

B. Mixed strategies. Just as in the one-cost case for $\rho_1 < \rho^*_2$, the equilibrium may also take the form of a mixture of the “pure” rules described above. That is, either type $\beta_H$ randomizes between $P$ and $G$ in state $c_1 = c_H$, or type $\beta_L$ randomizes in state $c_1 = c_L$. Denoting the mixture
of pure strategy rules \( R_i \) and \( R_j \) as \( R_{ij} \), and randomization as \( P/G \), the four possible cases are:\(^{19}\)

\[
\begin{array}{ccc}
R_{02} & \beta_H & \beta_L \\
c_L & P & P/G \\
c_H & G & G
\end{array}
\begin{array}{ccc}
R_{03} & \beta_H & \beta_L \\
c_L & P & G \\
c_H & P/G & G
\end{array}
\begin{array}{ccc}
R_{12} & \beta_H & \beta_L \\
c_L & P & P \\
c_H & P/G & G
\end{array}
\begin{array}{ccc}
R_{13} & \beta_H & \beta_L \\
c_L & P & P/G \\
c_H & G & G
\end{array}
\]

5.2 Decisions in Period 2

In the last period there are no more reputational concerns, so the unique equilibrium is \( R_0 \) (given Assumption 6). At the start of period 2 the individual 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

\[
\phi \equiv \pi(B - c_L) + (1 - \pi)b.
\]

Note that payoffs are evaluated here ex ante, i.e. at a point where the individual’s preferences between \( P \) and \( G \) are not yet subject to the distortion \( 1/\beta \) which he will experience when actually confronted with the craving \( c \). By contrast, his preferences between the payoff \( \rho \phi + (1 - \rho)b \) to be expected from undertaking \( W \), and the immediate gratification \( a \) afforded by \( NW \), already reflect the temptation to “take the easy route” and avoid testing his will altogether. Thus, he chooses \( W \) only if his self—reputation \( \rho_2 \) is above the threshold \( \rho^*_2 \) defined by

\[
\rho_2^* \phi + (1 - \rho^*_2)b \equiv \frac{a}{\gamma}, \tag{8}
\]

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

Assumption 7 \( \phi \equiv \pi(B - c_L) + (1 - \pi)b > a/\gamma > b. \)

The probability \( p_2(\rho) \) that willpower is put to the test (\( W \) is selected ) in period 2 is thus still given by (3), except that the threshold \( \rho^*_2 \) is now defined by (8). The resulting ex–ante value functions for each type are now \( V^H_2(\rho) \equiv p_2(\rho)\phi + (1 - p_2(\rho))a \) and \( V^L_2(\rho) \equiv p_2(\rho)b + (1 - p_2(\rho))a \) respectively. Note that \( V^H_2 \) is non–decreasing in \( \rho \), and so is \( V^L_2 \) whenever \( b > a. \)

---

\(^{19}\)The case \( R_{01} = R_{23} \) would correspond to situations where both types use mixed strategies. We show, however, that this does not occur, because both cannot be made indifferent by the individual’s single task–selection strategy in the next period.
5.3 Decisions 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 being easy). The individual therefore perseveres if and only if:

\[
\frac{c_H}{\beta_i} \leq B - b + \delta \left[ V^i_2(\hat{\rho}^+ - \hat{\rho}^-) \right],
\]

where \( \hat{\rho}^+ \) and \( \hat{\rho}^- \) denote the posteriors beliefs (given in the appendix) following the events \( P \) and \( G \) respectively, given that extenuating circumstances can plausibly be invoked (\( \hat{c} = c_H \)).

Let us now turn to the case where \( c_1 = c_L \), which leads to the recall state \( \hat{c} = c_L \) (no possible excuse) with probability \( \nu \), and to \( \hat{c} = c_H \) with probability \( 1 - \nu \). The individual therefore perseveres if and only if:

\[
\frac{c_L}{\beta_i} \leq B - b + \delta \nu \left[ V^i_2(\hat{\rho}^+ - \hat{\rho}^-) \right] + \delta (1 - \nu) \left[ V^i_2(\hat{\rho}^+ - \hat{\rho}^-) \right],
\]

where the updating rules \( \rho^+_2 \) and \( \rho^-_2 \) in the recall state \( \hat{c} = c_L \) are also given in the appendix. This expression makes clear the “fudging” role of \( \nu < 1 \). The first two terms on the right–hand side give an expression similar to (4) in the one–cost case, with the probability \( \lambda \) that a lapse is recalled now replaced by the probability \( \nu \) that a lapse cannot be rationalized away. The final term in (10) is new, and represents the more moderate loss in reputation that occurs when a case of hardship can be plausibly invoked (\( \hat{c} = c_H \)), knowing of course that this excuse could be the product of accidental misinterpretation or deliberate fabrication.

- **Restrictions on beliefs.** We shall again focus our attention on equilibria satisfying monotonicity in beliefs, that is, such that observing \( P \) always raises (weakly) the probability that the individual is a strong–willpower type, while observing \( G \) always lowers it (weakly). Formally, this means that \( \hat{\rho}^+_2 \geq \rho_1 \geq \hat{\rho}^-_2 \) and \( \rho^+_2 \geq \rho_1 \geq \rho^-_2 \) for all \( \rho_1 \), both on and off the equilibrium path (this is the only restriction we impose on beliefs following a probability zero event).

- **Dominant strategies.** Because \( B - b > c_L/\beta_H \), because \( V^H_2(\rho) \) is increasing in \( \rho \), and because of monotonicity in beliefs, it is clear from (9) that when \( c_1 = c_L \), playing \( P \) is a dominant strategy for type \( \beta_H \). Conversely, since \( V^L_2(\rho) - V^L_2(\rho') \leq \max \{b - a, 0\} \) for all possible \( \rho \) and \( \rho' \) with \( \rho' \leq \rho \), condition (10) shows that when \( c_1 = c_H \), playing \( G \) is a dominant strategy for type \( \beta_L \), as long as Assumption 6 is complemented by:

Assumption 8 \( \frac{c_H}{\beta_L} > B - b + \delta (b - a) \).

\[20\]This following assumption rules out compulsion by the weak type (\( \beta_L \) choosing \( P \) even when \( c_1 = c_H \) due to reputational concerns, given that \( b > a \)). Since the insights would be exactly the same as for compulsion by the strong type, we impose the assumption simply to cut down on the number of equilibria.
The only possible equilibrium strategies in period 1 when the W activity is undertaken are thus the four pure behavioral rules described in Section 5.1, together with their four “mixtures”. Since they all differ only through the prescribed actions in the two states \((\beta^H, c_H)\) and \((\beta^L, c_L)\), 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; \(q^H\) and \(q^L\) will of course be functions of \(\rho\).

We are now ready to examine when each rule will emerge as equilibrium behavior, depending in particular on the individual’s self-confidence \(\rho_1\) and cognitive parameters \(\lambda\) and \(\nu\). We shall first present the key results and intuitions in a relatively informal manner (Section 5.4), then state our main propositions characterizing the equilibrium set and its comparative statics (Section 5.5).

5.4 Basic Intuitions: Self–Restraint, Compulsiveness, and Excuses

As can be seen from equations (9)–(10) and the values \(V^H_2, V^L_2\) derived earlier, the maximum net benefits which types \(\beta^L\) and \(\beta^H\) respectively can ever expect from persevering rather than giving up in period 1 are

\[
\begin{align*}
C_L & \equiv B - b + \delta \max\{b - a, 0\} \\
C_H & \equiv B - b + \delta (\phi - a)
\end{align*}
\]

The term \(B - b\) corresponds to the difference in direct payoffs, while the terms in \(\delta\) correspond to the maximum loss in reputation value which playing \(G\) may bring about, if it induces a sure switch from the \(W\) to the \(NW\) task next period (“maximum demoralization”). 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 \(\beta^L\);
- if \(c_H/\beta^H > C_H\), caving in when the cost is \(c_H\) is a dominant strategy for type \(\beta^H\);

It will therefore prove fruitful to divide the \((c_L/\beta^L, c_H/\beta^H)\) plane into four regions, delimited by these two cutoffs; see Figure 3. A number of important intuitions can already be obtained:

- In Region II, where both of the above inequalities hold, the impulsive behavior (rule \(R_0\)) is the unique equilibrium.
- In Region I, both types give in when \(c = c_H\). When \(c = c_L\) the high–willpower type \(\beta^H\) always perseveres, so the key issue is whether the low–willpower type \(\beta^L\) is willing to incur the cost \(c_L/\beta^L\) in order to hide behind (pool with) the \(\beta^H\) type. His motivation to do so would be to induce, or preserve, a decision by next period’s self to choose \(W\) rather than \(NW\). \(^{21}\) Depending on whether the low type finds it preferable to exercise self–restraint because of this reputational concern, or to let himself be “found out” by caving in, the equilibrium will therefore be \(R_2, R_0\).

\(^{21}\) As explained earlier this is valuable for a weak type only when \(b > a\), meaning that, ex ante, some self–control is better than none. Conversely, when \(b < a\) Region I (as well as Region III) vanishes.
Figure 3: perfect inference ($\nu = 1$).

The equilibrium in bold is Pareto–dominant (in Region IV, this requires $b \geq a$).
or $R_{02}$ (in the last case he is just indifferent, and randomizes). In summary, behavior in Region I is quite similar to that studied in Section 4.

- Consider now Region IV, which displays a more novel type of behavior, namely compul-
  siveness. Here the weak-willed individual always caves in, so the key behavior is that of the
  strong-willed one $\beta_H$ in the high-cost state $c_H$. Should he persevere nonetheless, in order to sig-
  nal his type (separate from the weak type), or give up? Depending on his choice, the equilibrium
  will be $R_0, R_3$ or their mixture $R_{03}$.

  When the initial reputation $\rho_1$ is high enough, all three rules can be equilibria, depending on
  Self 2’s expectations (which are self-fulfilling) of what a strong type would do. Indeed, suppose
  that Self 2 “takes no excuses”, interpreting any lapse as revealing the weak type $\beta_L$, even when
  the recall state is $\hat{c} = c_H$. The strong type will then have no choice but to “hang tough” so as
  not to be mistaken for the weak one—which would induce NW rather than W next period. The
  “zero tolerance” rule $R_3$ is thus self-enforcing on the strong type; with the weak type there will
  be a lapse, followed by a complete collapse of self-restraint (e.g., binging) next period.22 Suppose
  now, on the contrary, that Self 2’s interpretation of lapses when plausible excuses exist ($\hat{c} = c_H$) is
  more “forgiving,” in the sense that he expects the strong-willed type to give in whenever $c_1 = c_H$
  (rule $R_0$). Observing a lapse is then less damaging to self-reputation, so if the initial prior was
  high enough the high-willpower type will indeed be able to “relax” when faced with $c_1 = c_H$, without jeopardizing next period’s motivation.23

  Comparing these two self-sustaining rules raises the question of which one is better. From
  an ex-post point of view, the individual with type $\beta_H$ in state $c_H$, is clearly always better off
  when given “the benefit of the doubt” ($R_0$) than when challenged to prove himself ($R_3$). 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 $\beta_H$ does not appear) corresponds well 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 which is out
  of proportion with any current benefits, and thus results in a net cost for the current period –
  possibly even lower intertemporal welfare. To an outside observer who does not properly account
  for self-reputational concerns, the individual appears to be acting as if his $\beta$ was greater than
  1, rather than smaller. In reality, it is precisely the fear that $\beta$ might be too low that gives rise
  to these compulsions. Baumeister et al. (1994, p. 85–86) summarize a similar view commonly
  held by psychologists: “Obsessions and compulsions are attempts to compensate for some self-

  22Baumeister et al. (1994) point out that “zero-tolerance beliefs”, such as “Just Say No” may severely backfire,
  because even minor transgressions of the rule can leads to large collapses of self-esteem and self-regulation. This
  is referred to as “lapse-activated snowballing”.

  23The $W$ decision is always the one he prefers, since $\phi > a$ (whether $b \geq a$).
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.”

As suggested above, for high enough degrees of self-confidence compulsive behavior will only be one of several equilibria, and may thus be avoided. For low enough initial self-confidence, however, $R_3$ will be the only equilibrium in Region IV (e.g., Figure 3), and the individual will not be able to avoid costly self-signalling whose sole purpose is to “reassure himself” of his strength of will. From here on, when speaking of “compulsiveness” we shall primarily be referring to the more interesting case where such behavior is harmful ex-ante (and a fortiori, ex-post), compared to the reputation-free, impulsive, benchmark (which could coexist as another equilibrium, or be sustainable only for different cognitive parameters). That is, we shall be assuming that $c_H > B - b$.

Finally, another important and intuitive result that also accords well with the above quotation is that less reliable inference about motives and circumstances (a low $\nu$) makes compulsiveness more likely. For a person who, in retrospect, cannot easily distinguish between legitimate excuses and illegitimate ones, or who is “too good” at coming up with plausible rationalizations, the flexible, state-contingent rule $R_2$ is more difficult to sustain, and self-restraint can only be achieved at the cost of some compulsion.

5.5 Equilibrium and comparative statics

We now formally solve the intrapersonal signalling game. As one might expect there may be multiple equilibria (three at most), sustained by different, self-confirming interpretations and predictions 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 $W$ activity in period 1, both the strong-willed type $\beta_H$ and the weak-willed type $\beta_L$ are better off if $R$ is played rather than $R'$ (each with its continuation value in period 2).

Finally, while the equilibrium strategies and beliefs can be derived for any value of $\nu \in [0, 1]$ (this is done in the appendix), the exposition will be considerably simplified by focussing here on the polar cases $\nu = 1$ and $\nu = 0$, which convey the main intuitions. In particular, contrasting the behaviors and welfare implications that emerge in these two cases will bring into sharp focus the cognitive foundations of personal rules.
5.5.1 Perfect Inference \((\nu = 1)\)

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

**Proposition 2** For \(\nu = 1\), the set of equilibria in period 1 is described by Figure 3. Multiple equilibria may arise for \(\rho_1 > \rho^*_2\), in which case the Pareto–dominant equilibrium is indicated in bold.

In Region I, where \(c_H/\beta_H > C_H\) but \(c_L/\beta_L < C_L\), the equilibrium is identical to that of the one cost case (with \(\lambda \equiv 1\)), described in Proposition 1. Complete self–control is achieved for \(\rho > \rho^*_2\), while for \(\rho < \rho^*_2\) the agent persists only with a probability \(q^L\) which is such that the posterior following an observation of \(P\) is exactly equal to \(\rho^*_2\). Indeed, with \(c_H/\beta_H > C_H\), both types give in when \(c_1 = c_H\). Since this high cost realization can be perfectly distinguished from that of \(c_1 = c_L\), the whole problem in Regions I and II reduces to that of one cost, \(c = c_L\).

In Region IV we obtain the potential for compulsiveness, as informally argued earlier. For low values of \(\rho_1\) the individual does not give himself the “benefit of the doubt” even when the recall state is \(\hat{c} = c_H\), so the \(\beta_H\) type must use adverse circumstances to “prove himself”. For \(\rho_1 > \rho^*_2\), on the other hand, \(R_0\) is an alternative feasible equilibrium. Furthermore this “forgiving” equilibrium is clearly preferred by type \(\beta_H\) to \(R_3\). It is also preferred by type \(\beta_L\) whenever \(b > a\), which is the case illustrated on Figure 3. 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.

Region III, finally, entails a mixture of 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 own resolve, on the other hand, compulsiveness will be the price to pay today for securing some of the benefits of future self–restraint (rule \(R_{13}\)).

5.5.2 Unreliable Inference \((\nu = 0)\)

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 (thus maintaining “plausible deniability”, as in covert operations).

**Proposition 3** The set of equilibria in period 1 is described by Figure 4. Multiple equilibria may arise for \(\rho > \rho_1\), in which case the Pareto–dominant equilibrium is indicated in bold.
Figure 4: unreliable inference ($\nu = 0$).
The equilibrium in bold is Pareto–dominant
(in Region IV, this requires $b \geq a$).
The two new thresholds which appear on Figure 4 are defined as follows:

\[ \rho_1 = \frac{\rho_2^+}{\rho_2^+ + (1 - \pi)(1 - \rho_2^+)} > \rho_2^* > \frac{\rho_2^+}{\rho_2^+ + (1 - \rho_2^+)/\pi} \equiv \bar{\rho_1}. \]

The higher threshold is such that, for \( \rho_1 > \bar{\rho_1} \), a lapse when the recall state is \( \hat{c} = c_H \) (plausible excuses exist) will never bring reputation below \( \rho_2^* \), if it is expected that the strong–willed type gives in whenever \( c_1 = c_H \). The lower threshold is such that, for \( \rho_1 < \bar{\rho_1} \), perseverance when the recall state is \( \hat{c} = c_H \) will never bring reputation above \( \rho_2^* \), if it is expected that the weak–willed type perseveres whenever \( c_1 = c_L \).

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

- One of the most important 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 II and III). The only rules which 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: \( R_{13} \). This general lesson is perhaps most apparent in Region III.
• In Region I, the fall in $\nu$ leads to a general *weakening of self-restraint*. As illustrated on Figure 5, the loss of is more drastic the higher the initial self-confidence $\rho_1$, so that self-restraint eventually *decreases* toward zero as initial self-reputation improves. The intuition is that when the “self-monitoring technology” is very ineffective, the weak type can take full advantage of the “principal’s” (Self 2’s) high level of trust $\rho_1$, and misbehave without any adverse consequences. When the initial reputation is low, however, Self 2 applies a tougher standard when deciding whether or not to attempt the $W$ activity in period 2, so the weak type must actually “work at” passing for a strong one; he will therefore persevere with positive probability in the low-cost state $c_L$ (when $b > a$).

• In Region IV, by contrast, less reliable inference leads to a *tightening of self-restraint*, or *increased compulsiveness*. As shown on Figure 6, this occurs when the initial reputation is in the intermediate range, $\rho_1 \in [\rho_2^*, \bar{\rho}_1]$. With $\nu = 0$ (a low value of $\nu$ more generally), observing a lapse $G$ together with “plausible excuses” $\hat{c} = c_H$ is worse news about one’s willpower $\beta$ than in the earlier case of $\nu = 1$, because there is now a chance that the true $c_1$ was actually $c_L$. Distrusting these excuses, the individual is then tougher on himself, demanding more “proof” that willpower is indeed high. Consequently the “forgiving” equilibrium $R_0$ now exists only for $\rho_1 > \bar{\rho}_1$, as
opposed to $\rho_1 > \rho_2^*$. For $\rho_1 \in [\rho_2^*, \bar{\rho}_1]$ the compulsive rule $R_3$ prevails in its place.

This leads to another interesting result: the individual may be better off if he could forget his lapses altogether, rather than remember them but not being able to reliably ascertain whether or not there was a valid excuse or extenuating circumstances. Indeed, with $\lambda = 0$ the only equilibrium is always $R_0$, which (given $c_H > B - b$) is always better than $R_3$, both ex ante and ex post. In other words, for an individual with intermediate degrees of confidence in his willpower, blissful ignorance of his past failings may be preferable to imperfect or manipulable self-monitoring and analysis.

6 Cooperation in short–run relationships

Our analysis may also shed light on an old puzzle in economics: why are people often “nice” (tipping, doing favors, restraining their emotions when annoyed, etc.) to others whom they will never meet again? In the absence of gains from repeated interaction, the standard explanation is some form of altruism. While certainly plausible, this rationale goes only so far, as it does not elucidate the factors giving rise to such behavior. For instance, it does not explain why the people one interacts with at random are the primary beneficiaries of such altruistic behavior—as opposed to, say, starving third-world children.

We propose here an alternative (and complementary) hypothesis, namely that such punctual cooperation is actually an attempt by the individual to raise his own, selfishly defined, welfare. An important determinant of success in societal life is one’s ability to consistently cooperate with family members, workmates, neighbors or other people one interacts with. However, the benefits from cooperation in such long-term relationships are often jeopardized by one’s salient short-term interests or emotions (temptation to cheat, jealousy, anger, etc.) which induce opportunistic or impulsive behavior toward others. It is therefore important for an individual to have confidence in his power of will with respect to social interactions: only if he feels that he can sustain a good relationship in the future will he find the energy to start or maintain one today. Viewed from this perspective, cooperation in one-shot relationships is a form of self-restraint carried out primarily for self-signalling purposes. Ceteris paribus, the individual would prefer to reserve cooperation for those with whom he will interact again, but his concern for self-reputation makes him extend the scope of such behavior to short-term relationships.

A simple variant of our model can be used to formalize this idea. Consider an individual who is engaged in a one-shot encounter, but knows that later on he will have opportunities to engage in (or abstain from) more durable relationships. We represent the short-term interaction

\footnote{Failure to cooperate will likely be detected sooner or later, and lead to the withdrawal of cooperation by others, if not to more severe forms of retaliation.}
as stage II ("afternoon") of the first period of our basic game, with perceived payoffs $-c_1/\beta + B_1$ for cooperation ($P$) and $b_1$ for defection ($G$). One should typically think of $B_1$ and $b_1$ as small, possibly even equal to zero. The long-run relationship is, for now, represented in a reduced form by the two stages ("full day") of period 2 in our basic game: first, the agent decides whether or not to "invest" in the relationship ($W$ or $NW$), foregoing the payoff $a/\gamma$. If he chooses $W$, later on he is faced with the usual choice of whether to behave opportunistically or not: $P$ or $G$, with contemporaneously perceived payoffs $-c_2/\beta + B_2$ and $b_2$, such that $B_2 - c_2 > a/\gamma > b_2$ (Assumption 6). Thus, for the moment we take it as a given that sustained cooperation in a long-run relationship pays off; one may imagine for instance that the partner plays a tit-for-tat strategy, with a delay between each player’s actions. The individual faces the same imperfect information and recall problems as before concerning his willpower $\beta \in \{\beta_L, \beta_H\}$.

One can then understand cooperation in short-run relationships as a purely compulsive form of behavior. Indeed, let $c_1/\beta_H > B_1 - b_1$, so that no one enjoys cooperating per se. Weak—willed types $\beta_L$ will always defect, but strong—willed types $\beta_H$ may still want to incur the costs of short—term cooperation, for fear that not doing so will undermine the favorable self—image required to invest in long-run relationships.\footnote{This separating equilibrium, similar to the compulsive rule $R_3$ seen earlier, occurs when $c_1/\beta_H < B_1 - b_1 + \delta(B_2 - c_2 - a)$ but $c_1/\beta_L > B_1 - b_1 + \delta(b_2 - a)$.}

Finally, the ad-hoc assumption that the partner in the long-run relationship plays tit-for-tat responses can easily be endogenized by “unfolding” the second stage of the game into a full—fledged repeated interaction. Thus: a) the individual faces the $P/G$ choice at all dates $\tau = 1, 2, \ldots, T$ where the relationship is still ongoing; (b) at the end of each date $\tau$, the partner has the choice between staying and quitting (forever, without loss of generality). In equilibrium, the partner stays until he observes defection. Under appropriate parameter configurations, the equilibrium of this game is for the strong type to cooperate (play $P$) forever and for the weak type to defect (play $G$) at the first opportunity.

The self—reputation perspective on cooperation may also help explain people’s apparently odd preferences concerning the beneficiaries of their generosity. We are more likely to be reminded of the face or complaint of the taxi driver whom we did not tip, the lost tourist we did not help out, or the beggar we passed by, than of the “disappointment” of an anonymous starving child (indeed, NGO campaigns are often aimed at remedying precisely this point).\footnote{See Rabin (1995) for a related point in a model where agents derive direct utility from their beliefs, and Bodner and Prelec (1997) for a model where agents with metapreferences over their own tastes (e.g., generosity, honesty) engage in self—signalling.} Consequently, the absence of cooperation is much more of a threat to self-reputation in the former case than in the latter.

\footnote{This separating equilibrium, similar to the compulsive rule $R_3$ seen earlier, occurs when $c_1/\beta_H < B_1 - b_1 + \delta(B_2 - c_2 - a)$ but $c_1/\beta_L > B_1 - b_1 + \delta(b_2 - a)$.}
7 Cognitive Rules

7.1 Emotion and Attention Control

Faced with the potentiality of a weak will, the individual responds with a variety of cognitive strategies. The first two, emotion and attention control, have very broad applicability, in that they represent reactions to the self–regulation problem in general, and not necessarily related to self-reputation concerns. The third one, which we shall terms “resolutions”, is directly linked to the monitoring of lapses and self–learning.

*Emotion control* refers to strategies of training oneself, in advance, to care less about the impulsive urge or emotion, or even associate negative images with impulsive actions (e.g., cigarettes and fatty foods with visions of diseased lungs and clogged arteries), and positive images with delayed-gratification actions (receiving praise or awards, achieving fame, etc.). While it is clear how emotion control can benefit a time-inconsistent individual by reducing the probability of lapses, it may also have negative consequences. First, it makes lapses less enjoyable (spoils the fun); to the extent that giving in is sometimes desirable or unavoidable, this has a direct welfare cost. More subtly, emotion control could have the same impact as external constraints: it may deter specific types or instances of impulsive behavior, but without letting the individual develop a self-reputation by putting his will to the test. Of course, the lack of development of a self-reputation may not matter if the individual is able to implement emotion control uniformly over time and activities; but he may be able to alter his affect only for some activities or at some points of time.

*Attention control* (or, following Freud, “repression”) is related to emotion control, in that it affects the individual’s perception of one of the alternatives. It consists in thinking about something else or distracting oneself when starting to think about a tempting activity (action G).

As described by Ainslie (1992) “When deciding whether or not to pursue a given activity, a person does not call up all his knowledge at once, but begins with a label with which he has categorized that knowledge; the aspect he is likely to remember first is its emotional meaning. He is apt to evaluate his options for further information processing according to the likely payoffs for those options.

Baumeister et al. (1994, p. 85) testify to the importance of attention control as a robust findings of the literature on self–regulation: “In all spheres of self–regulation –controlling emotion, appetites and desires, performances, thought processes, and the rest– the management of atten-

---

27 This strategy can affect choices in at least two ways. First, it may alter the individual’s perceptions or awareness of the payoffs associated with different choices. Alternatively, it can directly affect payoffs; for example, the image of a hungry child in the third world, or of the animal slaughter process, may reduce the individual’s payoff from eating a particular dish, even when not eating it will not do anything to solve the underlying problem.
tion emerged as a significant factor. Not only did it seem ubiquitous, but it also seemed widely effective.”

7.2 Resolutions

From now on, we will narrow down the discussion of cognitive rules to strategies that consists in setting “ex ante” desirable modes of behavior for oneself. These behavioral targets or resolutions may be specific (“I never eat candy,” “I will not start eating before others even when I am hungry,” “I finish every exercise routine which I start”), or define a broader mandate (“I am on a diet,” “I behave politely,” “I stick to my choices”). They may even have a universal character. For example, Weber viewed the Calvinist philosophy as a single, comprehensive side-bet aimed at systematic self-control.28

Viewed a bit cynically, the practice of religion and the rehearsal of philosophical or moral precepts are partly aimed at solving the individual’s time-inconsistency problem in social relationships. Humans realize that a key to their individual welfare is the ability to cooperate with family members, workmates, bosses, neighbors, and other people with whom they interacts. The benefits from cooperation, however, are jeopardized by individuals salient short-term interest, which dictates opportunistic behavior or makes it difficult to control one’s emotions. Religions, philosophies and morals all attempt to counteract such impulsive behaviors that compromise the long-term benefits from a cooperative relationship.

Of course, resolutions do not only serve to restrain behaviors in social contexts. Their primary aim is intrapersonal self-discipline. Indeed, in a reinterpretation of our model, the ex—ante and ex—post manipulations of awareness interact in an interesting way. Consider for instance Darwin’s personal cognitive rule:

“...I had during many years followed the Golden Rule, namely, that whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without fail and at once; for I had found by experience that such (contrary and thus unwelcome) facts and thoughts were far more apt to escape from memory than favorable ones.”


Evidently, the failure to record contrary evidence constituted a lapse, and Darwin’s personal rule of keeping written records amounted to raising the probability (λ) that he would later on remember this lapse. He was concerned that his selective memory (whether due to the desire

---

28See Ainslie (1992, p. 203) for a discussion, and Bodner and Prelec (2001) for a model based on multiple intratemporal selves.
to avoid writing down unpleasant things, or to an ex–post desire to maintain self-esteem, as in Bénabou and Tirole 2000a), while serving his short-term interests, would reduce his long-term welfare.

7.3 Why Do Resolutions Matter?

An individual cannot just select a good behavioral rule; we saw that for given memory parameters \((\lambda, \nu)\) only certain ones are self–enforcing. What he can affect, to some extent, are his cognitive processes. In the same way the he may “ex post” (at the end of date 1 in our model) be able to alter the probability that a known information (lapse, excuse or lack thereof), is later on brought back to awareness, he may also “ex ante” (at date 0 in our model) be able to make it more likely that certain types of future events will be remembered.

As noted earlier, people who fail at self–regulation are often those with poor self–monitoring, and conversely much of behavioral therapy involves endowing the patient with a more reliable, less manipulable “technology” for monitoring his actions, such as keeping a journal or talking regularly with the therapist. Twelve–step programs and weekly confession in church are also, in large part, self–monitoring devices. Neither his peers in the group nor the priest in the confessional can (or even try) to force the alcoholic or sinner to refrain from his vice (contrast this to, say, a rehabilitation clinic). Admittedly there is some social or moral pressure, but the individual could avoid it by lying about his recent drinking or bad actions. He may perceive some probability of being found out and condemned by his peers, or punished by divine intervention, but the direct and inevitable effect of participating in such “pure talk” sessions is that the individual is forced to think about his own behavior; even if he chooses to misrepresent it.

Other cognitive strategies include selecting a candidate rules with good mnemonic properties (e.g., only one cigarette after each meal, rather than just three per day) and rehearsing them often (e.g., religious principles) so as to make lapses more salient (raising \(\lambda\)) or more difficult to rationalize away (raising \(\nu\)). For instance, a resolution of “always sticking to one’s choices” will make any lack of perseverance more memorable. In this respect, a resolution is similar to a mnemonic device like a journal: both create another, “redundant” channel of activation of the lapse-related memory.\(^{29}\)

Cognitive rules can also raise the probability of future awareness through the following mechanisms: a) the initial encoding and future rehearsal of the information that action \(G\) has been taken is more likely to be effective if it is not passive and rather involves a “deep-level judgement” –that is, “considering the meaning rather than the surface form of stimuli improves memory for the stimuli” (Anderson (2000a), p. 198); b) The future consideration of the broader rule may, by

\(^{29}\)See Anderson (2000b, p212) for a discussion of redundancy in memory processes.
associativeness, force the retrieval of the particular action and thereby improve memory; c) cognitive rules may facilitate retrieval by helping the individual search his memory more efficiently (by “organizing”) material.

7.4 Incomplete Self-Contracts

When formulating a cognitive rule (consciously or unconsciously), the individual often faces a trade-off between making it precise (“I do not eat candy,” “I will stick to my diet except for two days a year: Thanksgiving and my spouse’s birthday”, or loose (“I will try to stay thin,” “I will follow the diet except on special occasions”).

The benefit of a precise rule is that it limits the scope for excuses, and thus encourages self-control. In our model, what constitutes a lapse is more likely to be remembered if the rule is not subject to ex post, self-serving reinterpretations. Precise rules may also prevent compulsive behavior, in the same way as an increase in \( \nu \) can eliminate it in our model. Of course, a rule will never be precise enough to account for all possible circumstances. In the context of our model, the individual should persevere “on average” in a given state. But there may be undescribed substates in which caving in would be desirable (the compliance cost is \( c_H \)). Because these circumstances have not been described earlier, caving in (action \( G \)) is a violation of the letter of the rule, even though it may comply with its spirit.

The cost of a precise rule is of course the ex-ante cost of thinking through and rehearsing a complex, state-contingent behavioral prescription. A “bright-line” rule can be viewed as one that tries to combine simplicity and precision. Thus “not smoking at all” would probably be suboptimal for most people if they could cheaply set precise state-contingent rules for their smoking behavior. Yet they often choose not to smoke at all, behavior which is both precise and simple to monitor.

7.5 Universality vs. Lapse Districts: the “Multimarket Contact” Tradeoff

A lapse in one activity is likely to spill over to another activity, because of its impact on self-reputation. To some extent, this linkage between activities amounts to increasing the concern for self-reputation (a higher \( \delta \) in our model). Because lapses in one activity are treated as precedents for another, such “multimarket contact”\(^30\) is conducive to both self-restraint and compulsion. The self-restraint effect argues in favor of universality, while the compulsive effect calls for keeping activities separate and lapses confined to “lapse districts”\(^31\).

But it is by no means easy to create lapse districts in order to avoid excessive spillovers of lapses from certain dimensions of one’s life onto other ones. After all, something has been learned

\(^{30}\)Economic analyses of the behavior of a firm when the latter’s action in one market affects its reputation in another market include Bernheim-Whinston (1990) and Fudenberg-Kreps (1987).

\(^{31}\)The term is borrowed from Ainslie (1999), who uses at an analogy the “vice districts” tolerated in most cities.
about the power of one’s will, that is generally relevant for other decisions.32 Again, awareness
management may be of some help here. In order not to draw from his failure to exercise self-
restraint in activity $A$ the inference that his $\beta$ is low, leading to dim prospects for activity $B$, the
individual may try to convince himself that his cost of perseverance in activity $A$ is especially
high. That is, if we extend our model to allow the distribution of cost $c$ in a given activity to be
imperfectly known to the individual, the individual may rehearse any piece of information that
has the effect of shifting the posterior distribution about $c$ toward higher values (in the sense of
first-order stochastic dominance). Such an awareness strategy amounts to admitting that one is
“hopeless” in that dimension (in the same way as alcoholics are advised by Alcoholics Anonymous
to regard themselves as “helpless against alcohol”), thus creating a lapse district in order to avoid
propagation to other activities.

8 Conclusion

---

32Baumeister et al. (194, p. 11) state that “self-regulatory capacity is a central, powerful, stable, and beneficial aspect of personality,” then describe research which has shown that children’s capacity to delay gratification is significantly correlated with a host of important outcomes and attitudes late in life, such as resourcefulness, cooperativeness, ability to deal with stress, etc.
APPENDIX

Proof of Proposition 1

Consider first the weak type’s decisions when confronted with the W activity, and let \( q_1 \) denote his probability of perseverance.

Case 1: \( q_1 = 1 \). This implies that \( \rho_2^+ = \rho_1 \), while \( \rho_2^- \) can be any \( \rho' \leq \rho \). Optimality in (4) 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.

Case 2: \( q_1 \in (0, 1) \). This implies \( \rho_2^+ \in (\rho_1, 1) \) and \( \rho_2^- = 0 \). Furthermore, (4) must now hold with equality,

\[
\frac{c}{\beta_L} = B - b + \delta \lambda \left[ V^L_2(\rho_2^+) - a \right].
\]  

which can only occur if

\[
\rho_2^+ \equiv \frac{\rho_1}{\rho_1 + (1 - \rho_1)(q_1 + (1 - q_1)(1 - \lambda))} = \rho_2^*,
\]

requiring \( \tilde{\rho}_1(\lambda) < \rho_1 < \rho_2^* \), and if the mixing probability \( \rho_2^* = p_2(\rho_2^*) \) that will result in period 2 satisfies:

\[
\frac{c}{\beta_L} = B - b + \delta \lambda p_2^*(b - a).
\]

These conditions uniquely determine \( q_1 \) and \( \rho_2^* \) in \([0, 1]\) as given in Proposition 1.

Case 3: \( q_1 = 0 \). This implies again that \( \rho_2^- = 0 \), and thus one must have \( c/\beta_L \geq B - c + \delta \left[ V^L_2(\rho_2^+) - a \right] = V^L_2(\rho_2^+) - a \). 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 now 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 task \( W \). Indeed, this yields \( B - c \) in period 1 and \( \delta [\rho_1(B - c) + (1 - \rho_1)b] \) in period 2, against \( a/\gamma \) in period 1 and the same expected payoff in period 2 if the NW is chosen instead (there is then no new information, \( \rho_2 = \rho_1 \), so W is chosen in period 2). Consider now the case where \( \tilde{\rho}_1(\lambda) < \rho_1 < \rho_2^* \). Choosing \( W \) rather than NW now leads to expected net gains of \( \Delta_1 \) in period 1 and \( \Delta_2 \) in period 2, where:

\[
\Delta_1 \equiv \rho_1(B - c - a/\gamma) + (1 - \rho_1) [q_1(B - c) + (1 - q_1)b - a/\gamma]
\]

is increasing in \( \rho_1 \), both directly and through \( q_1 \), and the same is true for
\[ \Delta_2 / \delta = \rho_1 [p^*_2 (b - c) + (1 - p^*_2) a] + (1 - \rho_1) \{ [q_1 + (1 - q_1) (1 - \lambda)] [p^*_2 b + (1 - p^*_2) a] + (1 - q_1) \lambda a \} - a \]
\[ = p^*_2 \{ \rho_1 (b - c - a) + (1 - \rho_1) [q_1 + (1 - q_1) (1 - \lambda)] (b - a) \}. \]

By continuity, the total gain \( \Delta_1 + \Delta_2 \) 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 \( \rho^*_1 \) is greater or smaller than the threshold \( \rho_1 = \tilde{\rho}_1 (\lambda) \) where \( q_1 = 0 \). 

**Bayesian Updating in the Two-Cost Case**

Let us denote as \( q^i (\rho, c) \) the perseverance strategy (probability of playing \( P \)) of each type \( i = H, L \) when confronted with cost \( c \in \{ c_H, c_L \} \) in the \( W \) activity in period 1, and given prior beliefs \( \rho \). Following an observation of \( c_H \) in period 1, Bayes’ rule implies:

\[
\frac{\hat{\rho}_2^+}{1 - \hat{\rho}_2^+} = \left( \frac{\rho}{1 - \rho} \right) \left( \frac{1 - \pi) q^H (\rho, c_H) + \pi (1 - \nu) q^H (\rho, c_L) \right),
\]

(A.2)

\[
\frac{\hat{\rho}_2^-}{1 - \hat{\rho}_2^-} = \left( \frac{\rho}{1 - \rho} \right) \left( \frac{1 - \pi) (1 - q^H (\rho, c_H) + \pi (1 - \nu) (1 - q^H (\rho, c_L))} \right).
\]

(A.3)

Similarly, following an observation of \( c_L \):

\[
\frac{\rho_2^+}{1 - \rho_2^+} = \left( \frac{\rho}{1 - \rho} \right) \left( \frac{q^H (\rho, c_L)}{q^L (\rho, c_L)} \right),
\]

(A.4)

\[
\frac{\rho_2^-}{1 - \rho_2^-} = \left( \frac{\rho}{1 - \rho} \right) \left( \frac{1 - q^H (\rho, c_L)}{1 - q^L (\rho, c_L)} \right).
\]

(A.5)

These expressions can be simplified once it has been shown that \( q^H (\rho, c_L) = 1 \) and \( q^L (\rho, c_H) = 0 \) are dominant strategies. In particular, \( \rho_2^- = 0 \). 

**Proof of Propositions 2 and 3**

We derive the necessary and sufficient conditions under which rule can be sustained in equilibrium, for the general case \( \nu \in (0, 1) \). We then let \( \nu \) tend 0 and to 1 in the formulas, which yields the results stated in the text.

**1) When is \( R_0 \) an equilibrium in period 1?**

Under \( R_0 \) the updating rules imply \( \hat{\rho}_2^+ = \hat{\rho}_2^- = 1 \), \( \rho_2^- = 0 \) and

\[
\frac{\hat{\rho}_2^-}{1 - \hat{\rho}_2^-} = \left( \frac{\rho_1}{1 - \rho_1} \right) \chi,
\]

39
where
\[
\chi \equiv \frac{1 - \pi}{1 - \pi \nu} = \Pr [c = c_H | \hat{c} = c_H \text{ was recalled}] \quad \text{(A.6)}
\]
represents the “reliability” or “credibility” of ex–post excuses. The optimality conditions (9)–(10), together with the previously computed values of \(V^i_2\), now require that:
\[
\begin{align*}
\frac{c_H}{\beta_H} &\geq B - b + \delta (\phi - V^H_2(\hat{\rho}^-)) \\
\frac{c_L}{\beta_L} &\geq B - b + \delta \nu (b - a) + \delta (1 - \nu) (b - V^L_2(\hat{\rho}^-))
\end{align*}
\]
Let us therefore define \(\bar{\rho}_1\) as the value of \(\rho_1\) which leads to the posterior \(\hat{\rho}^- = \rho^*_2\), that is:
\[
\bar{\rho}_1 \equiv \frac{\rho^*_2}{\rho^*_2 + (1 - \rho^*_2)\chi}. \quad \text{(A.7)}
\]
Note that \(\bar{\rho}_1 > \rho^*_1\), and that \(\bar{\rho}_1\) is decreasing in \(\chi\). The two equilibrium conditions are met when either:

a) \(\rho_1 < \bar{\rho}_1\) and
\[
\begin{align*}
\frac{c_H}{\beta_H} &\geq B - b + \delta (\phi - a) = C_H \\
\frac{c_L}{\beta_L} &\geq B - b + \delta (b - a) = C_L. \quad \text{(A.8)}
\end{align*}
\]

b) \(\rho_1 > \bar{\rho}_1\) and
\[
\frac{c_L}{\beta_L} \geq B - b + \delta \nu (b - a) \quad \text{(A.10)}
\]

- For \(\nu = 0\) we find therefore that \(R_0\) is an equilibrium in all of regions I to IV for \(\rho_1 > \bar{\rho}_1\), and in region IV for all every value of \(\rho_1\). For \(\nu = 1\) note that \(\chi = 1\), implying \(\bar{\rho}_1 = \rho^*_2\). Consequently, \(R_0\) is an equilibrium only in regions II (for \(\rho_1 > \bar{\rho}_1\)) and IV (for any \(\rho_1\)).

2) When is \(R_1\) an equilibrium in period 1?

Under \(R_1\) the updating rules imply \(\rho^+_2 = \rho_1\), \(\rho^-_2 = \text{any} \; \rho' \leq \rho_1\), \(\hat{\rho}^- = 0\) and
\[
\frac{\hat{\rho}^+_2}{1 - \hat{\rho}^+_2} = \left(\frac{\rho_1}{1 - \rho_1}\right) \left(\frac{1}{1 - \chi}\right),
\]
where $\chi$ was defined in (A.6). The equilibrium conditions (9)–(10) now take the form:

$$\frac{c_H}{\beta_H} \leq B - b + \delta (V^H_2(\hat{\rho}_2^+) - a)$$

$$\frac{c_L}{\beta_L} \leq B - b + \delta\nu (V^L_2(\rho_1) - V^L_2(\rho')) + \delta(1 - \nu) (V^L_2(\hat{\rho}_2^+) - a)$$

Given Assumption 6, the first condition requires that $c_H/\beta_H \leq B - b + \delta (\phi - a) = C_H$ and $\hat{\rho}_2^+ \geq \rho_2^*$. Define therefore $\rho_1$ as value of $\rho_1$ which leads to the posterior $\hat{\rho}_2^+ = \rho_2^*$, that is,

$$\rho_1 \equiv \frac{\rho_2^*}{\rho_2^* + (1 - \rho_2^*)(1 - \chi)}.$$  \hspace{1cm} (A.11)

Note that $\rho_1 < \rho_2^*$, and that $\rho_1$ is decreasing in $\chi$. We must have $\rho_1 > \rho^*_1$, so the second equilibrium condition takes the form:

$$\frac{c_L}{\beta_L} \leq B - b + \delta\nu (V^L_2(\rho_1) - V^L_2(\rho')) + \delta(1 - \nu) (b - a).$$  \hspace{1cm} (A.12)

For $\rho_1 > \rho_2^*$, it can be met with $\rho' \leq \rho_1$ as long as

$$\frac{c_L}{\beta_L} \leq B - b + \delta (b - a) = C_L,$$  \hspace{1cm} (A.13)

For $\rho_1 \in (\rho^*_1, \rho^*_2)$ the second term in (A.12) is zero, so the requirement becomes:

$$\frac{c_L}{\beta_L} \leq B - b + \delta(1 - \nu) (b - a).$$  \hspace{1cm} (A.14)

To summarize, first it must be that $c_H/\beta_H \leq C_H$. Second, when $c_L/\beta_L < B - b + \delta(1 - \nu) (b - a)$ this equilibrium exists for all $\rho \in (\rho^*_1, 1)$; when $B - b + \delta (1 - \nu) (b - a) < c_L/\beta_L < B - b + \delta (b - a)$, it exists for all $\rho \in (\rho^*_2, 1)$. In all other cases it does not exist.

- In particular, when $\nu = 0$ the equilibrium exists only in region III, for $\rho_1 > \rho^*_1$. When $\nu = 1$, implying $\rho_1 = 0$ it exists in region III for $\rho > \rho^*_2$.

3) **When is $R_2$ an equilibrium in period 1?**

Under $R_2$ the updating rules imply $\rho_2^+ = \rho_1$, $\rho_2^- = \rho_1$, and $\hat{\rho}_2^+ = \hat{\rho}_2^- = \rho_1$. The equilibrium conditions (9)–(10) now take the form $c_H/\beta_H \geq B - b$, which always holds, and

$$\frac{c_L}{\beta_L} \leq B - b + \delta\nu (V^L_2(\rho_1) - V^L_2(\rho')).$$

This requires that $\rho_1 > \rho_2^* > \rho'$; since $\rho' \leq \rho_1$ is unconstrained, only the first of these two
inequalities matters. Finally, it must be that:

\[
\frac{c_L}{\beta_L} \leq B - b + \delta \nu (b - a). \tag{A.15}
\]

• With \( \nu = 0 \), \( R_2 \) is therefore never an equilibrium. With \( \nu = 1 \), it is an equilibrium for \( c_L/\beta_L \leq C_L \) (regions I and III), provided that \( \rho_1 > \rho_2^* \).

4) When is \( R_3 \) an equilibrium in period 1?

Under \( R_3 \) the updating rules imply \( \rho_2^+ = \hat{\rho}_2^+ = 1 \), \( \rho_2^- = \hat{\rho}_2^- = 0 \). The equilibrium conditions (9)–(10) now take the form:

\[
\frac{c_H}{\beta_H} \leq B - b + \delta (\phi - a) = C_H \tag{A.16}
\]

\[
\frac{c_L}{\beta_L} \geq B - b + \delta (b - a) = C_L. \tag{A.17}
\]

• Thus, whether for \( \nu = 0 \) or \( \nu = 1 \), \( R_3 \) is an equilibrium in region IV, for all values of \( \rho_1 \).

5) When is \( R_{02} \) an equilibrium in period 1?

Under \( R_{02} \) the updating rules imply \( \rho_2^- = 0 \) and

\[
\frac{\rho_2^+}{1 - \rho_2^+} = \frac{\hat{\rho}_2^+}{1 - \hat{\rho}_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1}{q^L} \right)
\]

\[
\frac{\hat{\rho}_2^-}{1 - \hat{\rho}_2^-} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1 - \pi}{1 - \pi + \pi (1 - \nu) (1 - q^L)} \right)
\]

The equilibrium conditions (9)–(10) now take the form:

\[
\frac{c_H}{\beta_H} \geq B - b + \delta \left( V_2^H (\hat{\rho}_2^+ - V_2^H (\hat{\rho}_2^-)) \right)
\]

\[
\frac{c_L}{\beta_L} = B - b + \delta \nu \left( V_2^L (\rho_2^+ - a) + \delta (1 - \nu) (V_2^L (\rho_2^+ - V_2^L (\hat{\rho}_2^-)) \right)
\]

The second condition cannot be an equality (except with measure zero) unless either \( \rho_2^+ \) or \( \hat{\rho}_2^- \) is equal to \( \rho_2^* \).

Case 1: \( \rho_2^+ = \rho_2^* \), which uniquely defines \( q^L \) as long as \( \rho_1 < \rho_2^* \). The equilibrium conditions become:

\[
\frac{c_H}{\beta_H} \geq B - b + \delta p_2 (\rho_2^*) (\phi - a)
\]

\[
\frac{c_L}{\beta_L} = B - b + \delta p_2 (\rho_2^*) (b - a)
\]
Abbreviating $p_2(\rho_2^*)$ as $p_2^*$, the second condition yields

$$p_2^* = \frac{c_L/\beta_L - B + b}{\delta(b-a)},$$
so the equilibrium requirements finally become:

$$\frac{c_L}{\beta_L} \leq B - b + \delta (b - a) = C_L, \quad \text{(A.19)}$$
$$\frac{c_H}{\beta_H} \geq B - b + \left(\frac{c_L}{\beta_L} - B + b\right) \left(\frac{\phi - a}{b-a}\right). \quad \text{(A.20)}$$

In the $(c_L/\beta_L, c_H/\beta_H)$ plane, boundary for the latter inequality is the line $L_1$, with slope 1, that goes from the point $(B - b, B - b)$ to the point $(B - b + \delta (b - a), B - b + \delta (\phi - a))$.

**Case 2:** $\hat{\rho}_2^2 = \rho_2^*$, which by the updating rules uniquely defines $q^L$ as long as

$$\rho_2^* < \rho_1 < \frac{\rho_2^*}{\rho_2^* + (1 - \rho_2^*)\chi} = \bar{\rho}_1. \quad \text{(A.21)}$$

The conditions then become:

$$\frac{c_H}{\beta_H} \geq B - b + \delta (1 - p_2^*) (\phi - a)$$
$$\frac{c_L}{\beta_L} = B - b + \delta [\nu + (1 - \nu) (1 - p_2^*)] (b-a).$$

The latter yields:

$$1 - p_2^* = \left(\frac{c_L/\beta_L - B + b}{\delta(b-a)} - \nu\right) \left(\frac{1}{1 - \nu}\right), \quad \text{(A.22)}$$
as long as

$$B - b + \nu \delta (b-a) < c_L/\beta_L < B - b + \delta (b-a) = C_L. \quad \text{(A.23)}$$

The first condition then requires:

$$\frac{c_H}{\beta_H} \geq B - b + \left(\frac{c_L/\beta_L - B + b - \nu \delta (b-a)}{1 - \nu}\right) \left(\frac{\phi - a}{b-a}\right) \quad \text{(A.24)}$$

In the $(c_L/\beta_L, c_H/\beta_H)$ plane, the boundary for the second one is the line $L_2$, with slope $1/(1 - \nu)$, that goes from the point $(B - b + \nu \delta (b-a), B - b)$ to the point $(B - b + \delta (b-a), B - b + \delta (\phi - a))$.

- For $\nu = 0$, $R_{02}$ therefore exists in regions I and III$^+$ for $\rho_1 < \rho_2^*$ (Case 1) as well as for
$\rho_2^* < \rho_1 < \tilde{\rho}_1$ (Case 2), and thus for all $\rho_1 < \tilde{\rho}_1$. For $\nu = 1$, in which case $\tilde{\rho}_1 = \rho_2^*$, it exists in regions I and III+ for $\rho_1 < \rho_2^*$ (Case 1).

6) When is $R_{03}$ an equilibrium in period 1?

Under $R_{03}$ the updating rules imply $\rho_2^+ = 1$, $\rho_2^- = 0$, $\hat{\rho}_2^+ = 1$ and

$$\hat{\rho}_2^- = \frac{\rho_1}{1 - \rho_1} \left( \frac{(1 - \pi) (1 - q^H)}{1 - \pi + \pi (1 - \nu)} \right) = \chi \left( 1 - q^H \right) \left( \frac{\rho_1}{1 - \rho_1} \right).$$

The equilibrium conditions (9)–(10) now take the form:

$$\frac{c_H}{\beta_H} = B - b + \delta \left( \phi - V_2^H (\hat{\rho}_2^-) \right)$$

$$\frac{c_L}{\beta_L} \geq B - b + \delta \nu (b - a) + \delta (1 - \nu) \left( b - V_2^L (\hat{\rho}_2^-) \right)$$

The first condition requires that $\hat{\rho}_2^- = \rho_2^*$, which uniquely determines $q^H$ as long as $\rho_1 > \tilde{\rho}_1$ defined earlier in (A.7). Then, $\phi - V_2^H (\hat{\rho}_2^-) = (1 - p_2^*)(\phi - a)$, so

$$1 - p_2^* = \frac{c_H / \beta_H - B + b}{\delta (\phi - a)}, \quad \text{(A.25)}$$

requiring that:

$$c_H / \beta_H < B - b + \delta (\phi - a) = C_H. \quad \text{(A.26)}$$

The second equilibrium condition then becomes:

$$\frac{c_L}{\beta_L} \geq B - b + \delta \nu (b - a) + (1 - \nu) \left[ c_H / \beta_H - B + b \right] \left( \frac{b - a}{\phi - a} \right) \quad \text{(A.27)}$$

In the $(c_L / \beta_L, c_H / \beta_H)$ plane, the boundary for this inequality is again the line $L_2$, with slope $1/(1 - \nu)$, that goes from the point $(B - b + \nu \delta (b - a), B - b)$ to the point $(B - b + \delta (b - a), B - b + \delta (\phi - a))$.

- For $\nu = 0$, $R_{03}$ therefore exists in regions III$^-$ and IV for $\rho_1 > \tilde{\rho}_1$. For $\nu = 1$, in which case $\tilde{\rho}_1 = \rho_2^*$, it exists in region IV only, for $\rho_1 > \rho_2^*$.  

44
7) When is $R_{12}$ an equilibrium in period 1?

Under $R_{12}$ the updating rules imply $\rho_2^+ = \rho_1$, $\rho_2^- = \text{any } \rho' \leq \rho_1$, and

$$\frac{\hat{\rho}_2^-}{1 - \hat{\rho}_2^-} = \left( \frac{\rho_1}{1 - \rho} \right) (1 - q^H),$$

$$\frac{\hat{\rho}_2^+}{1 - \hat{\rho}_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{(1 - \pi)q^H + \pi(1 - \nu)}{\pi(1 - \nu)} \right).$$

Conditions (9)–(10) now take the form:

$$\frac{c_H}{\beta_H} = B - b + \delta (V_2^H(\hat{\rho}_2^+) - V_2^H(\hat{\rho}_2^-))$$

$$\frac{c_L}{\beta_L} \leq B - b + \delta \nu (V_2^L(\rho_1) - V_2^L(\rho')) + \delta(1 - \nu) (V_2^L(\hat{\rho}_2^+) - V_2^L(\hat{\rho}_2^-))$$

The first condition requires either Case 1 or Case 2 below.

**Case 1:** $\hat{\rho}_2^- = \rho_2^*$, which then uniquely defines $q^H$ as long as $\rho_1 > \rho_2^*$. Then,

$$1 - p_2^* = \frac{c_H/\beta_H - B + b}{\delta(\phi - a)},$$

requiring

$$c_H/\beta_H < B - b + \delta (\phi - a) = C_H. \quad (A.28)$$

The second equilibrium condition then becomes:

$$\frac{c_L}{\beta_L} \leq B - b + \delta \nu (b - V_2^L(\rho')) + (1 - \nu) (c_H/\beta_H - B + b) \left( \frac{b - a}{\phi - a} \right).$$

This can be satisfied with $\rho' \leq \rho_1$ as long as

$$\frac{c_L}{\beta_L} \leq B - b + \delta \nu (b - a) + (1 - \nu) (c_H/\beta_H - B + b) \left( \frac{b - a}{\phi - a} \right). \quad (A.29)$$

In the $(c_L/\beta_L, c_H/\beta_H)$ plane, the boundary for the latter inequality is again the line $\mathcal{L}_2$, with slope $1/(1 - \nu)$, that goes from the point $(B - b + \delta \nu (b - a), B - b)$ to the point $(B - b + \delta (b - a), B - b + \delta (\phi - a))$. 

45
Case 2: \( \hat{\rho}_2^+ = \rho_2^* \), which then uniquely defines \( q^L \) as long as \( \rho_1 < \rho_1 < \rho_2^* \). Then,

\[
\frac{c_H}{\beta_H} = B - b + \delta \rho_2^* (\phi - a) \leq B - b + \delta \nu \times 0 + \delta (1 - \nu) \rho_2^* (b - a),
\]

which uniquely determines \( \rho_2^* \) as long as

\[
\frac{c_H}{\beta_H} \leq B - b + \delta (\phi - a) = C_H
\]

\[
\frac{c_L}{\beta_L} \leq B - b + \delta (1 - \nu) (c_H / \beta_H - B + b) \left( \frac{b - a}{\phi - a} \right).
\]

In the \((c_L / \beta_L, c_H / \beta_H)\) plane, the boundary for this latter inequality is the line \( L_3 \), with slope \( 1 / (1 - \nu) \), that goes from the point \((B - b, B - b)\) to the point \((B - b + \delta (1 - \nu) (b - a), B - b + \delta (\phi - a))\).

- Putting together Cases 1 and 2, we see that when \( \nu = 0 \), \( R_{12} \) exists only in region III for \( \rho > \rho_2^* \) (Case 1) as well as for \( \rho_1 < \rho_1 < \rho_2^* \) (Case 2); hence for all \( \rho_1 > \rho_2^* \). When \( \nu = 1 \) it exists in all of region III for \( \rho > \rho_2^* \) (Case 1).

8) When is \( R_{13} \) an equilibrium in period 1?

Under \( R_{13} \) the updating rules imply \( \rho_2^- = \hat{\rho}_2^- = 0 \),

\[
\frac{\rho_2^+}{1 - \rho_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1}{q^L} \right)
\]

\[
\frac{\hat{\rho}_2^+}{1 - \hat{\rho}_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1 - \pi + \pi (1 - \nu)}{\pi (1 - \nu) q^L} \right) = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1}{q^L} \right) \left( \frac{1}{1 - \pi} \right).
\]

The equilibrium conditions (9)–(10) now take the form:

\[
\frac{c_H}{\beta_H} \leq B - b + \delta \left( V^H_{2} (\hat{\rho}_2^+) - a \right)
\]

\[
\frac{c_L}{\beta_L} = B - b + \delta \nu \left( V^L_{2} (\rho_2^+) - a \right) + \delta (1 - \nu) \left( V^L_{2} (\hat{\rho}_2^+) - a \right).
\]

The second condition cannot be an equality (except with measure zero) unless either \( \hat{\rho}_2^+ \) or \( \rho_2^+ \) is equal to \( \rho_2^* \).

Case 1: \( \rho_2^+ = \rho_2^* \), which then uniquely defines \( q^L \) as long as \( \rho_1 < \rho_2^* \). Since \( \rho_2^+ > \rho_2^+ \) always,
the equilibrium conditions then become

\[
\frac{c_H}{\beta_H} \leq B - b + \delta (\phi - a) = C_H
\]

\[
\frac{c_L}{\beta_L} = B - b + \delta [\nu p_2^* + 1 - \nu] (b - a).
\]

Hence:

\[
p_2^* = \frac{c_L/\beta_L - B + b - \delta (1 - \nu)(b - a)}{\delta \nu (b - a)},
\]

requiring:

\[
B - b + \delta (1 - \nu)(b - a) < c_L/\beta_L < B - b + \delta (b - a) = C_L.
\]

**(A.32)**

Case 2: \(\hat{\rho}_2^+ = \rho_2^*,\) which then uniquely defines \(q^L,\) as long as \(\rho_1 < \rho_1^*\) defined in (A.11). Since \(\hat{\rho}_2^+ > \rho_2^*\) always, the two conditions then become:

\[
\frac{c_H}{\beta_H} \leq B - b + \delta \rho_2^* (\phi - a)
\]

\[
\frac{c_L}{\beta_L} = B - b + \delta (1 - \nu) \rho_2^* (b - a).
\]

The latter condition determines \(p_2^*\) uniquely, as long as

\[
\frac{c_L}{\beta_L} \leq B - b + \delta (1 - \nu)(b - a).
\]

Finally, the first condition requires

\[
\frac{c_H}{\beta_H} \leq B - b + \left( \frac{c_L/\beta_L - B + b}{1 - \nu} \right) \left( \frac{\phi - a}{b - a} \right)
\]

In the \((c_L/\beta_L, c_H/\beta_H)\) plane, the boundary is again the line \(L_3,\) with slope \(1/(1 - \nu),\) that goes from the point \((B - b, B - b)\) to the point \((B - b + \delta (1 - \nu) (b - a), B - b + \delta (\phi - a)).\)

- Therefore, when \(\nu = 0, R_{13}\) exists only in region III only for \(\rho < \rho_1^*\) (Case 2). When \(\nu = 1\) it exists in all of region III for \(\rho_1 < \rho_2^*\) (Case 1)

**9) When is \(R_{01}\) an equilibrium in period 1?**

Under \(R_{01}\) the updating rules imply \(\rho_2^- = 0\) and
\[
\frac{\rho_2^+}{1 - \rho_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{1}{q^L} \right) \tag{A.33}
\]
\[
\frac{\hat{\rho}_2^+}{1 - \hat{\rho}_2^+} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{(1 - \pi)q^H + \pi(1 - \nu)}{\pi(1 - \nu)q^L} \right)
\]
\[
\frac{\hat{\rho}_2^-}{1 - \hat{\rho}_2^-} = \left( \frac{\rho_1}{1 - \rho_1} \right) \left( \frac{(1 - \pi)(1 - q^H)}{1 - \pi + \pi(1 - \nu)(1 - q^L)} \right)
\]

Note that \(\hat{\rho}_2^+ > \rho_2^+ > \rho_2 > \hat{\rho}_2^-\). The equilibrium conditions are then:

\[
\frac{c_H}{\beta_H} = B - b + \delta \left[ V_2^H(\hat{\rho}_2^+) - V_2^H(\hat{\rho}_2^-) \right]
\]
\[
\frac{c_L}{\beta_L} = B - b + \delta \nu \left[ V_2^L(\rho_2^+) - a \right] + \delta(1 - \nu) \left[ V_2^L(\hat{\rho}_2^+) - V_2^L(\hat{\rho}_2^-) \right].
\]

The first condition cannot be an equality unless either \(\hat{\rho}_2^+\) or \(\hat{\rho}_2^-\) is equal to \(\rho_2^*\); in that case, the equality determines at most one suitable \(p_2^*\). The second condition cannot be an equality unless either \(\rho_2^+\) or \(\hat{\rho}_2^+\) or \(\hat{\rho}_2^-\) is equal to \(\rho_2^*\); in either case, the equality again determines at most one suitable \(p_2^*\). Except with measure zero in the \((c_L/\beta_L, c_H/\beta_H)\) space, these two values of \(p_2^*\) will be different; thus an equilibrium of this type cannot exist. Intuitively, no single mixing strategy can make both types indifferent.
References


[34] Prelec, D. “Values and Principles: Some Limitations on Traditional Economic Analysis,”


