Centre for
Economic
and Financial
Research
at
New Economic
School

October 2008

Goal Setting as a Self-Regulation Mechanism

Anton Suvorov
Jeroen van de Ven

Working Paper No 122

CEFIR / NES Working Paper series

Goal Setting as a Self-Regulation Mechanism

Anton Suvorov[*]and Jeroen van de Ven[†]

October 7, 2008

Abstract

We develop a theory of self-regulation based on goal setting for an agent with present-biased preferences. Preferences are assumed to be reference-dependent and exhibit loss aversion, as in prospect theory. The reference point is determined endogenously as an optimal self-sustaining goal. The interaction between hyperbolic discounting and loss aversion makes goals a credible and effective instrument for self-regulation. This is an entirely internal commitment device that does not rely on reputation building. We show that in some cases it is optimal to engage in indulgent behavior, and sometimes it is optimal to set seemingly dysfunctional goals. Finally, we derive a condition under which proximal (short term) goals are better than distal (long term) goals. Our results provide an implicit evolutionary rationale for the existence of loss aversion as a means of self-control.

JEL codes: D00, D80, D90.

Keywords: self-regulation, goals, time inconsistency, loss aversion, indulgence, compulsiveness, proximal and distal

[*]CEFIR and New Economic School, Moscow; asuvorov@nes.ru. Address: NES, Nakhimovsky Prospekt, 47, Suite 1721, 117418 Moscow, Russia.

[†]University of Amsterdam (UvA); j.vandeven@uva.nl. Address: University of Amsterdam, ACLE, Roetersstraat 11, 1018 WB Amsterdam, the Netherlands.

1 Introduction

It is a prevalent idea in psychology that goals helps to motivate oneself and increase persistence (Locke and Latham, 1990; Heath et al., 1999). Simplifying, the principal claim in that literature is that a person who plans to run 5 miles in the park will run 5 miles and then stop, while if that same person had set the more ambitious goal for herself to run 10 miles, she would not stop until she made it to the 10 miles mark. Such personal goals, without incentives imposed by others, have been largely neglected in the economics literature. The reason is simple: based on the rational paradigm, with its focus on optimal choices and the usual neglect of self-control and self-regulation issues, economics makes no difference between a personal "goal" and a "strategy" or "decision".

In the realm of intrapersonal conflicts, goals can have substance. We show that goals can be part of self-regulation strategies: one of the intrapersonal "selves" can set goals for other selves. Understanding the mechanisms that may induce fulfillment of these goals despite the divergence of preferences then becomes an interesting and important topic. This provides an alternative to other self-regulation mechanisms discussed in the literature, such as external commitments (e.g. using illiquid assets (Laibson (1997))), internal commitments and personal rules (Bénabou and Tirole (2004)), manipulation of self-confidence and self-esteem (Carrillo and Mariotti (2000), Bénabou and Tirole (2003)). The mechanism of goalsetting differs in that it works as an internal commitment device that does not rely on reputation building.

One objective of this paper is to investigate goal setting in a formal game-theoretic model that allows to evaluate some trade-offs that arise when one chooses more or less ambitious goals. A central element of our model is time inconsistency. Many studies suggest that people become less patient when payoffs are nearby in the future.[1] Impatience leads to temptations to spend savings, to procrastinate important but unpleasant tasks, to deviate from a diet, etc. By setting goals for the future, the person's current self can try to control the behavior of her subsequent incarnations.

A central issue is of course the credibility of such goals, which can be considered

[1] See, e.g., a survey in Frederick et al. (2002).

as a constraint imposed by rationality. In this paper we show that the presence of *loss aversion* can ensure credibility of goals and make them an effective instrument of self-regulation.[2] Having established this, we then derive some features of goals that are of central interest in the psychology literature. In particular, we show that (1) self-control can be further enhanced by (ex ante) costly self-rewards or 'indulgence'; (2) goals can be dysfunctional in some circumstances; and (3) rigid 'proximal' short-term goals can be better than more flexible 'distal' long-term goals.

The second objective of this paper is to give an implicit evolutionary argument for the possible origins of loss aversion. Indeed, loss aversion seems a dysfunctional feature of human preferences: in many situations it leads to lower materials payoffs, or, using evolutionary terminology, lower "fitness". One might then ask why people would be endowed with such preferences. A recent thriving literature investigates the origins of our preferences.[3] Such evolutionary models can generate the formation of intertemporal preferences with hyperbolic discounting: Dasgupta and Maskin (2005) and Samuelson and Swinkels (2006) show that such preferences can be evolutionarily optimal in the presence of uncertainty. In contrast, we are not aware of the literature where the formation of preferences exhibiting loss aversion is given an evolutionary argument, and our paper makes a step in this direction.[4] We show that some degree of loss aversion can be beneficial for mitigating self-control problems.

In our model, a person with present-biased preferences forms beliefs (i.e. sets a goal) at date 0 whether or not to exert effort on a task at date 1. Exerting effort implies immediate costs in terms of disutility, and yields a delayed reward at date 2. From an ex ante (date 0) perspective, the person is better off undertaking the task, but the bias towards the present implies that when the actual effort decision has to be made (date 1), the person is tempted to withdraw from the goal set and exert no effort.

Setting a high-effort-high-reward goal may not be feasible if the present bias is

[2] Relatedly, in a recent paper Kőszegi and Rabin (2008) show that loss aversion can help to make optimal consumption plans credible.

[3] See Robson (2001) and Robson (2002) for reviews.

[4] See, however, Brunnermeier (2004) for foundations of prospect theory based on bounded rationality.

strong enough. Loss aversion, however, makes it more costly for date-1 self to deviate from an ambitious goal since this would imply a loss with respect to the expected date-2 reward, and this loss "looms larger" than the gain in terms of saved date-1 cost of effort. Loss aversion is thus able to counteract the tendency to shirk arising due to impatience. On the other hand, with loss aversion there are ranges of parameters for which there are multiple equilibria, including 'bad' equilibria, in which the person sets low self-sustaining expectations. In this case, she will not exert effort even if effort is optimal from both self-0 and self-1's point of view in the absence of loss aversion. However, since we interpret the reference point as the goal deliberately set by the person, we mostly focus on optimal goals.

By showing how goals can mitigate self-control problems, our model provides an explanation for the apparent wide use of personal goals. Our results show why and when goal setting can be effective. In our framework, it is the *interaction* between loss aversion and hyperbolic discounting that creates demand for goal setting and makes it an effective strategy. Neither loss aversion nor hyperbolic discounting by themselves are sufficient to generate nontrivial impact of goals on motivation. We show that the set of parameters such that effective goal setting is possible expands with growing severeness of both loss aversion and hyperbolic discounting. In the basic model without uncertainty, goal setting allows to achieve, for a range of parameters, ex ante optimal (first best) behavior, infeasible in the absence of loss aversion.

We then extend the basic model and show that effective goal setting can be expanded further by providing rewards to oneself contingent on completing the task. We show, however, that these rewards have to be costly from an ex-ante point of view. This is reminiscent of indulgent behavior: treating oneself on a good bottle of St. Emilion wine after finishing a paper, a bottle that would otherwise be considered as too extravagant.

Then, uncertainty about disutility of effort is introduced in our model. We show that loss aversion creates overly rigid ("compulsive") goals. For some parameters, the person optimally chooses to always (never) work even though costs may be higher (lower) than benefits from every period's perspective. Kőszegi and Rabin (2006) get a similar result in a different context, and we extend it to a context with hyperbolic discounting and discuss the interplay between loss aversion and hyperbolic discounting.

Finally, we extend the model to the case where a task requires two units of effort, to be completed over the course of two periods. We analyze the trade-off between proximal and distal goals, and show that the optimal type of goals depends on the correlation of costs shocks over the two periods. In particular, proximal goals are preferred when costs are positively correlated over time, but may be worse under negative correlation.

The paper is organized as follows. The next section reviews some relevant literature. Section 3 presents the basic model and analyzes optimal goals. Section 4 extends the analysis to self rewards and indulgence. Uncertainty is introduced in section 5, and proximal and distal goals in section 6. Section 7 concludes.

2 Background

Our model rests on a few key ingredients (time inconsistency, loss aversion, reference point), and we briefly review the relevant literature in this section, and that of goal setting.

Time inconsistency The first premise is that people tend to procrastinate, or more generally that they have problems of self-control. This is a widely held belief, and an emerging literature provides abundant evidence of this. The most documented manifestation is a declining rate of time preferences: subjects at the same time prefer \$100 now to \$110 tomorrow, and \$110 in 31 days to \$100 in 30 days (see e.g. the extensive review by Frederick et al. (2002)). Among more recent contributions, Benhabib et al. (2007), for instance, reject exponential discounting in their experiments, and support the presence of a present bias. Relatedly, in experiments reported in Burger et al. (2008), participants had to complete 75 hours of studying over a 5 week period. Many participants delayed studying towards the last few weeks, although there was much heterogeneity among participants in this respect..

Most models of time inconsistent behavior predict that people could benefit from commitments, and would even be willing to pay for the opportunity to commit. Ariely and Wertenbroch (2002) find that students do better with more strict deadlines, and when given the opportunity many students self impose costly deadlines. Such a demand for commitment is also found by Bernatzi and Thaler (2004) and by Ashraf et al. (2004) for the case of savings.

Time inconsistent behavior can be interpreted in several ways, such as a conflict between consecutive selves as in models with quasi-hyperbolic discounting (e.g. Strotz (1955), Laibson (1997), O'Donoghue and Rabin (1999)) or a conflict between a long-run self and a sequence of myopic short-run selves as in "dual-self" models (e.g. Thaler and Shefrin (1981), Fudenberg and Levine (2006)). We follow the approach of quasi-hyperbolic discounting, which allows to capture time inconsistency in a concise and tractable way, even though it may not be completely accurate.[5]

Reference point One of the main elements of Kahneman and Tversky's (1979) prospect theory is the reference point. In prospect theory preferences are defined not over final outcomes, but depend on their comparison with a preset reference point. Prospect theory, on which we base our model, is incomplete in that it does not explain how a key element – the reference point – is determined. In this paper we assume as Heath et al. (1999) do, that goals set by the person herself serve as future reference points. However, Heath et al. (1999) disregard the constraints that are naturally imposed by a requirement that the goals set be credible (they also do not consider self-control problems). In contrast, we follow Kőszegi and Rabin (2006) and assume that the goal – the reference point – corresponds to rational expectations about the future behavior. Although people do systematically deviate from goals they set for themselves, we believe that goals cannot be set arbitrarily and assuming that they be realistic forecasts is a natural, although somewhat exaggerated, benchmark.

Loss aversion Loss aversion, another key component of prospect theory, refers to the tendency of people to feel losses (with respect to the reference point) stronger than same-size gains: in the words of Kahneman and Tversky (1979) 'losses loom

[5] See, however, Rubinstein (2003) for a critique of this approach. Benhabib et al. (2007) also find no evidence supporting this specific form. Gul and Pesendorfer (2001) find that many anomalies can in fact be explained in a framework where people have preferences over sets of lotteries.

larger than gains'. Loss aversion is demonstrated in many empirical studies (see Kahneman and Tversky (2000)). It is put forward as an explanation of different types of preference reversals observed in the lab (e.g. Tversky and Kahneman (1991)) and of the endowment effect observed both in the lab and in the field[6]. Loss aversion may also account for the disposition effect observed among individual investors – a tendency to sell winners too soon and to hold on to losers for too long (Odean, 1998). Such phenomena seem to lower payoffs. In a theoretical model Shalev (2002) also shows that stronger loss aversion leads to a lower share of the pie a player gets in a (à la Nash) bargaining game. Besides, loss aversion is associated with lower IQ (Frederick, 2005) and it evaporates as the experience of market interactions accumulates (List, 2003).

Goal setting There is a literature in psychology that upholds the belief that goals have an impact on motivation and thereby performance, even if they are not accompanied by extrinsic incentives such as bonuses, piece rates, or threats of sanctions. They argue that by merely changing the goal, motivation can be improved. Locke and Latham (1991) survey over 200 studies in their authoritative and voluminous book. Heath et al. (1999, p. 81) summarize this literature as follows: "performance increases have been documented using tasks ranging from cognitive, such as solving anagrams or thinking of creative uses for a common household object, to physical, such as cutting logs and pedaling a bicycle".

In the typical experiment, some individuals are assigned an easy task and others a more challenging task. For instance, Matsui et al. (1981) asked subjects to detect discrepancies between two lists of three digit numbers. Those assigned to an easy goal perform on average significantly worse than those assigned the more challenging goal. Hence, the mere suggestion of a more challenging goal improved performance.

While the volume of the empirical literature on goal setting is impressive, understanding of mechanisms through which goals affect behavior seems to be lagging behind. More knowledge is needed on when and why personal goals are effective. A

[6] The endowment effect (Thaler (1980)) is a general tendency of people to value goods they have more than similar goods they do not have. It has been demonstrated in many lab and field experiments (e.g. Kahneman, Knetsch and Thaler (1990), List (2003, 2004)), but its generality and robustnest are not indisputable (e.g. List (2003, 2004), Plott and Zeiler (2005)).

difficulty in interpreting the experimental results also stems from the fact that personal goals are hard to observe, and in many experiments the manipulation of personal goals may have been confounded with changes in other factors. For instance, subjects may simply feel they should adhere to the authority of the experimenter. Given this lack of structured approach, we believe that our paper contributes to understanding of the mechanisms which make personal goals credible and effective in changing behavior, and suggests a framework to evaluate different types of goals from both a positive and a normative perspective.

3 The basic model

Consider an agent that is facing a task. Completing the task requires effort, e, and disutility of effort is given by $c(e)$. A successfully completed task yields a payoff $v(e)$. For simplicity, we assume that effort is a binary decision: $e \in \{0, 1\}$. We normalize $c(0) = 0$ and $c(1) = c$, and similarly $v(0) = 0$ and $v(1) = v$. We also assume that $c < v$.

There are three periods, $t = 0, 1, 2$. At $t = 1$ the agent chooses effort and incurs the corresponding cost. The reward for the task is received at $t = 2$. Preceding these two periods, there is a goal-setting stage at $t = 0$ in which the agent forms expectations about effort and the corresponding reward. The goal set by the agent reflects the belief that the task will be undertaken. We allow for stochastic goals: the agent may plan to play a mixed strategy at $t = 1$. We denote the goal set at $t = 0$ as $\tilde{\sigma} = (\tilde{\sigma}_0, \tilde{\sigma}_1)$ with $\tilde{\sigma}_0 + \tilde{\sigma}_1 = 1$; $\tilde{\sigma}_1$ is the belief that effort will be exerted at $t = 1$ and $\tilde{\sigma}_0$ is the belief that no effort will be exerted at $t = 1$. We assume that the set goal cannot be adjusted at later dates.

Let u_t be the agent's instantaneous utility at time t. Instantaneous utility has two components. The first component, $m_t(e)$, is standard utility, consisting of direct payoffs received at time t: the disutility of effort $m_1(1) = -c$ and $m_1(0) = 0$, and reward for the task $m_2(1) = v$ and $m_2(0) = 0$. The second component is reference dependent. It consists of the received payoffs as compared to the realizations of the reference point, where the reference point is determined by the goal set: for a deterministic goal $\tilde{e}$ the reference point is $(m_1(\tilde{e}), m_2(\tilde{e}))$; for a stochastic goal the reference point is a respective lottery. We denote by $\sigma = (\sigma_0, \sigma_1)$ the mixed strategy in period 1 that assigns probability σ_i to pure strategy $e = i$. We assume that u_t takes the following expected utility form, adopting the framework by Kőszegi and Rabin (2006)[7]:

$$u_t(\sigma|\tilde{\sigma}) = \sum_{e=0,1} \sigma_e \sum_{\tilde{e}=0,1} \tilde{\sigma}_{\tilde{e}} \left[m_t(e) + \mu(m_t(e) - m_t(\tilde{e})) \right],$$

where

$$\mu(x) = \begin{cases} \eta x & \text{if } x > 0, \\ \eta \lambda x & \text{if } x \leq 0 \end{cases}$$

is the component of the utility function reflecting its reference-dependent nature. The parameter $\eta \geq 0$ reflects the weight attached to the reference point component, and $\lambda \geq 1$ reflects the degree of loss aversion. Note, in particular, that payoffs are evaluated against each possible realization of the stochastic reference point.

Let U_t the agent's intertemporal (expected) utility function from the perspective of time t. We consider a quasi-hyperbolic intertemporal utility function of the $(\beta, \delta)-$ form[8]:

$$U_t = \delta^t u_t + \beta \sum_{\tau=t+1}^{2} \delta^\tau u_\tau$$

where $0 < \beta \leq 1$ and, without loss, we set $\delta = 1$ for simplicity. Hence, all future payoffs are discounted at rate β.

To close the model we need to make some additional assumptions on how goals are set. First of all we assume that goals are self-sustaining: the person has rational expectations about her actual behavior. Second, we assume that the goal set at $t = 0$ maximizes intertemporal utility U_o – the person chooses the optimal self-sustaining goal.[9]

[7] This specification with a piecewise-linear reference-dependent term is a special case of Kőszegi and Rabin (2006). They also consider this special case in some applications.

[8] See Phelps and Pollak (1968), Laibson (1997), Rabin and O'Donoghue(1999).

[9] Self-sustaining goals correspond to Personal Equilibria (PE) of Kőszegi and Rabin (2006), while optimal self-sustaining goals correspond to their Preferred Personal Equilibria.

Definition 1 *(i) Goal $\tilde{\sigma}$ is* self-sustaining *if $U_1(\tilde{\sigma}|\tilde{\sigma}) \geq U_1(\sigma'|\tilde{\sigma})$ for any σ'.*
(ii) Self-sustaining goal $\tilde{\sigma}$ is optimal, *if $U_0(\tilde{\sigma}|\tilde{\sigma}) \geq U_0(\sigma'|\sigma')$ for any other self-sustaining goal σ'.*

Our main interest is in the cases where the goal set at $t = 0$ has a nontrivial impact on the behavior at $t = 1$. We say that goal setting is *effective* if the optimal self-sustaining goal from $t = 0$ perspective is different from the optimal goal from $t = 1$ perspective. If, on the other hand, the agent's optimum goal is the same from the perspective of $t = 1$ and $t = 0$, or if they are different but $t = 1$-optimal goal is not self-sustaining, the goal setting does not have impact on behavior.

3.1 The choice of effort with exogenous goals

We first analyze the agent's behavior at date 1 given any goal. Suppose the agent's goal is to exert effort with probability q, i.e. $\tilde{\sigma} = (1 - q, q)$. This means that with probability q the agent expects a reference point $(-c(1), v(1)) = (-c, v)$, and with probability $1 - q$ the agent expects a reference point $(-c(0), v(0)) = (0, 0)$. If the agents exerts effort, the utility evaluated at $t = 1$ is then given by:

$$-c - (1 - q)\eta\lambda c + \beta v + (1 - q)\beta\eta v. \tag{1}$$

Effort yields a direct payoff equal to $-c + \beta v$. With probability q effort is expected, and exerting effort therefore coincides with the reference point. With probability $1 - q$ no effort is expected. In that case, the disutility of efforts is a loss relative to the reference point, and the reward is a gain compared to the reference point. Similarly, the utility from not exerting effort is given by:

$$q\eta c - q\beta\eta\lambda v. \tag{2}$$

There is no direct payoff in this case. With probability $1 - q$ no effort is expected and there is no gain or loss relative to the reference point. With probability q, effort is expected and saving disutility of effort is a gain, and no reward is a loss compared to the reference point.

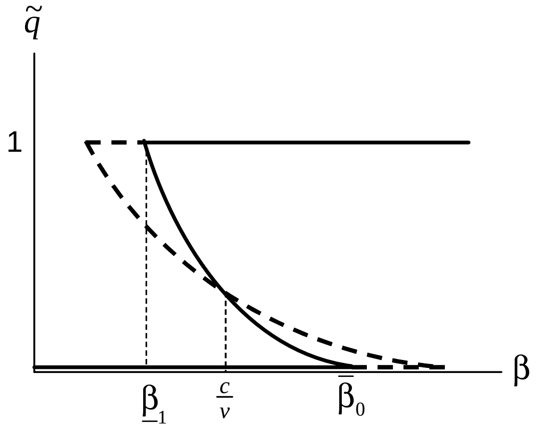

Figure 1: Self-sustaining goals and present-bias β. The dashed lines show the effect of an increase in loss aversion λ.

For any $q \in (0,1)$, for the goal to be self-sustaining, the agent must be willing to randomize between effort and no effort. This requires the agent to be indifferent. Combining (1) and (2), the agent is indifferent for $q = \tilde{q}$ where:

$$\tilde{q} \equiv \frac{c(1+\eta\lambda) - \beta v(1+\eta)}{\eta(\lambda-1)(c+\beta v)}. \tag{3}$$

Note that $\tilde{q} = 1$ for $\beta = \frac{c}{v}\frac{(1+\eta)}{(1+\eta\lambda)} \equiv \underline{\beta}_1$, and $\tilde{q} = 0$ for $\beta = \frac{c}{v}\frac{(1+\eta\lambda)}{(1+\eta)} \equiv \overline{\beta}_0$. Thus, $\overline{\beta}_0$ provides an upper bound for which $e = 0$ is sustainable, and $\underline{\beta}_1$ provides a lower bound for which $e = 1$ is sustainable. It is easy to see that $\overline{\beta}_0 > \underline{\beta}_1$ for $\gamma > 1$.

Proposition 1 *For any given goal, if $\beta < \underline{\beta}_1$ the unique self-sustaining goal is not to exert effort, $\tilde{\sigma} = (1,0)$; if $\beta > \overline{\beta}_0$, the unique self-sustaining goal is to exert effort, $\tilde{\sigma} = (0,1)$; and if $\beta \in [\underline{\beta}_1, \overline{\beta}_0]$, three self-sustaining goals exist: both pure strategy $\tilde{\sigma} = (0,1)$ and $\tilde{\sigma} = (1,0)$ and a mixed strategy one $\tilde{\sigma} = (1-\tilde{q}, \tilde{q})$, where $\tilde{q}$ is given by (3).*

The solid lines in Figure 1 illustrate the result. There is a simple intuition for why there is a range of multiple equilibria. A loss averse person weighs the losses more heavily than gains. If the goal is to exert effort, the reference point is to incur the cost of effort c and then get the reward v. The loss from not exerting effort would be in not getting the reward, and this loss, if unexpected, would outweigh the unexpected gain from lower disutility of effort. On the other hand, if, for the same parameters, the goal is not to exert effort, the reference point is to incur no cost and get no reward. The loss from exerting unexpected effort would outweigh the unexpected gain. Easy comparative statics shows that the more loss averse the agent is, the larger is the interval for which there are multiple equilibria, as illustrated by the dashed lines in figure 1 which correspond to a higher value of λ.

The foregoing argument also explains why the $\tilde{q}-$curve is downward sloping. A higher $\tilde{q}$ or β both make effort more attractive. To keep the agent indifferent between effort and no effort, $\tilde{q}$ must be lower for higher β.

3.2 Choice of effort with endogenous goal setting

Now suppose the agent can set his own goal. We focus on pure strategies, as it is easy to show that mixed strategies are never better[10]. Let e_t^* denote the optimal (not necessarily self-sustaining) pure-strategy goal from date t perspective. The reference point part drops out with pure strategies and self-sustaining goals, and one immediately gets the following lemma:

Lemma 1 *(optimal goals) $e_0^* = 1$ if and only if $v \geq c$, and $e_1^* = 1$ if and only if $\beta v \geq c$.*

Since we assumed that $c < v$, date-0 self would like to set $e_0^* = 1$. Note that $\underline{\beta}_1 < \frac{c}{v} < \overline{\beta}_0$ for $\lambda > 1$. If $\beta \in [0, \underline{\beta}_1)$, date-0 self's preferred goal $e_0^* = 1$ is not self-sustaining. If $\beta \in [\underline{\beta}_1, c/v)$, despite the conflict of interest between date-0 and date-1 selves ($e_0^* = 1$ and $e_1^* = 0$), by setting goal $\tilde{e} = 1$ the agent can induce herself to exert effort at date 1 (in the equilibrium which is optimal from date-0 perspective): in our terminology there is effective goal setting in this range of parameters. If $\beta \geq c/v$ there is no conflict of interest at different periods: $e_0^* = 1$ and $e_1^* = 1$.

Proposition 2 *(i) Effective goal-setting in which $\tilde{e} = 1$ occurs for $\beta \in [\underline{\beta}_1, c/v)$, and (ii) $\underline{\beta}_1$ is decreasing in λ and η with $\underline{\beta}_1 = \frac{c}{v} < 1$ if $\lambda = 1$ and $\underline{\beta}_1 \to 0$ as $\lambda \to \infty$.*

[10] To prove this, note that for any self-sustaining goal $\tilde{q} \in (0,1)$, $U_0((1-\tilde{q},\tilde{q})|(1-\tilde{q},\tilde{q})) = \tilde{q}(v-c) - \tilde{q}(1-\tilde{q})(\lambda-1)\eta(v+c)$. Clearly, since $v \geq c$, $U_0((1-\tilde{q},\tilde{q})|(1-\tilde{q},\tilde{q})) < v - c = U_0((0,1)|(0,1))$.

The shaded area (I) on Figure 2 illustrates effective goal setting for different combinations of β and λ. Two aspects are in particular noteworthy. First, it is the *interaction* of loss aversion and present-biased discounting that drives the result. For if *either* $\beta = 1$ or $\lambda = 1$ (or $\eta = 0$), generically there is no effective goal setting. When a person is not loss averse he or she evaluates the effort decision the same way independent of the goal set. Hence, there is no range with multiple equilibria, $\underline{\beta}_1 = \overline{\beta}_0$, and goal setting cannot be effective. On the other hand, as the degree of loss aversion λ (or the weight of the reference-dependent component of the utility function η) becomes arbitrarily large, effective goal-setting becomes possible for arbitrarily small discount factors β.

Furthermore, for the range of effective goalsetting, the agent achieves first best behavior from the ex ante point of view: the agent is induced to choose $e = 1$ at date 1 and the reference-dependent component of the utility function is 0. Outside the range of effective goalsetting, the model predicts the same behavior as if there were no reference point ($\eta = 0$): $e = 1$ if $\beta v > c$. Note that in this simple deterministic setting a higher degree of loss aversion (or a larger weight of the reference-dependent component) allows to overcome self-control problems with stronger present bias in intertemporal preferences (since it increases off-equilibrium losses from deviation) and, importantly, implies no utility loss on the equilibrium path. Of course, as our analysis in Section 5 shows, in a more realistic stochastic setting this is no longer true: stronger loss aversion expands the range of implementable goals, but also increases losses in reference-dependent part of utility on the equilibrium path.

Discussion The main conclusion we would like to draw so far is that loss aversion can help to overcome self-control problems associated with hyperbolic discounting. Despite its apparent inefficiency, loss aversion may thus have some evolutionary value.

Kőszegi and Rabin (2008) derive a very similar result in the context of consumption plans. Their starting point is different: they assume that people have preferences over changes in *beliefs*, and that these preferences exhibit loss aversion with respect to good and bad news – bad news is more unpleasant than good news is pleasant.

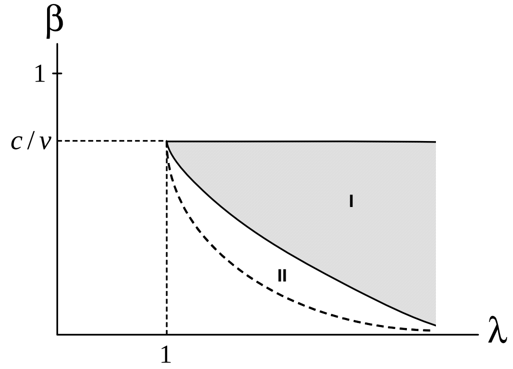

Figure 2: Effective goalsetting. Area I shows the range of effective goalsetting for different values of loss aversion λ and present bias β. Area II shows how the range of effective goalsetting expands with self-rewards.

Interestingly, that approach turns out to be analytically equivalent to the one in this paper. They show that in a simple consumption model, the optimal consumption path with the same consumption level in every period is not attainable, unless loss aversion is strong enough – in which case the optimal consumption plan is time consistent.

The self control mechanism in our model is entirely internal. By contrast, most of the other literature has focused on external commitment devices. Strategies might include commitment to contracts, making intentions public, or delegating authority to others. However, as Schelling (1984, p. 2) notes, "most of the tactics used to command one's own future performance probably do not depend on someone else's participation." An exception is the insightful paper by Bénabou and Tirole (2004) who also focus on internal commitments. In their analysis reputation building plays a crucial role, in combination with imperfect self-knowledge. In our view, our models are complementary.

Besides that, we also believe that we add an insight to the psychology literature. There is a vast literature in psychology on goal setting, and its relation to motivation and self-control. It is postulated that goals are an important regulator of effort. Baumeister et al. (1994) underline the importance of goals as being a prerequisite in effective self-regulation. Locke and Latham (1990) summarize numerous experiments that study the impact of goals on motivation, and find important effects. Our model provides one possible explanation of the causal effect of goals on motivation in terms of multiple equilibria: choosing the right goal can bring a person to an efficient intrapersonal equilibrium.

4 Self Rewards and Indulgence

Casual observation suggests many people engage in one or more forms of self-control. Paradoxically, it seems that the very same individuals trying to exercise self-control at the same time deliberately also engage in self-indulgent behavior. Those on a diet regularly treat themselves on a chocolate bar after depriving themselves, others go shopping after homework, get a pampering massage, or go out for a sumptuous meal after finishing a tiresome project at the office. We conjecture most readers are familiar with the idea of buying something you wouldn't "normally" buy for yourself. In this section we show that indulging can be an integral part of exercising self-control.

We now subdivide period 1 into two subperiods. In the first subperiod the agent chooses the effort level. In the second subperiod the agent decides to take a bonus or not. We assume that there is no discounting between these two subperiods (this assumption simplifies formulae without loss of generality). In period 0, the agent sets a goal in terms of effort and forms expectations about getting the bonus. We assume that expectations about getting the bonus can be made contingent on the choice of effort. In particular, the agent can set as goal to exert effort, and form expectations to take the bonus after effort only. If at date 1 the agent decides not to exert effort, then he updates his beliefs about getting the bonus in the second subperiod, and will not expect a bonus. This effectively ties the bonus to effort.

Our focus will be on the case where goalsetting by itself is not sufficient to eliminate the self-control problem, that is $\beta < \underline{\beta}_1$. Otherwise the bonus will not be essential for convincing self-1 to exert effort: self-0 may still use a bonus that is tied to effort, but efforts can be sustained even without the bonus so that the bonus does not effectively solve the self-control problem. In fact, in that case the bonus would only be used (in the equilibrium optimal for self-0) if self-0 liked to get the bonus anyway independent of effort.

We consider a bonus that yields a direct payoff b to the agent at $t = 1$, at cost γb incurred at $t = 2$. Indeed, if a person is not facing liquidity problems, then by buying a luxury good she deprives herself of some resources that would be available for *future* consumption. We also assume that self-0 has some control over the size of the bonus by choosing the level of γ; the cost-benefit ratio. We can think of different values of γ as characterizing different kinds of gifts – more or less extravagant. We impose a natural constraint $\gamma \geq 1$ that captures the idea that the bonus is costly; the following proposition shows, however, that this constraint is not binding. Self-0 can select the appropriate type of gift. The fact that the agent should take the bonus only if she has planned to do it (i.e. she has exerted effort) imposes some constraints on the cost-benefit ratio γ: not any bonus will do. The following result shows that the optimal bonus is one that self-0 would prefer not to receive if it were not required for creating incentives to exert effort.

Proposition 3 *(i) The optimal cost-benefit ratio of the bonus is given by* $\gamma^* = \frac{1+\eta}{\beta(1+\eta\lambda)}$*, and (ii) The expected net payoff from getting the optimal bonus is negative from date-0 perspective and positive from date-1 perspective:* $1 < \gamma^* < \frac{1}{\beta}$.

Proof. The beliefs set by self-0 about getting the bonus, must have some impact on the behavior of self-1. If self-1 would either always or never take the bonus no matter what were expected, then the bonus could not help to overcome the self-control problem. As with exerting effort, there is a range of parameters for which there are multiple self-sustaining equilibria. This is the case for $\gamma \in \left[\frac{1+\eta}{\beta(1+\eta\lambda)}, \frac{1+\eta\lambda}{\beta(1+\eta)}\right]$. Since self-0 wants to minimize the costs of the bonus, part 1 follows: the optimal γ^* is the lower bound of the interval of feasible values. Since we focus on $\beta < \underline{\beta}_1 = \frac{c}{v}\frac{1+\eta}{1+\eta\lambda} < \frac{1+\eta}{1+\eta\lambda}$, it follows that $\gamma^* = \frac{1+\eta}{\beta(1+\eta\lambda)} > 1$. Finally, $\beta\gamma^* = \frac{1+\eta}{(1+\eta\lambda)} < 1$. ■

Part 1 shows that the optimal cost-benefit ratio of the bonus is decreasing in β, η, and λ. A higher value of β reduces the self-control problem, and can reduce the costs of the bonus. Higher values of λ and η make the person more loss averse, increasing the scope of multiple equilibria, so that the bonus can be effective at lower values of γ.

The second part shows that the bonus is costly: self-0 would prefer not to take the bonus if not for creating incentives for self-1. So far we have shown that an effective bonus must be costly. We now show that effective bonuses indeed do exist and self-0 can use them to overcome the self-control problem. By conditioning the bonus on exerting effort, self-0 can convince self-1 to exert effort. The strategy of self-0 will therefore be to set $\tilde{e} = 1$ as the goal, and to expect the bonus if $e = 1$ and not otherwise. For this to be an optimal and credible strategy, some conditions have to be fulfilled.

With the optimal bonus, self-1 will take the bonus only if expected, hence only after choosing $e = 1$. Exerting effort becomes more attractive than without the bonus: it can be sustained if $\beta\frac{(1+\eta\lambda)}{(1+\eta)} + \frac{b}{v}(1-\beta\gamma)\frac{1}{1+\eta} \geq \frac{c}{v}$. Substituting for γ^*, we find that $e = 1$ is sustainable for:

$$\beta \geq \frac{c}{v}\frac{1+\eta}{1+\eta\lambda} - \frac{b}{v}\frac{\eta(\lambda-1)}{(1+\eta\lambda)^2} \equiv \underline{\beta}_1^b. \tag{4}$$

Note in particular that$\underline{\beta}_1^b < \underline{\beta}_1$, so that $e = 1$ is sustainable for a larger set of parameters than without a bonus. Even if this condition is satisfied, however, self-0 may be better off by setting $\tilde{e} = 0$ and never expecting a bonus since $\gamma^* > 1$. Self-0 is better off with the target $\tilde{e} = 1$ and expecting a bonus if $e = 1$ if

$$\beta \geq \left(\frac{1+\eta}{1+\eta\lambda}\right)\left(\frac{b}{v-c+b}\right). \tag{5}$$

Whether (5) or (4) is binding depends on the specific parameters. At $\lambda = 1$, (4) is the relevant lower bound provided $b < c$, but for larger values of λ, (5) may become binding.[11]

Proposition 4 *(i) If $\lambda > 1$, an ex ante costly bonus b can help to overcome self-control problems for some values of β if and only if $b < c$, and (ii) the lower bound of values of β is decreasing in λ.*

[11] To be precise, if $\frac{(v-c)(c-b)}{b(V-c+b)} < \frac{1}{1+n}$, then (5) becomes binding for a sufficiently large value of λ.

Proof. Part (i). If $b \geq c$, then $\frac{b}{V-c+b} \geq \frac{c}{v}$ and thus there can be no value of β satisfying (5) and $\beta < \underline{\beta}_1$. On the other hand, if $b < c$, then $\frac{c}{v} > \frac{b}{V-c+b}$ so there always exist values of β satisfying (5) and $\beta < \underline{\beta}_1$. Also, the right-hand side of (4) is strictly lower than $\underline{\beta}_1$, so there always exist values of β satisfying (4) and $\beta < \underline{\beta}_1$. Part (ii). Both (4) and (5) are decreasing in λ, hence so is the maximum of the two. ■

This result is illustrated in Figure 2. Area (II) between the dashed line and the shaded area of effective goalsetting is where bonuses are effective and desired. Since a bonus is costly, a larger bonus can make a person only better off for higher values of β, as the optimal cost-benefit ratio of the bonus decreases in β. When $b \geq c$ the minimum required β exceeds $\underline{\beta}_1$.

Discussion Based on survey evidence, Mick and Demoss (1990) conclude that personal accomplishments and achieved goals are important reasons to reward oneself. According to them, "self gifts can act as self contracts in which the reciprocity for the gift is effort and achievement" (Mick and Demoss 1990, p. 326). These gifts often have an indulgence character. A saying illustration of indulgence is a respondent who reported to purchase "things I would not normally buy myself" after quitting smoking (Mick and Demoss, 1990 p. 327). Kivetz and Simonson (2002) provide some further evidence, and conclude that consumers are more likely to prefer luxury goods after completing an effortful task. They argue that pleasurable consumption arises from "earning the right to indulge through hard work". While this evidence is suggestive, it raises the question how indulgence can be made contingent on effort, instead of always buying the good. Why is it that under normal circumstances one would not by the good? Our model provides a reason why such a contingent strategy is enforceable. It also gives an exact content to the idea of buying goods a person would not normally buy him/herself.

5 Uncertainty

So far we abstracted away from the uncertainty. More realistically, goals are set when there is still some uncertainty about the nature of the task, such as how much effort is needed for completion, or what are the prospects of success. When news arrives, a wrongly set goal may magnify the problem. Extra losses may result from either conforming to the unrealistic goal or from deviating from the goal.

In this section we introduce some uncertainty. Uncertainty is only relevant for goalsetting if the effort decision is possibly sensitive to any new information. Thus, while there might be uncertainty about the rewards, this is of less interest if the information is received after the effort decision. Instead, we focus on uncertainty about the cost of effort, which may reflect the difficulty of the task. Suppose disutility can be either c_l or $c_h > c_{l.}$ The probability that disutility is low is ρ. The costs are unknown at the goal-setting stage, but are revealed prior to the effort decision.

Denote by (q_l, q_h) the strategy profile of the agent, where q_l denotes the probability of working after c_l, and q_h denotes the probability of working after c_h. Let $\hat{q}_i$ denote the probability of working such that an agent is indifferent between working and not working after observing c_i. Clearly it cannot be the case that both $\hat{q}_l \in (0,1)$ and $\hat{q}_h \in (0,1)$, and neither that $\hat{q}_h > \hat{q}_l$. Hence, in addition to $(q_l, q_h) = (0,0)$ and $(q_l, q_h) = (1,1)$, we can have $(q_l, q_h) = (\hat{q}_l, 0)$ with (see the proof to proposition 5 in the appendix for derivations):

$$\hat{q}_l = \frac{c_l(1+\eta\lambda) - \beta v(1+\eta)}{\eta(\lambda-1)(c_l+\beta v)\rho}, \tag{6}$$

or $(q_l, q_h) = (1, \hat{q}_h)$ with:

$$\hat{q}_h = \frac{c_H(1+\eta\lambda) - \beta v(1+\eta) - \eta(\lambda-1)\rho(c_L+\beta v)}{\eta(\lambda-1)(c_H+\beta v)(1-\rho)}. \tag{7}$$

Again we restrict analysis to pure strategies, as mixed strategies never do better. The above equations determine four relevant threshold levels, $\overline{\beta}_{00,}$ $\underline{\beta}_{10}$, $\overline{\beta}_{10}$, $\underline{\beta}_{11}$, where β_{q_i,q_j} denotes the threshold value such that strategy (q_i, q_j) is *sustainable*, and the bar indicates whether it is an upper bound or a lower bound. Likewise, one can derive threshold levels for when strategy (q_i, q_j) is *preferred.* Let v_1 denote indifference between $(1,0)$ and $(0,0), v_2$ between $(1,1)$ and $(0,0)$, and v_3 between $(1,1)$ and $(1,0)$ (these thresholds, independent of β, are derived in the Appendix).

We are particularly interested in how uncertainty can lead to a situation in which optimal goals are rigid, seemingly dysfunctional. Specifically, we show that there are cases where the agents never works even if $v \geq \beta v > c_l$. In that case, the agent never works although effort after c_l is desirable in terms of direct payoffs from the perspective of *every* date. Similarly, we show that there are cases where the agent always works even though $c_h > v \geq \beta v$. With some abuse of terminology we refer to these two cases as *underdiligence* and *overdiligence*. Rather than characterizing all possible equilibria for different parameter configurations, we focus on those two interesting cases.

Figure 3 below illustrates a situation with both types of dysfunctional goals. Equilibrium strategies are indicated in some areas, with an asterisk in case it is a preferred equilibrium. In area I there is underdiligence. $v \leq v_1$ so the agent prefers never to work. This equilibrium is sustainable as $\beta \leq \overline{\beta}_{00}$. And since $\beta v \geq c_l$, it satisfies the conditions for an overly rigid goal. Very similar remarks apply for area II, in which there is overdiligence. Kőszegi and Rabin (2006) derive a similar result in a context without hyperbolic discounting, and indeed overly rigid goals exist for $\beta = 1$. Figure 3 shows how this result interacts with lower values of β. Note in particular that there are cases for which no overly rigid goals exist for $\beta = 1$, but do exist for lower values of β as in area III.

The reason behind the existence of overly rigid goals can be understood as follows. Consider area I, the situation in which there is underdiligence. When costs are low, there is a merit of exerting effort in terms of direct payoffs. Compared to never exerting effort, this shifts the reference points towards expecting efforts with a positive probability (if cost realizations happen to be low). This creates stochastic goals, implying deviations from the goal on the equilibrium path, hence – associated utility losses. To see this more clearly, considered the expected utility from exerting effort

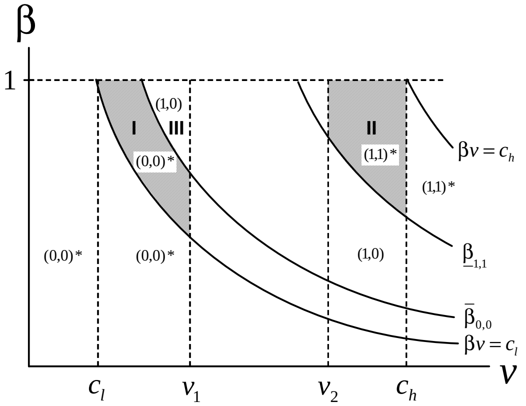

Figure 3: Under and overdiligence. Underdiligence arises in area I, overdiligence in area II.

after low costs (when the goal is (1,0)). It is given by:

$$\rho[\underbrace{-c_l + v}_{(1)} + \underbrace{(1-\rho)(-\eta\lambda c_l + \eta v)]}_{(2)} + (1-\rho)\underbrace{\rho(\eta c_l - \eta\lambda v)}_{(3)}, \tag{8}$$

The first effect, occurring with probability ρ, is the direct payoff from exerting effort and is positive by construction. If costs are low, that might be a deviation from the reference point. With probability $1-\rho$, no effort was expected so exerting effort brings higher disutility costs and rewards than expected. This is the second term, which might be positive or negative depending on λ. If costs are high, no effort is exerted but some effort was expected with probability ρ, in which case disutility costs and rewards are lower than expected. This is the third term, and this is negative since in area I $v > c_l$. By contrast, if the agent never exerts effort there is never a deviation from the reference point.

Seemingly dysfunctional goal setting only happens for $\lambda > 1$. For $\lambda = 1$, the upper bound for which never to exert effort is sustainable ($\overline{\beta}_{00}$) coincides with $\beta v = c_l$, and the lower bound for which always to exert effort is sustainable ($\underline{\beta}_{11}$) coincides with $\beta v = c_h$. As λ increases, there is more room for dysfunctional goals. Both the $\overline{\beta}_{00}-$curve and v_1 shift to the right, as is easy to show. Likewise, $\underline{\beta}_{11}$ and v_2 shift leftward. It will be clear that for some λ, $v_1 = v_2 = v_3$ and the agent is indifferent between all three strategies $(0,0)$, $(1,0)$, and $(1,1)$. For any higher value of λ, the strategy $(1,0)$ is always dominated, so the relevant question is whether or not $(1,1)$ is preferred to $(0,0)$. In the appendix we prove that if λ it is sufficiently high, then never exerting effort is always sustainable and preferred for any $v \in [c_l, c_h]$. The next proposition summarizes these results. Define $\lambda^* \equiv \max\left\{1 + \frac{1}{\eta(1-\rho)}, \frac{c_h}{c_l}\frac{1+\eta}{\eta} - \frac{1}{\eta}\right\}$, then:

Proposition 5 *Underdiligence exists for some $v \in [c_l, c_h]$ if and only if $\lambda > 1$, with $\beta \leq \max\{\overline{\beta}_{00}, 1\}$, where $\overline{\beta}_{00}$ decreases in v; overdiligence exists for some $v \in [c_l, c_h]$ if and only if $\lambda^* > \lambda > 1$, with $\beta \geq \underline{\beta}_{11}$, where $\underline{\beta}_{11}$ decreases in v.*

Proof. See Appendix. ■

6 Proximal and Distal Goals

By focusing on the simple decision to exert effort or not in a single period, we bypassed an important theme in the psychology literature. This concerns the question wether goals should be proximal or distal (Locke and Latham, 1990; Baumeister et al., 1994). That is, is it better to split the task up in small portions, such as weekly or monthly targets, or to set one single goal at the end of the project? This is essentially a question about flexibility. Proximal goals, if effective, induce greater commitments to take many small steps, eventually resulting in reaching overall success. But distal goals allow greater flexibility, and hence more responsiveness to unforeseen or uncertain shocks. Under distal goals, any absence due to illness or the interference of another more pressing task can be absorbed by working harder on other days.

Overall, the evidence whether proximal or distal goals are better is mixed (Baumeister et al., 1994). It is an open question how these data may be organized to account for the conflicting results. We believe that one natural way to pursue this question, is to relate the choice of proximal or distal goals to the distribution of shocks over time. Intuitively, distal goals allow for the absorption of an illness only if the illness does not persist over the complete time horizon.

We address the trade-off between proximal and distal goals by considering a simple extension to the main model with uncertainty. We subdivide period 1 in two distinct periods, 1a and 1b. The task is such that two units of effort are needed to complete it. The agent can either exert 1 unit of effort in every period, or both in one subperiod. If the agents exerts two units of effort in one period, total disutility is $2\psi c$, with $\psi \geq 1$. One unit of effort in a subperiod gives costs c. Thus, at each date costs of effort is a convex function, so spreading effort is less costly. As in the previous section, costs are uncertain and can be either low or high each period. We assume this time that the task has to be completed (not competing implies excessively high costs upon the agent). This will greatly reduce the number of possible cases.

The following definition conceptualizes proximal and distal goals in a precise manner.

Definition 2 *A proximal goal is such that* $e_{1a} = e_{1b} = 1$ *for any cost realization* c_{1a}*. A distal goal is such that (i)* $e_{1a} = e_{1b} = 1$ *if* $c_{1_a} = c_L$ *and (ii)* $e_{1a} = 0$ *and* $e_{1b} = 2$ *if* $c_{1a} = c_{H.}$

By this definition, distal goals are more flexible and responsive to costs than the rigid proximal goals. They do not dictate to exert effort in period $1b$ under all circumstances. This flexibility has its obvious benefits. The downside is that concentrating all efforts in one single period is assumed to be more costly than spreading efforts at any given costs as $\psi \geq 1$.

Of course, the first question is under which conditions any of the two strategies is feasible as an equilibrium strategy. But the expressions are tedious and are not our primary interest here. Rather, we would like to know which strategy the agent would choose in the event that both strategies are available to her. When only one or none type of strategies are optimal the choice is trivial.

What type of goal is better depends crucially on the correlation of shocks over time. It is insightful to consider the two extreme opposites of perfect positively correlated shocks and perfect negatively correlated shocks. Both these extremes have a useful interpretation. Perfect positively correlated shocks can for instance reflect ability of the agents for the task, which can either be high or low but is presumably relatively constant over the course of completion. On the other hand, when other obligations get in between, this will raise the costs of doing the task. Perhaps the agent knows that another task has to be completed during the same time interval as the current task, but is unsure about whether this has to be done in the first or second subperiod. This yields a perfect negative correlation of shocks over time.

Proposition 6 *Let proximal and distal goal be equilibrium strategies. Then, (i) if costs are perfect positively correlated over periods, proximal goals are preferred for any* $\psi \geq 1$*; (ii) if costs are perfect negatively correlated over periods proximal goals are preferred for* $\psi \geq \bar{\psi}(c_H, c_L, \lambda)$*, and distal goals otherwise, with* $\bar{\psi}(\cdot)$ *increasing in* c_H *and decreasing in* c_L *and* λ*.*

When costs are perfect negatively correlated, distal goals sometimes do better than proximal goals under our assumptions. Distal goals are more flexible, and prevent incurring high costs in the first period. The downside are the higher costs of concentrating all effort in one period, and this is worse for higher values of ψ. The critical value $\bar{\psi}$ beneath which distal goals are preferred, increases in c_H and decreases in $c_{L.}$ This is intuitive, since if the difference in costs is large, postponing after high costs and exert effort when costs are low becomes more beneficial. If $\Delta \equiv c_H - c_L \to 0$, a rough measure of the uncertainty level, proximal goals are preferred for any $\psi \geq 1$. A higher degree of loss aversion λ makes the deviations from the reference point more costly, which makes proximal goals more attractive for lower values of ψ since proximal goals have less extreme deviations from the reference point.

When costs are positively correlated, the flexibility of distal goals loses its bite. The agent postpones exerting effort after high costs, but when shocks are positively correlated, costs will also be high in the next period. Thus, postponing only results in even higher costs. In this case, proximal goals are always better for any $\psi \geq 1$.

7 Discussion and conclusions

Building on an extensive literature in psychology, we proposed a formal model that explains why goalsetting can be a credible and effective strategy of self-regulation.

We also demonstrated that this can explain some well-known forms of apparently irrational behavior, such as indulgent behavior and dysfunctional goals.

There are several interesting extensions of our model we plan to pursue in future research. One possible direction is to examine the relationship between external and internal commitment strategies: we focused on internal strategies, but in practice this approach has its limitations. The model can be extended to include external commitment strategies, and can be used to analyze the interaction.

We also plan to extend the model to explain systematic deviations from goals. It appears goals are often set unrealistically optimistic. For instance, DellaVigna and Malmendier (2006) find that many members of the gym club overestimate attendance and end up paying more per visit in monthly fees than they would have had to pay using per visit fees. Burger and Lynham (2006) study individuals that place bets on loosing a certain amount of weight, and find that the vast majority loses their bet. Future work could shed light on this, by weakening the strong requirement of rational expectations, or otherwise.

8 Appendix

Proof of proposition 5. We first derive the threshold levels given in equations (6) and (7). Consider the strategy $(q_l, 0)$, hence never effort when costs are high, and effort with probability q_l when costs are low. At $t = 1$, $e = 0$ after low costs gives utility:

$$\rho q_l(\eta c_l - \eta\lambda\beta v), \tag{9}$$

whereas $e = 1$ yields:

$$-c_l + \beta v + \rho(1 - q_l)(-\eta\lambda c_l + \eta\beta v) + (1 - \rho)(-\eta\lambda c_l + \eta\beta v). \tag{10}$$

Equation (6) follows from indifference, thus equating the two above expressions. For any $q_l < \hat{q}_l$, (0,0) is preferred, while for any $q_l > \hat{q}_l$, (1,1) is preferred. Hence, (0,0) can be sustained as long as $\hat{q}_l \geq 0$, or:

$$\beta \leq \overline{\beta}_{00} \equiv \frac{c_l}{v}\frac{1 + \eta\lambda}{1 + \eta}, \tag{11}$$

and (1,1) can be sustained as long as $\hat{q}_l \leq 1$, or:

$$\beta \geq \underline{\beta}_{10} \equiv \frac{c_l}{v}\frac{1+\eta\lambda-\rho\eta(\lambda-1)}{1+\eta+\rho\eta(\lambda-1)}. \tag{12}$$

Similarly, given strategy $(1, q_h)$, $e = 0$ after high costs gives:

$$\rho(\eta c_l - \eta\lambda\beta v) + (1-\rho)q_h(\eta c_h - \eta\lambda\beta v), \tag{13}$$

while $e = 1$ yields:

$$-c_h + \beta v + \rho(-\eta\lambda(c_h - c_l)) + (1-\rho)(1-q_h)(-\eta\lambda c_h + \eta\beta v). \tag{14}$$

From the above two equations, equation (7) follows. Hence, (1,0) can be sustained as long as $\hat{q}_h \geq 0$, or:

$$\beta \leq \overline{\beta}_{10} \equiv \frac{1}{v}\frac{c_h(1+\eta\lambda)-c_l\rho\eta(\lambda-1)}{1+\eta+\rho\eta(\lambda-1)}, \tag{15}$$

and (1,1) can be sustained as long as $\hat{q}_h \leq 1$, or:

$$\beta \geq \underline{\beta}_{11} \equiv \frac{1}{v}\frac{c_h(1+\eta+\rho\eta(\lambda-1))-c_l\rho\eta(\lambda-1)}{1+\eta\lambda}. \tag{16}$$

From the perspective at $t = 0$, the epxected utility of (0,0) is zero, the expected utility of (1,0) is:

$$\rho[-c_l + v + (1-\rho)(-\eta\lambda c_l + \eta v)] + (1-\rho)\rho[\eta c_l - \eta\lambda v], \tag{17}$$

and of (1,1):

$$\rho[-c_l + v + (1-\rho)(-\eta(c_l - c_h))] + (1-\rho)[-c_h + v + \rho(-\eta\lambda(c_h - c_l))]. \tag{18}$$

If $(1-\rho)\eta(\lambda-1) \geq 1$, (0,0) is always preferred to (1,0), and if $(1-\rho)\eta(\lambda-1) < 1$, (0,0) is preferred to (1,0) if:

$$v \leq v_1 \equiv c_l\frac{1+(1-\rho)\eta(\lambda-1)}{1-(1-\rho)\eta(\lambda-1)}. \tag{19}$$

(1,1) is preferred to (1,0) if:

$$v \geq v_2 \equiv c_h - 2c_l\frac{\rho\eta(\lambda-1)}{1+\rho\eta(\lambda-1)}, \tag{20}$$

and (1,1) is preferred to (0,0) if:

$$v \geq v_3 \equiv \rho c_l + (1-\rho)c_h + \rho(1-\rho)\eta(\lambda-1)(c_h - c_l). \tag{21}$$

The following facts are easy to verify. First, if $\lambda = 1, \overline{\beta}_{00} = c_l/v$, $\underline{\beta}_{11} = c_h/v$, $v_1 = c_l$, and $v_2 = c_h$. Furthermore, $\overline{\beta}_{00}$ and v_1 are increasing in λ, and $\underline{\beta}_{11}$ and v_2 are decreasing in λ. Thus, as $\lambda \downarrow 1$, $v_1 < v_2$ and it is always possible to find a region with both under- and overdiligence, as in figure 3. As $\lambda \to 1+1/(\eta(1-\rho))$, $v_1 \to \infty$, so for some $\lambda < 1 + 1/(\eta(1-\rho))$, $v_1 = v_2 = v_3$. (1,0) is then always dominated by (0,0) for any larger λ. Comparing (0,0) and (1,1), note that v_3 is increasing in λ, and $v_3 = c_h$ for $\lambda = 1 + 1/(\eta(1-\rho))$. Also, $\overline{\beta}_{00} \geq 1$ at $v = c_h$ if $\lambda \geq (c_h/c_l)(1+\eta)/\eta - 1/\eta$, in which case (0,0) is sustainable for any $v \in [c_l, c_h]$ and any β.

Now, if $\max 1 + \frac{1}{\eta(1-\rho)} \geq \frac{c_h}{c_l}\frac{1+\eta}{\eta} - \frac{1}{\eta}$, then $\lambda^* = 1 + \frac{1}{\eta(1-\rho)}$. In that case, if $\lambda < \lambda^*$ then $v_3 < c_h$ and (1,1) is preferred for $v \in (v_3, c_h]$ and sustainable for some values of β. If $1 + \frac{1}{\eta(1-\rho)} < \frac{c_h}{c_l}\frac{1+\eta}{\eta} - \frac{1}{\eta}$, then $\lambda^* = \frac{c_h}{c_l}\frac{1+\eta}{\eta} - \frac{1}{\eta}$. In that case, if $\lambda < \lambda^*$ then $\overline{\beta}_{00} < 1$ and (0,0) cannot be sustained for some values of β. As long as $\lambda > 1$, (1,1) is the equilibrium strategy for some values of β. On the other hand, whenever $\lambda \geq \lambda^*$, $(0,0)$ is preferred to any other strategy for any $v \in [c_l, c_h]$ and sustainable for any β in that range.

Proof of proposition 6. Consider first perfect positively correlated costs. Comparing ex ante expected utility at $t = 0$, it is straightforward to derive that proximal goals yield higher utility if:

$$\psi \geq 1 - \left[\frac{\rho\eta(\lambda-1)}{1+\rho\eta(\lambda-1)}\right]\frac{c_L}{c_H}.$$

The RHS is smaller than 1, so this is satisfied for any $\psi \geq 1$. If costs are perfect negatively correlated, proximal goals yield higher utility if:

$$\psi \geq \frac{c_H + c_L + 3(c_H - c_L)\rho\eta(\lambda-1)}{2c_L(1+\rho\eta(\lambda-1))}.$$

It is easy to verify that the RHS is increasing in c_H and decreasing in c_L and λ.

References

[1] Ainslie, G. (1992) *Picoeconomics,* Cambridge, UK: Cambridge University Press.

[2] Ariely, D. and K. Wertenbroch (2002), “Procrastination, Deadlines, and Performance: Self-Control by Precommitment,” *Psychological Science*, 13, 219-224.

[3] Ashraf, N., D. Karlan, and W. Yin (2004), Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines, *Quarterly Journal of Economics* 635-672.

[4] Baumeister, R., F. Heatherton and D. Tice (1994), *Losing Control: How and why people fail at self-regulation*, Academic Press, San Diego.

[5] Bénabou R. and J. Tirole (2003) “Intrinsic and Extrinsic Motivation”, *Review of Economic Studies*, 70: 489-520.

[6] Bénabou R. and J. Tirole (2004) "Willpower and Personal Rules", *Journal of Political Economy* 112(4): 848-886.

[7] Benhabib, J., A. Bisin and A. Schotter (2007), Present-Bias, Quasi-Hyperbolic Discounting, and Fixed Costs, mimeo.

[8] Benartzi, S., and R. Thaler (2004), Save More Tomorrow^{TM}: Using Behavioral Economics to Increase Employee Saving, *Journal of Political Economy* 112(1), S164–S182.

[9] Brunnermeier, M. K. (2004) "Learning to Reoptimize Consumption at New Income Levels: a Rationale for Prospect Theory", *Journal of European Economic Association*, 2(1): 98-114.

[10] Burger, N., G. Charness and J. Lynham (2008), Three Field Experiments on Procrastination and Willpower, mimeo.

[11] Burger, N. and J. Lynham (2007), “Betting on Weight Loss… and Losing: Personal Gambles as Commitment Mechanisms,” mimeo.

[12] Carrillo and Mariotti (2000) "Strategic Ignorance as a Self-Disciplining Device." *Review of Economic Studies*, 67(3), 529-44.

[13] Dasgupta P. and E. Maskin (2005) "Uncertainty and Hyperbolic Discounting", *American Economic Review*, 95(4), 1290-1299.

[14] DellaVigna, S. and U. Malmendier (2006), “Paying Not to Go to the Gym,” *American Economic Review*, 96, 694-719.

[15] Frederick, S. (2005) “Cognitive Reflection and Decision Making”, *Journal of Economic Perspectives*, 19(4), pp. 25–42

[16] Frederick S., G. Loewenstein and T. O’Donoghue (2002) "Time Discounting and Time Preference: a Critical Review", *Journal of Economic Literature*, 40, 351-401.

[17] Fudenberg, D. and D. Levine (2006), A Dual Self Model of Impulse Control, Harvard Institute of Economic Research Discussion Paper No. 2112.

[18] Gul, F. and W. Pesendorfer (2001), Temptation and Self-control, *Econometrica* 69(6), 1403-1435.

[19] Heath, C., R. Larrick and G. Wu (1999) "Goals as Reference Points", *Cognitive Psychology*, 98, 79-109.

[20] Kahneman D. and A. Tversky (1979) "Prospect Theory: an Analysis of Decision under Rusk", *Econometrica*, 47, 263-291.

[21] Kahneman D. and A. Tversky (2000) *Choices, Values and Frames.*

[22] Kivetz, R. and I. Simonson (2002) "Self-Control for the Righteous: Toward a Theory of Precommitment to Indulgence", *Journal of Consumer Research*, 29, 199-217.

[23] Kőszegi B. and M. Rabin (2006) "A Model of Reference-Dependent Preferences", *Quarterly Journal of Economics*, 121(4), 1133-1166.

[24] Kőszegi B. and M. Rabin (2008), Reference-Dependent Consumption Plans, *American Economic Review*, Forthcoming.

[25] Laibson D. (1997) “Golden Eggs and Hyperbolic Discounting”, *Quarterly Journal of Economics*, 112, 443-478.

[26] List J. (2003) “Does Market Experience Eliminate Market Anomalies?” *Quarterly Journal of Economics*, 118(1), pp. 41-71.

[27] List J. (2004) "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace", *Econometrica*, 72(2): 615-625.

[28] Locke and Latham (1990) *A Theory of Goal Setting and Task Performance*, Prentice Hall Englewood Cliffs, NJ.

[29] Matsui, T., A. Okada, and R. Mizuguchi (1981), Expectancy theory prediction of the goal theory postulate: The harder the goals, the higher the performance, *Journal of Applied Psychology* 66, 54–58.

[30] Mick, G. and Demoss (1990), Self-Gifts: Phenomenological Insights from Four Contexts *Journal of Consumer Research* 17(3), 322-332.

[31] Odean, T. (1998) "Are Investors Reluctant to Realize their Losses?", *Journal of Finance*, 53(5), 1175-1798.

[32] O’Donoghue, T. and M. Rabin (1999), Doing it Now or Later, *American Economic Review* 89(1), 103-124.

[33] Plott C. and K. Zeiler (2005) "The Willingness to Pay–Willingness to Accept Gap, the “Endowment Effect,” Subject Misconceptions, and Experimental Procedures for Eliciting Valuations", 95(3), 530-545.

[34] Robson (2001) "The Biological Basis of Economic Behavior", *Journal of Economic Literature*, 39, 11-33.

[35] Robson A. (2002) "Evolution and human Nature", *Journal of Economic Perspectives*, 16(2), 89-106.

[36] Rubinstein A. (2003) "Economics and Psychology? The Case of Hyperbolic Discounting", *International Economic Review*, 44(4), 1207-1216.

[37] Samuelson L. and J. Swinkels (2006) "Information, Evolution and Utility", *Theoretical Economics*, 119-142.

[38] Shalev J. (2002) "Loss Aversion and Bargaining", *Theory and Decision*, 52, 201-232.

[39] Stigler G. and G. Becker (1977) "De Gustibus Non Est Disputandum", *American Economic Review*, 67(2), 76-90.

[40] Strotz, R. H. (1955), Myopia and Inconsistency in Dynamic Utility Maximization, *Review of Economic Studies* 23(3), 165-180.

[41] Thaler, R. H. and H. M. Shefrin (1981), An Economic Theory of Self-Control, *Journal of Political Economy* 89(2), 392-406.

[42] Tversky and Kahneman (1991) "Loss Aversion in Riskless Choice: A Reference-Dependent Model", *Quarterly Journal of Economics*, 106(4), pp. 1039-1061.

Nakhimovsky prospekt 47, office 720, Moscow, 117418
Tel. +7 495 105-5002 Fax +7 495 105-5003
e-mail:cefir@cefir.ru http://www.cefir.ru