Self-Control in Peer Groups

Marco Battaglini [1], Roland Bénabou, [2] and Jean Tirole [3]

This version: June 2003[4]

[1]Princeton University (mbattagl@princeton.edu)

[2]Princeton University (rbenabou@princeton.edu)

[3]IDEI, Université de Toulouse (tirole@cict.fr), CERAS and MIT.

[4]We are grateful for helpful comments to Leonardo Felli, Ted O'Donoghue, John Morgan, Michele Piccione, Matthew Rabin and Tom Romer, as well as to seminar participants at the Congress of the European Economic Association, the Studienzentrum Gerzensee, the Harvard–MIT theory seminar, the London School of Economics, and the Stanford Institute for Theoretical Economics. Bénabou gratefully acknowledges financial support from the National Science Foundation (SES–0096431) and the hospitality of the Institute for Advanced Study over the academic year 2002-2003.

Abstract

People with a self-control problem often seek relief through social interactions rather than binding commitments. Thus, in self-help groups like Alcoholic Anonymous, Narcotics Anonymous etc., members are said to achieve better personal outcomes by mainly sharing their experiences. In other settings, however, peer influences can severely aggravate individual tendencies towards immediate gratification, as is often the case with interactions among schoolmates or neighborhood youths.

Bringing together the issues of self-control and peer effects, we study how observing the behavior of others affects individuals' ability to resist their own impulses towards short-run gratification. In particular, these purely informational spillovers tend to make agents' choices of self-restraint or self-indulgence mutually reinforcing. We thus identify conditions on initial self-confidence, confidence in others, and degree of correlation that lead to either a unique "good news" equilibrium where social interactions improve self-discipline, a unique "bad news equilibrium" where they damage it, or the coexistence of both.

We conduct a welfare analysis of group membership and show that individuals will generally find social interactions useful for self-control only if they have at least a minimal level of confidence in both their own and their peers' ability to resist temptation. At the same time, having a partner who is "too perfect" is no better than being alone, since one learns nothing from his actions. The ideal partner is shown to be someone with a slightly worse self-control problem, as this makes his successes more encouraging, and his failures less discouraging.

Our paper thus provides a psychologically grounded theory of endogenous peer effects, as well as of the importance of group morale.

Keywords: peer effects, social interactions, clubs, self-control, willpower, addiction, time-inconsistency, memory, psychology.

JEL Classification: C72, D82, D71, D91, J24.

I Introduction

The behavioral and economic implications of dynamic inconsistency in the preferences of a single decision maker have been the focus of much recent work. Yet people are typically immersed in social relations, and often seek relief from their self-control problems by intensifying some of these interactions. Such is the case with self-help groups like Alcoholic Anonymous, Narcotics Anonymous and similar organizations through which agents appear to achieve better personal outcomes by simply sharing their experiences. Conversely, there are cases where group influence can further aggravate individual tendencies towards immediate gratification, as in some peer interactions among schoolmates or youths living in the same neighborhood.

In this paper we study how observing the actions of others affects individuals' ability to exercise self-control. The paper can thus be seen as bringing together the literature on self-control and that on peer effects. While obviously complementary –as any parent might attest– these two issues have until now largely evolved as different areas of economic inquiry (one bordering on psychology, the other on sociology).

Support groups are an important social phenomenon. Organizations like Alcoholics Anonymous, Narcotics Anonymous, Gamblers Anonymous, Debtors Anonymous and the like have branches in many countries, and millions of members. Yet there has been no formal work on how these groups may (or may not) help people deal with their self-control or addiction problems. Economists are used to thinking about how entering contracts or implicitly binding agreements with others allows agents to achieve desirable commitment. This, however, is not at all what self-help groups are about. Among the fourteen points listed under "What Alcoholics Anonymous does *not* do" (emphasis added), one thus finds:[1]

1. "Furnish initial motivation."
2. "Keep attendance records or case histories."
3. "Follow up or try to control its members."
4. "Make medical or psychological diagnoses or prognoses."
5. "Engage in education about alcohol."

Analogous statements can be found in the programs of similar organizations, making it clear that one cannot view these groups as just standard commitment devices: they not only cannot, but do not even want to "control" their members, or provide them with "initial motivation". Their scope is in fact explicitly limited to fostering informational interaction (discussion) among members. Thus in "What does Alcoholics Anonymous

[1] The following correspond to points 1, 4, 6, 7 and 10 respectively in A.A.'s list, which can be found at http://www.alcoholics-anonymous.org/.

do?" it is clearly stated that "A.A. members *share their experience* with anyone seeking help with a drinking problem" (emphasis added).

Once this is acknowledged, the impact on individuals' behavior of confronting their experiences with those of others remains an important open question. If such organizations are useful for overcoming self-control problems, or sometimes perhaps detrimental, one needs a theory to explain why. Such a theory of social interactions will in fact have a much wider applicability than just self-help groups. As mentioned above, peer influences and role models often play a critical part in young people's choices –particularly with respect to activities that are subject to episodes of intense temptation: smoking, drinking, drug abuse, sexual behavior, procrastination of studying effort, etc. In such settings peers may be "good influences" or "bad influences," and the latter scenario is often correlated with cases of low or fragile self-esteem. Similar issues arise in a couple where both partners are trying to quit smoking or lose weight, or in a sports team where an athlete may be encouraged or discouraged to "push himself harder" by observing what his teammates do. Finally, a theory of peer effects in self-control should be normative as well as positive. In many cases joining a group is a voluntary decision; but sometimes it is compulsory, as with a judge ordering an addict to attend a twelve-step program.[2]

In this work we take the first steps towards such a theory, by developing a reputational model that combines the dynamics of self-control with social learning. The presence of peers makes this a theoretically novel problem, taking the form of a signaling game with *multiple senders* of correlated types. To our knowledge this class of games has not been studied before, and our analysis yields results on strategic interactions that are more general than the specific application of this paper.

There are two fundamental assumptions in our analysis. First, agents have incomplete information about their ability to resist temptation, and must therefore try and infer it from past actions. The lack of direct access to certain aspects of one's own preferences and the key role played by *self-monitoring* in people's regulation of their behavior are both heavily emphasized in the psychology literature (Ainslie (1992), (2001); Baumeister et al. (1994)). We build here on Bénabou and Tirole's (2000) formalization of these phenomena, which is based on the idea that imperfect recall gives rise to a concern for *self-reputation* that can allow time-inconsistent individuals to achieve greater self-restraint. By breaking self-imposed rules such as diets, exercise regimens or abstinence resolutions, the individual would reveal himself, in his own (future selves') eyes, as weak-willed –that is, incapable of resisting temptation. Such a loss in reputation would further undermine his resolve in the future, to the point where he may abandon all attempts at self-restraint: what is

[2]The judge may be concerned about externalities from the agent's behavior on others, or take into account the agent's own tendency to procrastinate.

the point of abstaining from drinking today if, based on recent experience, one is most likely to relapse tomorrow anyway? The fear of creating precedents that would adversely impact future morale and behavior thus creates an incentive to maintain a clean "track record," in order to influence one's future (selves') actions in a desirable direction.

The second key assumption, novel to this paper, is that agents' characteristics are correlated, so that there is also something to be learned from observing others' behavior. This is in fact considered to be an essential element in the success of support groups and similar programs, which are typically mono-thematic: alcohol, narcotics, anorexia, debt, depression, etc. The idea is that members are linked together by a common problem, and that sharing their experiences is useful. Thus, Alcoholic Anonymous clearly states that:

> *The source of strength in A.A. is its single-mindedness. The mission of A.A. is to help alcoholics. A.A. limits what it is demanding of itself and its associates, and its success lies in its limited target. To believe that the process that is successful in one line guarantees success for another would be a very serious mistake.*

In fact, "anyone may attend open A.A. meetings. But *only* those with *drinking* problems may attend *closed* meetings or become A.A. members" (italics in the original text).[3]

Observing the decisions of people similar to oneself is a source of additional information about the manageability or severity of the self-control problem. The information may turn out to be good news, if the others are observed to persevere (stay "dry", "clean," etc.), or bad news, if they are observed to cave in or have a relapse. When deciding whether or not to exercise costly self-restraint in the face of temptation, the individual will take into account the likelihood of each type of news, and how it would impact the "reputational return" on his own good behavior. Therefore, in addition to his self-confidence, a key role will now be played by his assessment of his peers' ability to deal with their own self-control problems, and of the degree to which they are correlated with his own. The fundamental difference with the single-agent case, however, is that the informativeness of others' actions is endogenous, and depends on everyone's equilibrium strategies. As a result of these strategic interactions our model, in which *peer effects are purely informational,* can give rise to multiple equilibria where agents' choices of self-restraint or self-indulgence are mutually reinforcing. More generally, we identify conditions on agents' initial self-confidence, confidence in others, and correlation between types (difficulty of the self-control problem) that uniquely lead to either a "good news" equilibrium where group

[3]Both quotations are from the official web page of Alcoholics Anonymous. Task-specific informational spillovers are also evident in Weightwatchers' groups' practice of weighing members each week and reporting to each one not just his or her own loss or gain, but also the average for the group.

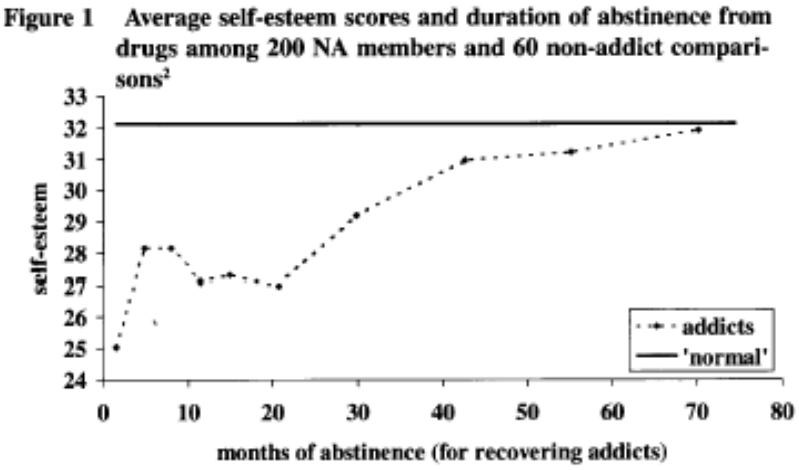

Figure 1: This plot is taken from Christo (1999). See footnote 5 for details.

membership improves self-discipline, a "bad news" equilibrium where it damages it, or to both.

In the first part of the paper we focus on a symmetric situation where individuals are ex-ante identical in all respects. We derive the equilibria of the joint signaling game, study their comparative statics, and conduct a welfare analysis to determine when group membership is preferable to, or worse than, staying alone. Three main results emerge. First, the situations under which peer influences improve, or on the contrary worsen, self-restraint, are fully characterized. Second, group interactions are beneficial only when peers' initial self-confidence is *above a critical level*; below that, they are actually detrimental. When beneficial, moreover, the peer group is not a mere commitment device: the welfare improvement occurs not only ex ante, but even ex post, inducing a *Pareto superior equilibrium* in which all types (weak and strong-willed) are better off. Third, as the degree of correlation between agents rises, self-restraint and welfare improve in the good news equilibrium, but deteriorate in the bad news equilibrium. At the same time, the range of initial beliefs for which both coexist tends to grow, creating a trade-off between the potential benefits from joining a community that shares common experiences and the ex-ante ambiguity of the outcome.[4]

In the second part of the paper we extend the analysis to heterogeneous "clubs". Are peers with a less severe self-control problem always more desirable? Would group members admit into their ranks someone who is even more susceptible to temptation

[4]This may explain the importance of a group's coordinator, whose "leadership" role would not necessarily stem from any superior abilities (greater trustworthiness, better judgement, etc.), but primarily from the need for coordination among peers.

than themselves? We establish a novel and even somewhat surprising –but in fact quite intuitive– result: the ideal peer is someone who is perceived to be *somewhat weaker* than oneself, in the sense of having a potentially worse self-control problem. Indeed, this somewhat pessimistic prior on one's partner makes his successes more encouraging, and his failures less discouraging: "if *he* can do it, then so can I. " More generally, we show that individuals value the "quality" of their peers *non-monotonically,* and will want to match only with those whom they expect to be neither too weak nor too strong. These results stand in sharp contrast to those of sorting or social-interactions models based on a priori specifications of agents' interdependent payoffs. Whereas these typically imply monotone comparative statics, our analysis of learning-based spillovers reveals a general tradeoff between the *likelihood* that someone else's behavior will be a source of encouraging or discouraging news, and the *informativeness* of these news.

As these discussions make clear, the dynamics of self-confidence play a key role in our theory of peer effects. First, self-restraint by one member today (e.g., abstinence) improves both his and the others' self-confidence tomorrow, and this in turn leads to more self-restraint by all in the future. Second, another of our main results is that agents will find self-help groups useful (and therefore join or remain in them) only if they have at least a minimal level of confidence in their own and their peers' ability to resist the temptation to relapse. While there is no systematic empirical literature on the subject, field studies of self-help groups have documented some robust stylized facts which our model fits rather well. This empirical evidence will be discussed in Section III.3, but its essence can already be conveyed by Figure 1, which plots the levels of self esteem of a random sample of Narcotics Anonymous members against their "clean time" in the group; a strong positive correlation between the two variables is clearly apparent.[5] Such findings lead Christo and Sutton (1994) to conclude that

> *"Addicts with greater cleantime tend to have lower anxiety and higher self-esteem. The presence of such successful individuals is likely to have a positive influence on newer Narcotics Anonymous members, helping to create an ethos of optimism and self-confidence."*

I.1 Related literature

Our paper connects two lines of research. First, there is now in economics a substantial empirical and theoretical literature on "peer effects". Many studies have found an

[5] Each point in Figure 1 represents the mean self-esteem score for 20 recovering addicts sampled within each of 10 ranges of abstinence time. For comparison, the mean score of 20 non-addicted students is shown as a horizontal thick line. The mean score of the 20 recovering addicts who have been abstinent for over 5 years is similar to the student comparison group.

influence of group characteristics on individual youths' behavior, whether in terms of academic achievement, drug use, teen pregnancy, employment, criminal activity and the like (e.g., Coleman (1988), Henderson et al. (1978), Dynarsky et al. (1989), Crane (1991), or Case and Katz (1991)). While many of the earlier studies regressing individual outcomes on group averages and observable individual controls were potentially subject to a self-selection bias, more recent ones exploiting "natural experiments" and random assignments have confirmed the existence of peer effects in academic and "lifestyle" decisions (e.g., Hoxby (2001), Sacerdote (2001)). Econometric studies are clearly essential to assess the existence and incidence of peer influences, but say little about how or why such effects occur. Similarly, nearly all the theoretical literature takes the existence of local complementarities as its starting assumption, and then explores what these imply for the equilibrium and optimal composition of groups. Thus, De Bartolome (1990) and Bénabou (1993) study how peer or neighborhood effects shape the functioning of a city and its schools; Bernheim (1994) examines how a concern for others' views of oneself leads to conformity; Brock and Durlauf (2001) and Glaeser and Scheinkman (2000) study how non-market interactions can lead to "social multipliers" and multiple equilibria.[6] The only previous work endogenizing peer effects is Banerjee and Besley's (1990) model of student testing, where complementarity arises from a common shock to the mapping between efforts and grades.[7] This paper and ours thus share the basic objective of deriving peer effects from informational externalities. On the other hand, the mechanism which they study is specific to a particular setting (student testing) and technology. It does not apply to smoking, drinking, crime and other impulsive behaviors where peer effects seem very important;[8] relatedly, it has the feature that being with peers –even bad ones– is always better than being alone.

The other literature to which our paper belongs is the one on self-control problems, due in particular to non-exponential discounting.[9] More specifically, a recent line of research has emphasized how the combination of self-control and informational concerns can account for a rich set of apparently irrational behaviors documented by psychologists, which we can gather under the heading of "motivated cognition". Carrillo and Mariotti (2000) established the important result that time-inconsistent individuals may have, ex ante, a negative value for information. Incorporating the idea of imperfect memory, Bénabou and

[6] See also Brueckner and Lee (1989) and Scotchmer (1994) for an analysis of related issues.

[7] The idea is that if a student's classmates work hard their grades will be more informative about the (unknown) difficulty of the test, which will then make his own grade more informative about his personal ability (which he seeks to discover). A more discriminating test, in turn, increases the informational return to his studying effort.

[8] See O'Donoghue and Rabin (2000) for a discussion of some of the relevant psychology literature.

[9] See, e.g., Strotz (1955), Phelps and Pollack (1968), Akerlof (1991) and Laibson (1994).

Tirole (2002) provide a theory of rational self-deception through endogenously selective recall. Building on the work of Ainslie (1992), they then develop in (2000) a model of self-reputation to explain the workings of personal rules such as diets, resolutions or moral principles. A closely related line of work by Bodner and Prelec (1997, 2001) examines the informational value of actions in a split self (ego-superego) model where the individual has "metapreferences" over his own, imperfectly known, tastes. Finally, our present work shares with Brocas and Carillo (2001) a concern with direct interactions between time-inconsistent agents. That paper analyzes how competition in the form of "patent races" can improve, and cooperation in joint projects worsen, individuals' tendency to procrastinate. It is thus very well suited to studying workplace, career and perhaps political interactions, but not peer influences in drug or alcohol abuse, sexual behavior or studying effort by youths. Furthermore, the strategic interdependence between agents is taken as a primitive of the problem. In our model *no individual's action directly enters another one's payoff,* so all externalities arise *endogenously* from inferences among peers who observe each other's behavior.

It should be clear, however, that our aim here is not to provide an all-purpose model of peer effects. We thus do not deny the presence in some settings of complementarities in payoffs, and recognize that groups also involve other forms of learning or signaling than those we focus on (e.g., in a seminar). But, as evident from the literatures on addiction, self-help organizations, and youths' risky behaviors, the interplay of peer influences and self-control is widely recognized as an important problem. This is therefore the *main object of our study,* leading us to endogenize peer effects in a model that brings together social learning and time inconsistency.

The paper is organized as follows. In Section 2 we present the model. In Section 3 we study symmetric equilibria and their welfare implications. In Section 4 we extend the analysis to asymmetric settings and equilibria. Proofs are gathered in the appendix.

II The model

II.1 Willpower and self-reputation

We start from the problem of a single decision maker who is uncertain about his own willpower, as in Bénabou and Tirole (2000). The canonical example is that of an alcoholic who must decide every morning whether he will try to abstain that day, or just start drinking right away. If he was sure of his ability to resist his cravings throughout the day and evening, when temptation and stress will reach their peak, he might be willing to make the effort. If he expects to cave in and get drunk before the day's end anyway, on

the other hand, the small benefits of a few hours' sobriety will not suffice to overcome his initial proclivity towards instant gratification, and he will just indulge himself from the start.

Formally, we consider an individual with a relevant horizon of two periods (the minimum for reputation to matter), $t = 1, 2$. Furthermore, each period 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" activity (NW), which will yield a known payoff a in subperiod I. This corresponds to indulging in immediate gratification (drinking, smoking, eating, shopping, slacking off, etc.) *without even trying* to resist the urge.[10]

2) Undertaking a "willpower-dependent" project or investment (W): attempting to exercise moderation or abstinence in drinking, smoking, eating, or buying; or taking on a challenging activity: homework, exercising, ambitious project, etc. Depending on his strength of will relative to the intensity of temptation that he will experience –as described below– the individual will then, at he beginning of subperiod II, either opt to *persevere* until completion (P), or *give up* along the way (G).

Perseverance entails a "craving" cost $c > 0$ 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 $a < b < B$. The assumption that $b > a$ means that *some* self-restraint (resisting for a while but eventually giving up) is better than none at all.

As explained below, c takes values c_L or c_H for different types of individuals, and is only imperfectly known by the individual himself. We assume that $c_H < B - b$, so that *ex ante,* attempting and then persevering in self-restraint would be the efficient action regardless of type.

The agents we consider, however, face a recurrent self-control problem which will often cause them to succumb to short-run impulses at the expense of their long-run interests. To formalize this weakness of will or temptation problem we assume that, in addition to a standard discount rate δ between periods 1 and 2, their preferences exhibit time-inconsistency, represented by the usual quasi-hyperbolic specification. Thus:

1) When deciding whether to choose the NW activity or to 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 to be expected from the second option, starting in the

[10]Note that NW need not yield a flow payoff *only* in subperiod I: a could be the present value, evaluated as of (t, I), of an immediate payoff plus later ones. The important assumption is that there be *some* immediate reward to choosing NW.

next sub-period. Accordingly, the agent discounts the latter at a rate $\beta < 1$; equivalently, he values the immediate gratification from NW at a/β 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. The immediate pain of craving looms larger than the benefits to be expected by not caving in, so in his decision-making the cost c gets magnified to c/β. Equivalently, all future payoffs are discounted by β[11]

The second critical assumption is that the individual has only limited knowledge of the intensity of the temptation to which he will be subject if he attempts to resist his impulses (W activity), and therefore of the ultimate outcome of his effort at self-control. This craving disutility is *revealed only through the experience* of actually putting one's will to the test. It cannot be accurately known in advance, nor reliably recalled through pure introspection or memory search. As a result, the agent in period 2 will have to try and *infer* his vulnerability to temptation from his actual behavior when confronted with it (self-monitoring): "how did I behave last night, and what kind of a person does that make me?" We shall discuss this assumption of persistent imperfect self-knowledge in more detail below. First we state it formally, and show how it combines with imperfect willpower (hyperbolic preferences) to generate an internal reputational "stake" in good behavior.

Recall that the intensity of temptation during the craving stage is equal to c/β, where the cost c is specific to the activity in question while β measures the individual's general ability to defer gratification. We seek here to emphasize the spillovers in self-learning that take place across agents facing similar challenges, rather than the role of willpower as a general trait. In particular, we observe that "self help" groups are typically monothematic: Alcoholics Anonymous, Narcotics Anonymous, Debtors Anonymous, etc. The same is true for students in the same grade or school.[12] We therefore assume that agents know their general degree of present-orientation, and for simplicity we take it to be the same $\beta < 1$ for everyone (similarly with δ). By contrast, the cost of perseverance c is imperfectly known, and potentially correlated across agents.

We take susceptibility to cravings to be a fixed characteristic of the individual, equal to either $c = c_L$ or $c = c_H$, with $c_L < c_H$. The low-cost type will also be referred to as a "strong-willpower type," or just a "strong type"; a high-cost individual will be referred

[11] In general, the salience or willpower parameters in the first and second subperiods could be different, since the individual's physical state (e.g., accumulated stress) and environment (presence of temptations) are likely to have changed; see Bénabou and Tirole (2000). For notational simplicity, we take them to be the same.

[12] Of course there are also other groups, whether voluntarily chosen or not, where learning about and building general "character" may be more important: families, churches, etc.

to as a "weak type." At the start of period 1 the agent initially does not know his type, but only has priors ρ and $1-\rho$ on c_L and c_H. Depending on whether he perseveres or gives in during period 1 (and, in a group context, depending also on what his peers do), this self-reputation gets updated up or down to ρ' at the beginning of period 2.

The two key psychological features of the problem that we study, namely the divergence in preferences between an individual's date-1 and date-2 selves (self-control problem) and the second self's lack of direct access to earlier preferences (imperfect recall), thus result in a simple *signaling game between temporal incarnations.* The presence of peers will add a *social dimension,* with signaling taking place across individuals as well.

We assume that resisting temptation is a dominant strategy for the low-cost (or strong) type. The high-cost (or weak) type, by contrast, would prefer to cave in, *if* he was assured that this would have no effect on his future behavior. Thus:

$$\frac{c_L}{\beta} < B - b < \frac{c_H}{\beta}. \tag{1}$$

If, on the other hand, a display of weakness today sets such a bad precedent that it leads to a complete loss of self-restraint tomorrow (a sure switch from W to NW), the weak type prefers to resist to his short-run impulses:[13]

$$\frac{c_H}{\beta} < B - b + \delta(b-a), \tag{2}$$

where the maximum reputational "stake" $b-a>0$ reflects the fact that even partial self-restraint (choosing W, but eventually defaulting to G) is better than none (choosing NW at the outset).

Turning now to the choice between W and NW at the start of period 2, it is clear that the individual will only embark on a course of self-restraint when he has sufficient confidence in his own fortitude: at that time, there are no longer any reputational concern that might cause the weak type to restrain his behavior. Therefore, the expected return from attempting W exceeds the immediate (and more salient) payoff from NW only if the individual's updated self reputation ρ' is above the threshold ρ^* defined by:

$$\rho^*(B-c_L) + (1-\rho^*)b \equiv \frac{a}{\beta}. \tag{3}$$

We assume $B - c_L > a/\beta > b$, so that $\rho^* \in (0,1)$. Note how, due to $\beta < 1$, the individual is always too tempted to take the path of least resistance, and not even attempt to exercise willpower: the ex-ante efficient decision would instead be based on a comparison

[13] The precedent-setting role of lapses in a personal rule as is emphasized by Ainslie (1992). Baumeister et. al. (1994) refer to it as "lapse-activated snowballing," and Elster (2001) as a "psychological domino effect". The self-reputation model presented here gives it a precise, formal content.

of $\rho'(B-c_L)+(1-\rho')b$ and a. A higher level of confidence ρ' in one's ability to successfully resist temptation is then a *valuable asset,* worthy of protection (self-restraint) in the previous period, because it helps offset the individual's natural tendency to "give up without trying". In particular, the distortion in the future self's preferences due to $\beta < 1$ creates an incentive for the weak type to *pool* with the strong one by persevering in the first period, so as to at least induce partial self-control in the second period.

We now come back to the assumption that the intensity of temptation c (more generally, c/β) is known only through direct experience, and cannot be reliably recalled in subsequent periods. First, a craving is by definition a "hot," internal, affective state, which is hard to remember later on from "cold" introspection. This is not only intuitive, but in line with experimental and field evidence on subjects' recollections of pain or discomfort (Kahneman et al. (1997)) and their (mis)predictions of how they will behave under conditions of hunger, exhaustion, drug or alcohol craving, or sexual arousal (Loewenstein and Schkade (1999)).[14] Second, the agent will often have, ex post, a strong incentive to "forget" that he was weak-willed, and "remember" instead that he was strong. Indeed, there is a lot of evidence that people's recollections of their past actions and performances are often *self-serving:* they tend to remember (be consciously aware of) 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.[15] In such circumstances of imperfect or even self-serving recall, introspection about one's vulnerability to temptation is unlikely be very informative, compared to asking *what one actually did* –a "revealed preference" approach familiar to economists.

The idea that individuals learn about themselves by observing their own choices, and conversely make decisions in a way designed to achieve or preserve favorable self-images, is quite prevalent in psychology (e.g., Bem (1972)). It is also supported by experimental evidence, such as Quattrone and Tversky's (1984) findings that subjects who were led to believe that tolerance for a certain kind of pain (keeping one's hand in near-freezing water) was diagnostic of either a good or a bad heart condition reacted by, respectively, extending or shortening the amount of time they withstood that pain.

II.1.1 Correlation in self-control problems

The central feature of our paper is that, instead of confronting his self-control problem alone, the agent is immersed in a social relationship where he can observe the actions of others. In some cases, like that of Alcoholic Anonymous, his exposure to a particular

[14] Loewenstein (1996) terms this kind of phenomenon "hot–cold empathy gaps".

[15] See Bénabou and Tirole (2002) for references and a model showing how the selectivity of memory or awareness arises endogenously in response to either a self-control problem or a hedonic value of self-esteem.

	c_L	c_H
c_L	$\rho^2 + \alpha\rho(1-\rho)$	$(1-\alpha)\rho(1-\rho)$
c_H	$(1-\alpha)\rho(1-\rho)$	$(1-\rho)^2 + \alpha\rho(1-\rho)$

Figure 2: The distribution of c and c'.

group may be a choice variable, decided upon by the agent himself or by some principal (judge imposing a sentence, social planner). In other situations, like public schools, there is no such choice and peer influences are essentially unavoidable. We shall therefore be interested in how group interactions affect self-discipline from both a positive and a normative viewpoint.

What makes the actions of others relevant is that agents face the same problem (trying to stay "dry" or "clean," to graduate, etc.), and the costs and rewards of perseverance are likely to be correlated among them. By observing B's actions, A can thus learn something about himself, or more generally alter his own expectations of success. If B successfully resists temptation these news are encouraging to A, while if B caves in or has a relapse they are discouraging. But since B himself is trying to control his own impulses, the informativeness of his actions about the underlying costs depends on his equilibrium strategy. The same is true for A's actions, leading to a strategic interdependence via informational spillovers that will turn out to have important consequences.

Figure 2 describes our simple model of the joint distribution for the costs (c, c') faced by two agents. The first parameter, ρ, is the prior probability of being a low cost (or high-willpower) type. The second one, α, is a measure of correlation –as evident in the conditional probabilities:

$$\begin{aligned} \pi_{LL} &\equiv \Pr(c' = c_L \,|\, c = c_L) = \rho + \alpha(1-\rho), \\ \pi_{HH} &\equiv \Pr(c' = c_H \,|\, c = c_H) = 1 - \rho + \alpha\rho, \end{aligned}$$

For $\alpha = 0$ we get back the single-agent case (types are independent), while for $\alpha = 1$ correlation becomes perfect. This simple stochastic structure also has the advantage that changes in α leave the unconditional probabilities unchanged, and vice-versa. This will allow comparative statics which cleanly separate the effects of initial reputation and of correlation. Finally, note that we have assumed a completely symmetric situation, in

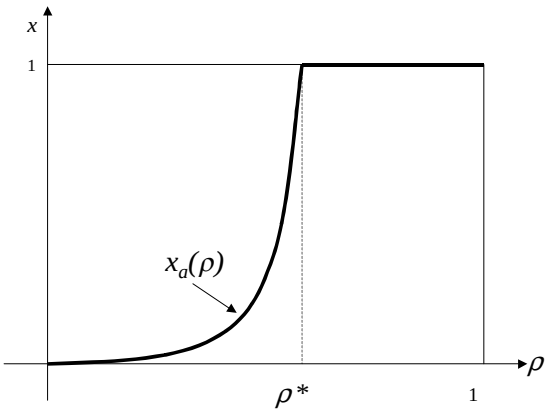

Figure 3: Self–restraint in the single-agent case

terms of both the unconditional and the conditional probabilities.[16] In particular, the two agents enter the game with the *same level of self-confidence* in their willpower, ρ, and this is common knowledge. Correspondingly, we shall focus attention on symmetric equilibria of this symmetric game. These assumptions simplify the analysis and the notation, but are not critical to our main results or the underlying insights. In Section IV, we shall in fact extend the analysis to asymmetric initial conditions and equilibria.

III Homogeneous peer groups

III.1 Main intuitions

1. The single-agent benchmark. We first briefly review the results from the one-agent case, as it provides an intuitive starting point to understand group interactions and evaluate their welfare implications. Given that the strong type always perseveres, the key question is whether, by also resisting temptation (choosing P), the weak type can induce his future self to opt for the willpower action. The basic result is illustrated on Figure 3, which shows how his ability to use such a pooling strategy is limited by the individual's initial self-confidence.[17] If Self 1 plays P with probability one, observing P is completely uninformative for Self 2, and his priors remains unchanged. Complete self-restraint (perfect pooling) is therefore an equilibrium only when the agents' initial

[16] This is why there is no need to index π_{LL} or π_{HH} by the identity of the agent on whose cost one conditions.

[17] The figure describes the weak type's strategy in the subgame where the decision node between P and G has been reached. This initial confrontation with cravings could be the result of an initial choice by the agent (requiring that initial self-confidence not be too low), of accidental circumstances (e.g., no alcohol or cigarettes were on hand that morning), or of a constraint imposed by someone else.

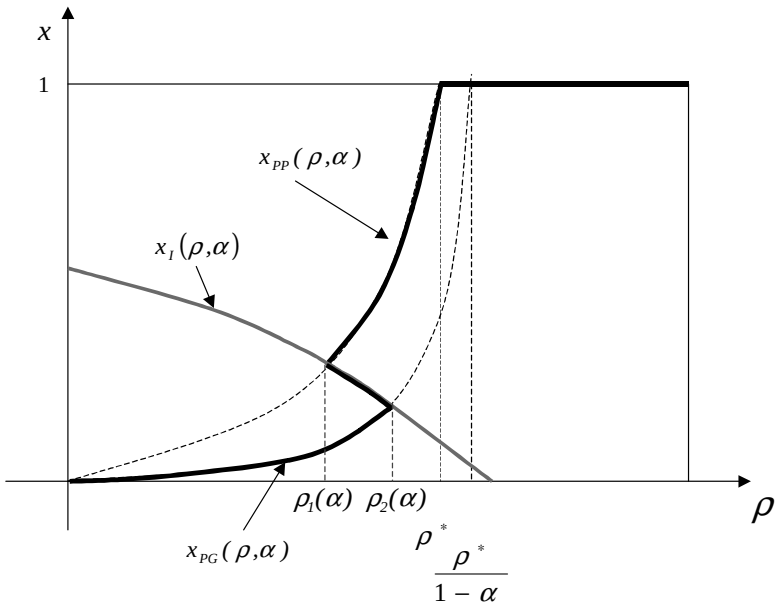

Figure 4: Equilibrium self-restraint for a moderate level of correlation

reputation ρ is above ρ^* defined in (3); in that case, choosing P successfully induces Self 2 to play W with probability one. When self-confidence is below ρ^*, however, Self 2 is more distrustful, and responds to an observation of P by selecting W only with a probability sufficiently small to eliminate the weak type's incentive to cheat (making him indifferent between playing P and G). Furthermore, the weak type's probability of pooling must be low enough that observing P is sufficiently good news to raise Self 2's posterior from ρ to ρ^*, where he is willing to randomize between W and NW. This *informativeness constraint*, $\Pr_{x,\rho}(c = c_L \mid P) = \rho^*$, uniquely defines the equilibrium strategy of the (weak) single agent as an increasing function $x_a(\rho)$, which starts at the origin and reaches 1 for $\rho = \rho^*$ (the subscript a stands for "alone").

2. Two agents with moderate correlation. Let us now bring together two individuals whose willpower types (craving costs) are correlated as in Table 1, and examine how this affects the behavior of weak types at the temptation stage. As mentioned earlier, we focus until Section 4 on the case where the two agents, denoted i and j, have the same initial self-reputation $\rho^i = \rho^j = \rho$, and play the same strategy $x^i = x^j = x$. A decision by one agent to persevere may now lead to two different states of the world: either the other agent also perseveres (event PP), or he gives in (event PG).

To build up intuition, let us first assume the correlation between types is relatively mild (α is small). By continuity, equilibrium behavior will not be too different from that of the single agent case; the issue is thus not by how much x changes, but the direction

in which it changes. The key new element is that *the expected return to resisting one's impulses now depends on what the other agent is likely to do,* and on how informative his actions are. Suppose, for instance, that agent i discovers himself to be tempted (a weak type), and consider the following three situations, corresponding to different ranges of ρ on Figure 3.

a) When initial reputation is low, j is most probably also a weak type, and will play a strategy close to $x_a(\rho) \approx 0$. Consequently, he is almost sure to be a source of "bad news," meaning an observation of G that will reduce i's hard-earned reputational gain from playing P. This *discouragement* effect will naturally lead agent i to persevere with lower probability $x_{PG}(\rho;\alpha) < x_a(\rho)$, as indicated by the thick curve emanating from the origin in Figure 4. Intuitively, agent i must now must counterbalance the bad news from j by making his own perseverance a more credible signal of actual willpower; this requires pooling with the strong type less often, which means adopting a lower x. The x_{PG} locus to which x adjusts thus corresponds again to an "informativeness constraint," along which Self 2's posterior following an observation of PG just equals ρ^*. This is why the equilibrium strategy x_{PG} rises with ρ, as in the single-agent case. The slope is less steep, however, due to the correlation of types: since i knows himself to be weak, and $\alpha > 0$, a higher prior ρ is only moderately better news about the other player.

b) When initial reputation is high –just below ρ^* or even above it– a similar reasoning indicates that being in a group will *increase* the equilibrium probability of perseverance, relative to the single-agent case. Indeed, with j now either a strong type or a weak type who exerts self-controls with probability close to $x_a(\rho) \approx 1$, agent i's playing P is most likely to lead to an observation of PP, resulting in an extra boost to his self-confidence and propensity to choose the willpower activity. Due to this *encouragement* effect x increases, until equilibrium is reestablished by the weak type playing P with a probability $x_{PP}(\rho;\alpha) \geq x_a(\rho)$ (with strict inequality unless $x_a(\rho)$ was already equal to 1). This is illustrated in Figure 4 by the thick curve which rises up to $(\rho^*, 1)$, and along which Self 2's posterior following an observation of PP just equals ρ^* (second informativeness constraint). In this case, the positive externality allows the agent to choose a less informative strategy.

c) Finally, consider situations where ρ is in some intermediate range. Conditional on i's playing P, neither PG nor PP can now be treated as having probability close to one (or zero). Both events are relevant to i's payoff, and which one ends up shaping equilibrium strategies (i.e., which informational constraint ends up binding) is no longer pinned down by the initial reputation. Instead, this is where the strategic nature of interaction is determinant, resulting in *multiple equilibria.* Intuitively, the higher the x^j used by agent j, the more likely the event PP in which agent i gains from having played

P, relative to the event PG in which he loses; therefore, the greater is i's incentive to increase x^i. Conversely, if x^j is low relative to x_a, the resulting reduction in agent i's return to choosing P will drive x^i down. Due to this strategic complementarity (which operates purely through shared informational spillovers on the decision of Self 2), the two equilibria $x_{PP}(\rho;\alpha)$ and $x_{PG}(\rho;\alpha)$ described earlier coexist over some intermediate range of ρ; see Figure 4. As usual, a third equilibrium $x_I(\rho;\alpha)$ then also exists in-between; we shall describe it more detail below.

2. Increasing the correlation. So far we have considered the case of moderate correlation, which is easier to analyze because strategies are close to those of a single agent. As the correlation of the agents increases, the x_{PG} locus pivots down, while the x_{PP} locus pivots up: what one agent does becomes more informative for the other, reinforcing all the effects described above and making the strategic interaction stronger.

We shall now more formally analyze the informational and incentive effects outlined above, and fully characterize the resulting equilibrium set.

III.2 Equilibrium group behavior

1. The informativeness constraints. Let $\mu_{PG}(x;\rho\,,\alpha)$ denote the posterior probability that agent i is a strong type, given that he chose P in the first period but agent j chose G, and that weak types are assumed to play P with probability x. Similarly, let $\mu_{PP}(x;\rho\,,\alpha)$ be the posterior following a play of P by both agents. Since strong types always play P, we have $\mu_{PG} < \mu_{PP}$ for all $\rho > 0$. It is also easy to see that, in any equilibrium:

$$\mu_{PG}(x;\rho,\alpha\,) \leq \rho^* \leq \mu_{PP}(x;\rho\,,\alpha), \tag{4}$$

unless $\rho > \rho^*$ and $x = 1$, in which case the first inequality need not hold. Indeed, if both posteriors were below ρ^* Self 2 would never play W, therefore weak types would always act myopically and choose G. Observing P would then be a sure signal of strength, a contradiction. Similarly, if both posteriors are above ρ^* weak types will always play P, since this induces Self 2 to choose willpower with probability one. But then priors remain unchanged, requiring $\rho > \rho^*$.[18] Naturally, both posterior beliefs are non-decreasing in the prior ρ. They are also non-increasing in x, since more frequent pooling by the weak type makes a signal of P less informative. Equation (4) thus defines two upward-sloping loci in the (ρ, x) plane, between which any equilibrium must lie:

$$x_{PG}(\rho;\alpha) \leq x \leq x_{PP}(\rho;\alpha), \tag{5}$$

[18] Formally, $\mu_{PP}(1;\rho\,,\alpha) = \rho$, requiring $\rho > \rho^*$. The event PG has zero probability and can be assigned any posterior in (ρ^*, ρ).

where:

$$x_{PP}(\rho;\alpha) \equiv \max\{x \in [0,1] \mid \mu_{PP}(x;\rho,\alpha) \geq \rho^*\}, \quad (6)$$

$$x_{PG}(\rho;\alpha) \equiv \min\{x \in [0,1] \mid \mu_{PG}(x;\rho,\alpha) \leq \rho^*\}. \quad (7)$$

We shall refer to these two curves as the *informativeness constraints* in the "good news" state PP and the "bad news" state PG, respectively. As illustrated on Figure 4, x_{PP} increases with ρ up to the threshold $\rho = \rho^*$, after which it equals 1. Along the increasing part, we have $\mu_{PP} = \rho^*$: the weak type is just truthful enough (x is just low enough) to maintain Self 2's posterior following the good news PP equal to ρ^*. In other words, he exploits these good news to their full extent. Above ρ^* the constraint $\mu_{PP} \geq \rho^*$ in (4) is no longer binding, allowing complete pooling. A similar intuition underlies the x_{PG} locus, which increases with ρ up to $\min\{\rho^*/(1-\alpha), 1\}$, and then equals 1. Along the increasing part, $\mu_{PG} = \rho^*$: the weak type is just truthful enough to exactly offset the bad news from the other player and maintain Self 2's posterior following PG at ρ^*. Naturally, since for any given (x, ρ) observing the event PG is worse news about one's type than just observing oneself playing P (and, conversely, PP is better news), the single-agent equilibrium strategy x_a lies between x_{PG} and x_{PP}.

These results already allow us to classify possible equilibria into three classes:

I. *Good News equilibrium.* When the equilibrium lies on the x_{PP} locus, the agent in period 2 undertakes W with positive probability only after the event PP. Accordingly, each agent's strategy is shaped by the informational constraint in the pivotal state, namely PP.

II *Bad News equilibrium.* When the equilibrium lies on the x_{PG} locus, the agent in period 2 will undertake W with positive probability even after PG, and with probability 1 after PP. It is now the informational constraint in the bad news case which is relevant, as the agent's posterior belief must not fall below ρ^* even after the other agent caved in.

III *Extreme News equilibrium.* When the equilibrium lies strictly between the x_{PG} and x_{PP} loci, Self 2's beliefs following PG and PP fall on opposite sides of ρ^*, so he will follow a pure strategy: choose W after PP, and NW after PG.

2. The incentive constraint. We now determine exactly when each scenario applies. Together with the two informativeness constraints, the other key requirement for an equilibrium is the weak type's *incentive constraint:* in order for him to be willing to mix between playing P and playing G, the net utility gains he can expect in the event PP

must just compensate the net losses he can expect in the event PG. Similarly, for him to play $x = 1$ the expected gain across the two events must be positive.

Let therefore $\Pi(x, y, y'; \rho, \alpha)$ denote the *net* expected gains to a weak type of choosing P rather than G when he believes other weak agents to use strategy x, and expects his own Self 2 to choose the willpower activity with probability y following event PP, and with probability y' following event PG. Since a weak type will reap payoff b under W rather than a under NW, we have:

$$\begin{aligned}\Pi(x, y, y'; \rho, \alpha) \equiv\ & B - b - \frac{c_H}{\beta} \\ & + \delta\left[(1-\alpha)\rho + (1-(1-\alpha)\rho)\left(xy + (1-x)y\prime\right)\right](b-a). \qquad (8)\end{aligned}$$

Note that $1 - (1-\alpha)\rho = \pi_{HH}$ is the conditional probability that the other agent is also a weak type (high cost of perseverance). A particularly important role will be played by $\pi(x; \rho, \alpha) \equiv \Pi(x, 1, 0; \rho, \alpha)$, which corresponds to Self 1's payoff when Self 2 plays a pure strategy in both events. In particular, this is what happens in the third type of equilibrium described above. The weak type's indifference between P and G then requires that

$$\pi(x; \rho, \alpha) \equiv B - b - \frac{c_H}{\beta} + \delta\left[(1-\alpha)\rho + (1-(1-\alpha)\rho)x\right](b-a) = 0.$$

It is easily seen that this equation uniquely defines a downward-sloping locus $x_I(\rho; \alpha)$ in the (x, ρ) plane; we shall refer to it as the weak type's *incentive constraint.* Given (1)-(2), x_I starts strictly between 0 and 1 and cuts the horizontal axis at some $\tilde{\rho}(\alpha)$ which may be above or below 1, depending on parameters. The intuition for the negative slope is quite simple: in $\pi(x; \rho, \alpha)$, the arguments ρ and x refer to the reputation and strategy of the *other* agent, say j. The more likely it is that j will persevere (the higher ρ or x), the greater the probability that i's playing P will pay off ex-post (event PP) rather than lead to net losses (event PG). In order to maintain indifference, a higher ρ must thus be associated with a lower x. For the same reason, a greater correlation α must result in a higher $x_I(\rho, \alpha)$.

Putting these results together with the earlier ones shows that:

- Bad News equilibria correspond to the portion of x_{PG} locus that lies *below* the incentive locus $\pi(x; \rho, \alpha) = 0$. Indeed, as $y = 1$ following PP, Self 2's mixing probability y_{PG} following PG must be such that $\Pi(x, 1, y_{PG}; \rho, \alpha) = 0$. Since $\Pi(x, 1, 1; \rho, \alpha) > 0$ by (2), such a y_{PG} exists if an only if $\pi(x; \rho, \alpha) = \Pi(x, 1, 0; \rho, \alpha) \leq 0$.

- Good News equilibria correspond to the portion of the x_{PP} cure that lies *above* the incentive locus. Indeed, there must exist a mixing probability y_{PP} for Self

2 such that $\Pi(x, y_{PP}, 0; \rho, \alpha) = 0$. Since $\Pi(x, 0, 0; \rho, \alpha) \leq 0$ by (1), this requires $\pi(x; \rho, \alpha) = \Pi(x, 1, 0; \rho, \alpha) \geq 0$.

- Extreme News equilibria correspond precisely to the portion of the incentive locus x_I which is "sandwiched" between the two informational constraints x_{PG} and x_{PP}.

To summarize, the set of symmetric equilibria in the two-agent game corresponds to the "inverted Z" configuration shown in bold on Figure 4. Formally:

Proposition 1 *The set of equilibria is fully characterized by two threshold functions* $\rho_1(\alpha) : [0,1] \to [0, \rho^*]$ *and* $\rho_2(\alpha) : [0,1] \to [0, \rho^*/(1-\alpha)]$ *such that:*

1. *For* $\rho < \rho_1(\alpha)$ *there is a unique equilibrium, which is of the "bad news" type:* $x = x_{PG}(\rho; \alpha)$.

2. *For* $\rho > \rho_2(\alpha)$ *there is a unique equilibrium, which is of the "good news" type:* $x = x_{PP}(\rho; \alpha)$.

3 *For* $\rho \in [\rho_1(\alpha), \rho_2(\alpha)]$ *there are three equilibria, namely* $x_{PG}(\rho; \alpha)$, $x_I(\rho; \alpha)$, *and* $x_{PP}(\rho; \alpha)$.

Moreover, for any $\alpha > 0$, $\rho_1(\alpha) < \rho_2(\alpha)$, *but as correlation converges to zero, so does the measure of the set of initial conditions for which there is a multiplicity of equilibria:* $\lim_{\alpha \to 0} |\rho_2(\alpha) - \rho_1(\alpha)| = 0$.

Figure 5 provides a convenient representation of these results in the (ρ, α) space.[19] The horizontal axis thus measures the degree of correlation between the agents, and the vertical one their initial level self-confidence.

The set of parameters where multiplicity occurs is the area between $\rho_1(\alpha)$ and $\rho_2(\alpha)$. As correlation declines to zero it shrinks to a point, and in the limit we get back the unique equilibrium of the single-agent case. This is quite intuitive, since without correlation in preferences what the other agent does is irrelevant. Clearly, in our model all *the externalities are in beliefs, not in payoffs.* Thus, for any $\alpha > 0$ the set of ρ's giving rise to indeterminacy has positive measure, and as α rises it becomes quite large. Even for α very small, when there is essentially a unique equilibrium, it is not always of the same type: if ρ is high it is a Good News equilibrium, while if ρ is low it is a Bad

[19] We focus there on the equilibrium set for $\rho \leq \rho^*$, which is the interesting case. Above ρ^* there is always the Pareto-dominant $x_{PP} = 1$ equilibrium, plus possibly (when $\rho^*/(1-\alpha) < 1$) the x_{PG} and x_I equilibria.

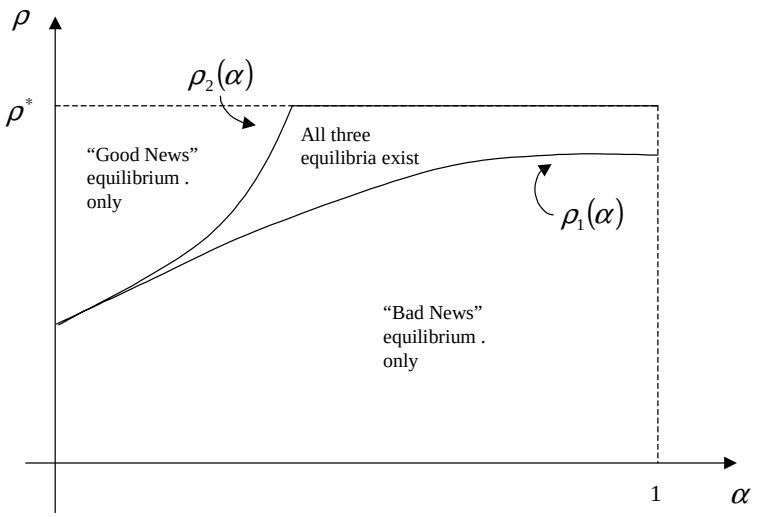

Figure 5: Equilibrium outcomes

News equilibrium. Although these equilibria converge to the same limit $x_a(\rho)$, they are qualitatively different in terms of welfare, as we shall see below.

The nature and shape of the thresholds $\rho_1(\alpha)$ and $\rho_2(\alpha)$ are easily understood from our earlier analysis of the informativeness and incentive constraints. As mentioned previously (and illustrated by the comparison of Figures 3 and 4), the function $x_{PP}(\cdot\,;\alpha)$ shifts up with α, while $x_{PG}(\cdot;\,\alpha)$ shifts down. Quite intuitively, greater correlation magnifies both the "discouragement" and the "encouragement" effects of the other agent's choosing G or P, respectively. The incentive constraint $x_I(\cdot,\alpha)$ shifts up with α: a greater likelihood that the other agent is also a weak type reduces expected profits $\pi(x;\rho,\alpha)$, and this must be compensated by a strategy that makes good news more likely (a higher x). From these results it is immediate that $\rho_2(\alpha)$, which is the intersection of $x_{PG}(\cdot\,;\alpha)$ and $x_I(\cdot\,;\alpha)$ is increasing in α. The slope of $\rho_1(\alpha)$ is ambiguous, however.

Note that the PG equilibrium exists only when ρ is not too high, or when α is large enough, making a weak agent relatively pessimistic about his partner's type. Otherwise the "dominant" event is PP, which in that equilibrium calls for Self 2 to choose willpower with probability one in the next period. But this implies that a weak type always gains from playing P, violating the incentive constraint. A higher degree of correlation thus makes PG easier to sustain. Similarly, to sustain the PP equilibrium ρ must be high enough, making a weak agent relatively optimistic about his partner's type.

III.3 Welfare Analysis

In this section we compare welfare levels across the equilibria that may arise in a group, and relate them to the single-agent benchmark. This last point is particularly important because it will show when groups do indeed provide valuable "help," and when they can actually do damage. Correspondingly, we will be able to identify situations in which people want to join such a group, or remain alone.

We shall evaluate welfare both ex ante and from the (interim) points of view of weak-willed and strong-willed individuals. The ex-ante approach is quite standard in the asymmetric information literature when there are different, a priori unknown types. On the other hand, in the present context it is also interesting to analyze welfare at the interim stage, i.e. after the agent's type is revealed (and even if he later forgets it). Suppose for instance that a judge must decide whether an agent with a substance abuse problem should be compelled to join a support group like Alcoholics Anonymous. First, the judge may have evidence on the agent's type (spousal complaint, police reports, etc.) that the agent either does not have or is in denial about; the judge will then want to maximize the agents' expected welfare, or perhaps social welfare, given that knowledge. Second, even behind a veil of ignorance the judge may be concerned about the fairness of the decision, in the sense of valuing differently the welfare of weak and strong types.

In computing interim welfare we shall be concerned with utilities which are not distorted by temporary impulses –and this for two reasons. First, these are the two basic components of ex-ante welfare, which simply weights them according to the individual's initial self-confidence. Thus, in deciding *whether to join a group,* the individual computes the (undistorted) payoffs W^s and W^w that he will reap if he turns out to be strong or weak, and then bases his membership decision on the weighted sum, $W = \rho W^s + (1-\rho)W^w$. Second, this focus on "long-term" welfare (without the temporary bias created by β at the temptation stage) is quite standard in the literature on hyperbolic discounting (e.g., Laibson (1994), O'Donoghue and Rabin (2000)).[20]

Throughout the analysis, we shall focus on the case where $\rho < \rho^*$, which is the more interesting one.[21] We also assume that in period the willpower activity is undertaken, so that there is indeed a decision to be made about whether to resist or succumb to temptation.[22] As a point of reference, we first consider welfare in the benchmark case

[20] One could also easily compute each type's utility as perceived at the moment of temptation (when it is revealed, or rather re-experienced), but this would not be very interesting, or informative. In particular, recall that in all the equilibria (for $\rho < \rho^*$) the weak type is, from this "short-term utility" point of view, indifferent between persevering and giving in. Therefore being in a group or not, and what happens in a group, makes no difference to him.

[21] When $\rho > \rho^*$, indeed, there is no self control problem.

[22] See footnote 17.

where the agent is alone (equivalently, $\alpha = 0$). Ex-ante welfare then equals $\rho W_a^s + (1-\rho) W_a^w$, where W_a^w and W_a^s are the weak and strong type's (interim) expected values of the subgame that starts just after the willpower activity been undertaken. For the weak type,

$$W_a^w = b + \delta a + x_a \left[B - b - c_H + \delta y_a (b - a) \right], \tag{9}$$

where x_a denotes his first-period perseverance strategy, and y_a the second-period self's probability of choosing the willpower option following P. Note how we have decomposed the payoff into what is obtained under G (which always reveals a weak type, and hence leads to NW next period), and the expected surplus that arises from choosing P. Next, for the weak type to be indifferent, it must be that $B - c_H/\beta + \delta \left[y_a b + (1 - y_a) a \right] = b + \delta a$; substituting into (9) yields:

$$W_a^w = b + \delta a + x_a \left(\frac{1-\beta}{\beta} \right) c_H. \tag{10}$$

In equation (9), c_H is not magnified by $1/\beta$ because we are evaluating "long term" welfare, that is, welfare without the temporary distortion in preferences that arises at the moment of temptation. By contrast, the indifference condition must hold at the temptation stage. The second term in (10) thus reflects the *value of the self-discipline* (resisting temptation) that is achieved through the reputational mechanism. Turning now to expected welfare for a strong type, we can write:

$$W_a^s = B - c_L + \delta \left[y_a (B - c_L) + (1 - y_a) a \right]. \tag{11}$$

Because $y_a < 1$ for all $\rho < \rho^*$, the strong type's average payoff per period is always less than $B - c_L$, which is what he would achieve under perfect information, or in a one-shot context.[23] He is thus hurt by the reputational game, whereas we saw that the weak type gains by achieving greater self-control. There is therefore a sense in which the strong type "cross-subsidizes" the weak type in this single-agent equilibrium.

We now turn to the two leading cases discussed above: welfare in the Good News equilibrium and in the Bad News equilibrium. Since the analysis of the "Extreme News" equilibrium is technically very similar, it is presented in the Appendix. Readers who are not interested in the derivation of the welfare results may want to go directly to Section III.3.3, which summarizes and discusses the main insights.

III.3.1 Welfare in a Good News equilibrium

From Proposition 1 we know that for $\rho > \rho_1(\alpha)$ there is always an equilibrium in which the weak type perseveres with probability x_{PP}, and in period 2 the willpower option is

[23] This loss is independent of $\rho < \rho^*$, since the mixing strategy y_a that makes a weak Self 1 indifferent is easily seen to have this property. The loss disappears for $\rho \geq \rho^*$.

chosen with positive probability y_{PP} only when *both* agents have persevered. The weak type's expected surplus is then described by

$$W^w_{PP} = b + \delta a + x_{PP}\left[B - b - c_H + \delta \Pr_{PP}(P \mid w)\, y_{PP}(b-a)\right], \tag{12}$$

where $\Pr_{PP}(P \mid w) = 1 - \pi_{LL} + \pi_{LL}\, x_{PP}$ denotes the probability that –in this PP equilibrium– player j will choose P, given that player i is a weak type. Using again the weak type's indifference condition $\pi(x_{PP}; \rho, \alpha) = 0$ to simplify this expression yields:

$$W^w_{PP} = W^w_a + (x_{PP} - x_a)\left(\frac{1-\beta}{\beta}\right) c_H. \tag{13}$$

From our earlier results we know that $x_{PP} > x_a$: in the Good News equilibrium, the (weak) agent achieves greater self-control than when left to his own devices. As result, his welfare is higher. Turning now to the strong type, we have:

$$W^s_{PP} = B - c_L + \delta a + \delta\, \Pr_{PP}(P \mid s)\; y_{PP}\,(B - b - c_L),$$

where $\Pr_{PP}(P \mid s) = \pi_{LL} + (1 - \pi_{LL})x_{PP}$ is the equilibrium probability that j will choose P, given that i is a strong type. Next, subtract (9) and note that for the weak type to be indifferent *both* after event P in the single-agent game and after event PP in a group setting, it must be that $y_a = y_{PP} \Pr_{PP}(P \mid w)$. Thus:

$$W^s_{PP} = W^s_a + \delta y_{PP}\left[\Pr_{PP}(P \mid s) - \Pr_{PP}(P \mid w)\right](B - a - c_L), \tag{14}$$

Thus, as long as $\alpha > 0$, the strong type is also strictly better off: $W^s_{PP} > W^a_s$. The intuition is straightforward. In the single-agent case, the probability that persevering will induce the willpower option W next period is the same for the weak and the strong types, because all that matters is the individual's own action. With two agents, however, the payoff to i's playing P is contingent on what agent j does, which in turn depends on j's type. Since being weak suggests that the other agent is also weak, a weak player i has a lower chance of seeing his perseverance pay off than a strong one. To maintain his willingness to persevere, this lower-odds payoff must be greater, meaning that the second-period self must choose W with higher probability than in the single agent case: $y_{PP} > y_a$. This yields no extra surplus for the weak type, who remains indifferent, but generates rents for the strong type. In summary, we have:

Proposition 2 *In the Good News equilibrium that exists for all (ρ, α) with $\rho > \rho_1(\alpha)$, joining a group is strictly better then staying alone from an interim point of view (i.e., for both types), and therefore also ex ante.*

What it is interesting here is the fact that joining a group can unambiguously improve the welfare of *both types,* rather than just transfer surplus from one to the other. This is somewhat surprising because, as mentioned earlier, joining a group generally entails a trade-off between the positive informational spillover received when the other agent perseveres, and the negative one suffered when he does not. In a PP equilibrium, however, the latter's impact on the weak type's welfare is just compensated by an increase in y_{PP}, relative to y_a. The positive spillover, meanwhile, allows each agent to engage in more pooling (increase x): even though each signal of P is now less informative, their concordance (event PP) remains sufficiently credible to induce the willpower action next period. In summary, the weak type benefits by achieving greater self-discipline in period 1, and the strong type gains from a greater exercise of willpower in period 2.

As seen earlier, however, such a virtuous equilibrium does not exist when initial self-confidence is too low; and even when it does, it may not be chosen due to coordination failure. We therefore now turn to the Bad News scenario (the intermediate case of the Extreme News equilibrium is given in the appendix).

III.3.2 Welfare in a Bad News equilibrium

The derivations in this case are very similar to the previous one, but yield substantially different results. For the weak type we have again:

$$W^w_{PG} = W^w_a + (x_{PG} - x_a) \left[\frac{1-\beta}{\beta} \right] c_H. \tag{15}$$

Since $x_{PG} < x_a$, the weak type is now worse off in a group, compared to staying alone. The intuition is simple: when the other agent gives in (state PG) this is bad news about one's own type. In order to *offset this damage,* the fact that one has persevered must be a more credible signal of being a strong type, which means that a weak type must exert self-restraint less often (x must be smaller). This, of course, only worsens the inefficiency from time-inconsistent preferences. Things are quite different for the strong type, however. Using the same steps as previously, we can write:

$$W^s_{PG} = W^s_a + \delta \left[\Pr_{PG}(P \mid s) - \Pr_{PG}(P \mid w) \right] (1 - y_{PG}) (B - a - c_L). \tag{16}$$

From this condition it is evident that the strong type is better off than staying alone, although whether by more or by less than in the PP equilibrium depends on the parameters. We have:

Proposition 3 *In the Bad News equilibrium that exists for all (ρ, α) with $\rho < \rho_2(\alpha)$, the weak type is (from an interim perspective) strictly worse off than alone, and the strong type strictly better off.*

		$\rho < \rho_1$	$\rho_1 < \rho < \rho_2$	$\rho > \rho_2$
Interim	**Strong type**	Better off	Better off / Better off	Better off
	Weak type	Worse off	Better off / Worse off	Better off
Ex ante		Worse off iff $\rho < \hat{\rho}$	Better off / Worse off iff $\rho < \hat{\rho}$	Better off

Figure 6: The value of joining a group. In the middle column there are three equilibria; the first entry refers to the best equilibrium for the type under consideration, the second to the worst one.

In contrast to the Good News equilibrium, group membership now has opposite effects on the interim utility of the two types, so its net ex-ante value is a priori ambiguous. Intuition suggests, however, that joining should be beneficial when (and only when) agents' level of self-confidence ρ is sufficiently high. This is essentially correct, except that ρ matters not *per se,* but mostly in relation to ρ^*, the level required to attempt the willpower activity next period. In the (most interesting) case where ρ^* is neither too close to 0 nor to 1, there is indeed a well-defined self-esteem cutoff for joining a group.

Proposition 4 *Assume that agents expect a Bad News equilibrium. There exist two values* $0 < \underline{\rho}^* < \overline{\rho}^* < 1$ *such that for all* $\rho^* \in \left]\underline{\rho}^*, \overline{\rho}^*\right[$, *agents prefer joining a group to staying alone if and only if their self-confidence* ρ *exceeds a cutoff* $\hat{\rho} \in (0, \rho^*)$, *which increases with* ρ^*.[24]

III.3.3 The value of joining a group: summary and some empirical evidence

Figure 6 summarizes the results from the previous sections. When $\rho > \rho_2(\alpha)$ there is a unique equilibrium; it is of the Good News type, and is Pareto superior to the outcome achievable by staying alone. In other words, the agent is better off not just *ex ante* but also at the *interim* stage. For $\rho_1(\alpha) \leq \rho \leq \rho_2(\alpha)$, however, such gains are not guaranteed since all three equilibria are possible. When $\rho < \rho_2(\alpha)$, finally, the unique equilibrium is the Bad News one, in which the strong type gains at the expense of the weak one.

[24] For $\rho^* < \underline{\rho}^*$(resp. $\rho^* > \underline{\rho}^*$), joining is always preferable to (resp., worse than) staying alone, independently of $\rho \in [0, \rho^*]$. Recall that ρ^* is given by (3) as a simple function of the model's parameters.

Two lessons can be drawn from our analysis:

1. Not all types benefit from joining a group. In particular, if $\rho \leq \rho_2(\alpha)$ there is always an equilibrium in which the weak and strong types have conflicting interests: the former would be better off alone, while the latter would gain from being in a group. Consequently, at the ex-ante stage, individuals will form (homogenous) groups only in a Good News equilibrium or, in a Bad News equilibrium, when self confidence is high enough.

2. There is, in some sense, a tradeoff between the potential benefits achievable through group membership, and the uncertainty over whether these gains will be reaped, or turn into losses, due to the multiplicity of possible equilibria. As the correlation α increases, so do the benefits of group membership in the best (Good News) equilibrium, but the set of parameters for which there is a multiplicity of possible outcomes also expands. Moreover, the welfare of the weak type under the worst (Bad news) equilibrium decreases with α.

As seen in the introduction, there is direct evidence of a positive correlation between individuals' level of self-esteem and their "clean time" in support groups. An example is the study of 200 members of Narcotics Anonymous illustrated in Figure 1.[25] De Soto et al. (1985) perform a similar exercise for a completely different group and in a different context (a sample of 312 members of Alcoholics Anonymous), and obtain similar results. The correlation thus seems to be a robust stylized fact. A plausible interpretation of this evidence is that group interactions help individuals sustain desirable behavior, which in turn raises their self-esteem. This would be in line with our results, since the building up and maintenance of self-confidence is the very focus of the model. The authors of these studies, however, do not test whether the observed correlation is due to beneficial effects of the programs or to self-selection: it could be that agents with lower self-esteem drop out because they do not find group membership useful (in expected terms).[26] This

[25] Analogous results are obtained with "anxiety" instead of "self-esteem". See Christo and Sutton (1994) and Christo (1999). Note that the time on the horizontal axis is the number of months of abstinence, as opposed to simple tenure in the group. There is thus no contradiction between the upward trend and the requirement, common to all Bayesian learning models, that individual beliefs should follow a martingale (hence also average beliefs, absent selective attrition of the type discussed below).

[26] The large impact of the first 3 months apparent on Figure 1 gives credence to the self-selection hypothesis, at least at the beginning, since it is widely believed that such programs need longer periods of time to produce beneficial effects. The gap between the indices of self-confidence of the average member of Narcotics Anonymous and of the average member of the reference group (60 non-addict students of a London Polytechnic) is reduced by nearly 30 per cent between the first and second observations. Later on, however, selective attrition is unlikely to account for most of the upward trend, as De Soto et al. (1994) cite a number of follow-up (longitudinal) studies of alcoholics that do show individual improvements in self-esteem during abstinence time.

interpretation of the data (which does not conflict with the first one) is also in line with our theoretical predictions: as seen above, agents with relatively high self confidence always find group membership beneficial, while it is only those with low self-confidence who may find peer interactions detrimental, and prefer isolation.

IV Heterogeneous peer groups

We have until now focussed attention on the symmetric equilibria that arise in a homogenous group. We now consider the more general case where peers may differ in their preferences, willpower, or incentives to exercise self-restraint. Such heterogeneity leads to asymmetric equilibria, which we fully characterize. Conversely, we show that asymmetric equilibria cannot arise in a homogenous group. This extended analysis allows us to answer two important questions about the nature of peer interactions. The first is whether an individual can free-ride on others' behavior, increasing his self-control at their expense.[27] The second and central issue is the impact on each individual's behavior and welfare of the group's or "club's" composition. For instance, when an agent's self-control problem becomes less severe –due to better time-consistency, external incentives, or lower temptation payoffs– does this help or hurt his peers? Would anyone accept a partner whom they perceive to be weaker than themselves?

IV.1 Equilibrium behavior

We consider a more general, possibly asymmetric correlation structure between the two agents' costs, represented by a joint distribution $F(c^1, c^2)$ over $\{c_H, c_L\} \times \{c_H, c_L\}$. Individuals' unconditional expectations or initial self-confidence levels will still be denoted as $\rho^i \equiv \Pr_F(c^i = c_L)$, and the conditional probabilities as $\pi^i_{LL} \equiv \Pr_F(c^i = c_L \,|\, c^j = c_L)$ and $\pi^i_{HH} \equiv \Pr_F(c^i = c_H \,|\, c^j = c_H)$, for $i = 1, 2$.[28] We only impose a general condition of positive correlation between agents' craving costs (monotone likelihood ratio property):

$$\frac{\Pr_F\left((c_H, c_L)\right)}{\Pr_F\left((c_L, c_L)\right)} < \frac{\Pr_F\left((c_H, c_H)\right)}{\Pr_F\left((c_L, c_H)\right)}. \tag{17}$$

We also allow for differences in agents' preferences parameters such as a^i, b^i, B^i, β^i, etc.[29] As a result, their self-confidence thresholds for attempting the willpower activity in the

[27] In a symmetric equilibrium everyone uses the same strategy, and therefore benefits equally from the spillovers provided by the group. When different strategies are allowed, one may imagine a case in which agent 1's actions (say) are so informative that agent 2 can afford to be less informative.

[28] We shall similarly denote $\pi^i_{HL} \equiv 1 - \pi^i_{LL}$ and $\pi^i_{LH} \equiv 1 - \pi^i_{HH}$. Condition (17) below is then equivalent to $\pi^i_{LH}/\pi^i_{LL} < 1 < \pi^i_{HH}/\pi^i_{HL}$, for $i = 1, 2$.

[29] The two agents could also face different costs in each state, c^i_H and c^i_L, with joint distribution F over $c^1_H, c^1_L \times c^2_H, c^2_L$, as long as the conditions (1), (2) and $0 < \rho^*_i < 1$ are satisfied for both of them.

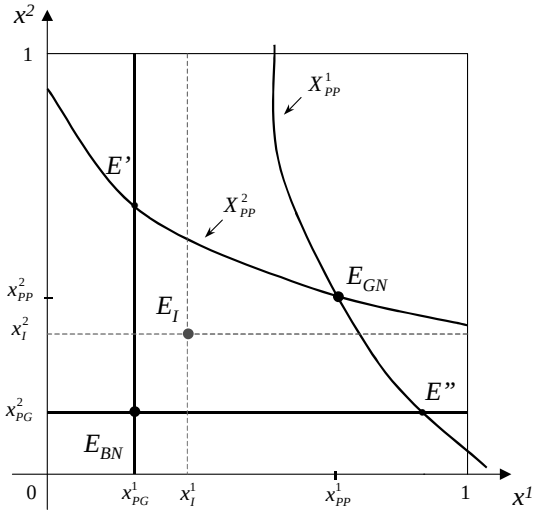

Figure 7: Good News, Bad News and Extreme News equilibria

second period, defined by (3), may be different. We shall denote them as ρ^{i*}, and focus on the interesting case where $\rho^i < \rho^{i*}$ for all i; one can think of $\rho^{i*} - \rho^i$ as the individual's "demand for self-confidence". Finally, the two individuals may now use different self-restraint strategies (probability of perseverance by a weak type), which we shall denote as x^1 and x^2. Although it is much more general than the symmetric case considered earlier, this game can be analyzed using the same key concepts and intuitions.

1. Informativeness constraints. In keeping with previous notation, let $\mu^i_{PP}(x^i, x^j)$ and $\mu^i_{PG}(x^i)$ denote individual i's posteriors about his own type when both agents persevered in the previous period, and when he persevered but the other agent did not.[30] The same simple reasoning as in Section III.2 (namely, the choice of W in the second period must be state-contingent) shows that, in any equilibrium, these beliefs must satisfy:

$$\mu^i_{PG}(x^j) \leq \rho^{i*} \leq \mu^i_{PP}(x^i, x^j). \tag{18}$$

As shown in the appendix and illustrated on Figure 7, each equation $\mu^i_{PP}(x^i, x^j) = \rho^{i*}$ uniquely defines a downward-sloping function $x^i = X^i_{PP}(x^j)$, with $(X^1_{PP})^{-1}$ steeper than X^2_{PP}. As long as the two agents are not excessively different from one another, there is then a unique intersection $E_{GN} = (x^1_{PP}, x^2_{PP}) \in [0,1] \times [0,1]$, where both (weak) agents play their "good news" strategies.[31] Similarly, each equation $\mu^i_{PG}(x^i) = \rho^{i*}$ has a unique

[30] Clearly, $\mu^i_{PG}(x^i)$ is independent of x^j : once agent j has given in, his type is completely revealed. The functions μ^i_{PP} and μ^i_{PG} depend of course on the joint distribution F, as do the profit functions Π^i defined below. For notational simplicity we shall leave this dependence implicit.

[31] For simplicity, we shall focus on this case from here on. The case where any of the intersections

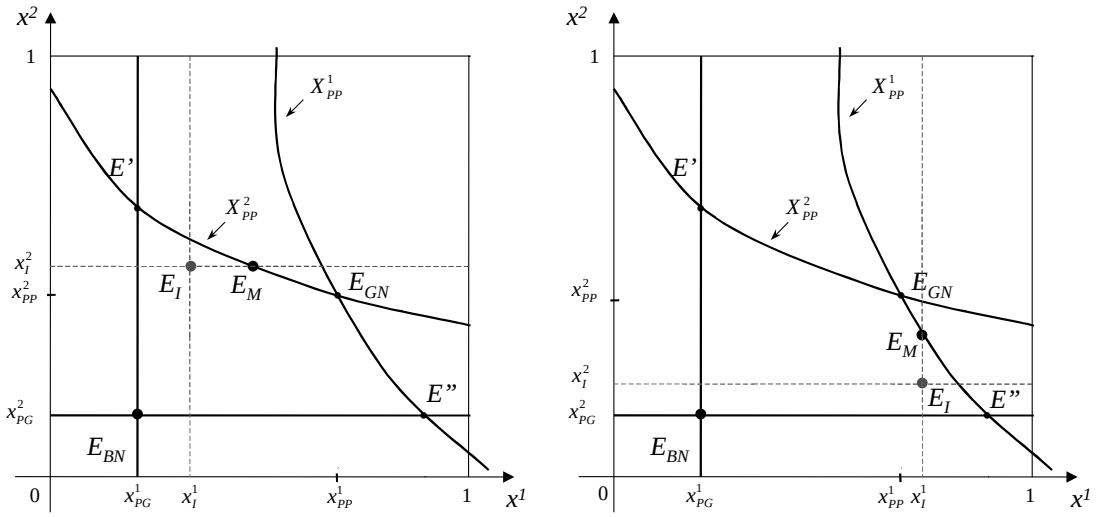

Figure 8: Mixed, Extreme News, and Bad News equilibria

solution $x^i = x^i_{PG}$, which corresponds on Figure 7 to a straight horizontal or vertical line. At the intersection $E_{BN} = (x^1_{PG}, x^2_{PG})$, both (weak) agents play their "bad news" strategies. Quite intuitively, each of these lines lies closer to the origin than the corresponding X^i_{PP} curve, so that together the four constraints in (18) define a "permissible" region $E_{BN}E'E_{GN}E''$ within which any equilibrium must lie:

$$x^i_{PG} \leq x^i \leq X^i_{PP}(x^j). \tag{19}$$

2. Profitability constraints. Let $\Pi^i(x^j, y^i_{PP}, y^i_{PG})$ denote the net expected gains to a weak agent i if he chooses P rather than G, given that the other (weak) agent uses strategy x^j and that agent i's own second-period self will choose the W activity with probabilities y^i_{PP} and y^i_{PG} following the events PP and PG respectively. Let $\pi^i(x^j) \equiv \Pi^i(x^j, 1, 0)$, and denote as x^j_I the solution to the linear equation $\pi^i(x^j) = 0$.[32]

Clearly, in any equilibrium it must be that $\Pi^i(x^j, y^i_{PP}, y^i_{PG}) \geq 0$, with equality unless $x^i = 1$. Following a reasoning similar to that of Proposition 1, we can combine this condition with the second-period selves' optimal behavior to show that

$$\begin{cases} \text{if } \rho^{i*} < \mu^i_{PP}(x^i, x^j) \text{ then } \pi^i(x^j) \leq 0, \\ \text{if } \rho^{i*} > \mu^i_{PG}(x^i) \text{ then } \pi^i(x^j) \geq 0, \end{cases} \tag{20}$$

x^i_{PP} occurs outside the $[0,1] \times [0,1]$ box (implying a corner solution for i's equilibrium strategy) is easily analyzed using the techniques developed in this section, and it yields the same intuitions.

[32] We do not a priori constrain x^j_I to lie in $[0,1]$.

for $i = 1, 2$. Given our definitions, these conditions translate into:

$$\begin{cases} \text{if } x^j > x_I^j \text{ then } x^i = X_{PP}^i(x^j); \\ \text{if } x^j < x_I^j \text{ then } x^i = x_{PG}^i. \end{cases} \tag{21}$$

The two incentive-constraint loci $x^1 = x_I^1$ and $x^2 = x_I^2$ divide the (x^1, x^2) plane into four quadrants. By (21), we see that:

1) The only possible equilibrium inside the Northeast (respectively, Southwest, Northwest, or Southeast) quadrant is the point E_{GN} (respectively, E_{BN}, E', or E''), and it is indeed an equilibrium when it lies in the said quadrant.

2) The only possible equilibria along the quadrant boundaries are: i) $E_I = (x_I^1, x_I^2)$, when it lies inside the region $E_{BN}E'E_{GN}E''$; ii) the point $E_M \equiv \left(\left(X_{PP}^2\right)^{-1}(x_I^2), x_I^2\right)$ when it lies on the upper boundary of that region, as on the left panel of Figure 8; iii) the point $E_M \equiv \left(x_I^1, \left(X_{PP}^1\right)^{-1}(x_I^1)\right)$ when it lies on the right boundary of that same region, as on the right panel of Figure 8.

These simple conditions allow us to completely derive the set of equilibria, depending on the location of E_I in the (x^1, x^2) plane. We provide these general results in the appendix, and focus here on the case where all three possible types of equilibria coexist, so that we can analyze the comparative statics of each one. It is easily seen from (21) that a necessary and sufficient condition for multiplicity is that the point E_I lie in the permissible region of Figures 7–8, that is,

$$x_{PG}^i < x_I^i < X_{PP}^i(x_I^i), \text{ for } i = 1, 2. \tag{22}$$

Proposition 5 *When condition (22) holds, the equilibrium set S is determined as follows.*
i) If $x_{PG}^i < x_I^i < x_{PP}^i$, for $i = 1, 2$, then $S = \{E_{BN}, E_I, E_{GN}\}$.
ii) If $x_{PG}^1 < x_I^1 < x_{PP}^1$ but $x_I^2 > x_{PP}^j$ then $S = \left\{E_{BN}, E_I, E_M \equiv \left(\left(X_{PP}^2\right)^{-1}(x_I^2), x_I^2\right)\right\}$.
iii) If $x_{PG}^2 < x_I^2 < x_{PP}^1$ but $x_I^1 > x_{PP}^1$ then $S = \left\{E_{BN}, E_I, E_M \equiv \left(x_I^1, \left(X_{PP}^1\right)^{-1}(x_I^1)\right)\right\}$.
When (22) does not hold there is a unique equilibrium, as specified in the appendix.

Thus, under condition (22) there is an equilibrium where both agents are in a "bad news" regime, another one where both are in an "extreme news" regime, and a third one where at least one of them is in a "good news" regime. In the last case the other agent plays either a "good news" strategy (we can then unambiguously refer to the equilibrium as a Good News equilibrium) or else an "extreme news" strategy (we refer to this as a Mixed equilibrium, hence the M subscript). Such a Mixed equilibrium occurs when E_I is located inside the permissible region, but either higher than or to the right of E_{GN}; see Figure 8. In such a situation the informativeness constraint $\pi^j = 0$ is binding on one agent

and the incentive constraint $\mu^i_{PP}(x^i, x^j) = \rho^{i*}$ on the other, so that the equilibrium lies at their intersection. Intuitively, this corresponds to a situation where agent i's self-control problem is significantly worse than agent j's.

Conversely, note that in a symmetric game the two agents' incentive constraints are symmetric, so their intersection E_I must lie on the diagonal. The same is true for the informativeness constraints in each state and their respective intersections E_{GN} and E_{BN}.

Consequently, we have:

Corollary 1 *In a homogeneous peer group (ex-ante identical agents), there can be no asymmetric equilibria.*

This result is interesting because it makes clear that when agents are ex-ante identical, none of them can free ride on the other, i.e. engage in more pooling with strong types (choose a higher x_1, which is beneficial ex ante) with the expectation that the other agent will make up for the reduced informativeness of the joint outcome by adopting a more separating strategy (a low x_2). In a symmetric group, therefore, no one ever gains (ex ante) at the expense of others. This will no longer be the case when the agents are heterogeneous.

IV.2 Comparative statics and welfare analysis

We now examine how a change in the severity of the self-control problem of one individual affects the behavior and welfare of his peers. Note that since the type and actions of agent i do not directly enter the payoff of agent j, a change in i's parameters can affect j *only through the informational content* of the jointly observed behavior.

One might think that having a partner who finds it easier (or faces better incentives) to exert self-restraint is always beneficial. The insights already obtained from our model suggest that this need not be true. A person who never gives in to temptation, either because he is never really tempted (strong type), or is able to exercise nearly perfect self control (x close to 1, due for instance to a high self-reputational stake), provides no informational spillover at all to his partners. Being with someone who is "too perfect," or always acts that way, is thus no better than being alone, and therefore less desirable than being matched to someone with more imperfect self-control. Of course, one would also expect that an excessively weak partner will be undesirable, as he is likely to generate only bad news. In line with these intuitions, we shall demonstrate that *individuals value the "quality" of their peers non-monotonically.*

The fact that the only externalities in the model are informational implies that, from the point of view of agent 2, a sufficient statistic for all the preference parameters of agent 1

is his self-reputation threshold ρ^{1*}, defined by (3). A lower degree of willpower β^1, a lower long-run payoff from perseverance B^1, or a higher payoff from the no-willpower option a^1 all translate into a higher self-confidence "hurdle" ρ^{1*} that agent 1 must achieve if he is to choose W in the second period. Together with the joint cost distribution $F(c^1, c^2)$, this is all that agent 2 needs to know about his peer. In our analysis we can therefore simply examine the effects on agent 2 of variations in ρ^{1*}, without having to specify their ultimate source.[33]

Rather than examine the local comparative statics of each equilibrium separately, we shall integrate them into a more interesting *global* analysis, allowing us in particular to ask what type of partner is (ex-ante) optimal. Specifically, we gradually raise ρ^{1*} from 0 to 1, and track the equilibrium with the highest level of self-control as it evolves from the Good News type to the Mixed type that is its natural extension, and finally to the Bad News type.[34] The key results are illustrated on the right panel of Figure 9.

Proposition 6 *In a heterogenous peer group where the equilibrium with the most self-control is always selected:*

i) Each agent's ex-ante welfare W^i is hump-shaped with respect to the severity of his partner's potential self-control problem, as measured by ρ^{j*}.

ii) The partner who maximizes agent i's welfare is one who is expected to be a little weaker than him, that is, who has a ρ^{j*} somewhat above ρ^{j*}.

iii) Group membership is strictly preferable to isolation only if the partner is neither too strong nor too weak compared to oneself (ρ^{j*} belongs to an interval that contains ρ^{i*}).

These results reflect a very intuitive tradeoff between the *likelihood* that the peer's behavior will be a source of encouraging or discouraging news, and the *informativeness* of his perseverance or giving up. The first effect tends to make a stronger partner preferable, since he is more likely to behave well, and thus be a source of good news. The second effect favors having a weaker partner, since low expectations make his successes more meaningful, and his failures less so. Figure 9 shows that for relatively low values of

[33] One might think about also varying agent 1's initial self-confidence (and reputation) ρ^1, but this turns out not to be a very meaningful exercise. Indeed, ρ^1 cannot be varied without also altering either agent 2's own self-confidence ρ^2, or the entire correlation structure between the agents: by Bayes' rule, $\rho^2 = \rho^1 \pi^2_{LL} + (1-\rho^1)(1-\pi^2_{HH})$. For instance, if it is common knowledge that both agents are always of the same type ($\pi^i_{HH} = \pi^i_{LL} = 1$), then $\rho^1 \equiv \rho^2$. Conversely, for ρ^2 to remains unaffected, the conditional probabilities π^2_{LL} and π^2_{HH} must decrease in just the right way. Intuitively, if an agent's view of his peer changes he must also revise his own self-view, or the extent to which their preferences are correlated.

[34] The comparative statics of the Extreme and Bad News equilibria are also obtained in the process. It is important to note that while we focus here (for completeness) on the case where all three equilibria coexist, all our results (Proposition 6 below) *apply unchanged* when there is a *unique* equilibrium that is of the Good News or Mixed type. (See the proof of Proposition 5 and Figure 10 in the Appendix for the parameter configurations that lead to uniqueness).

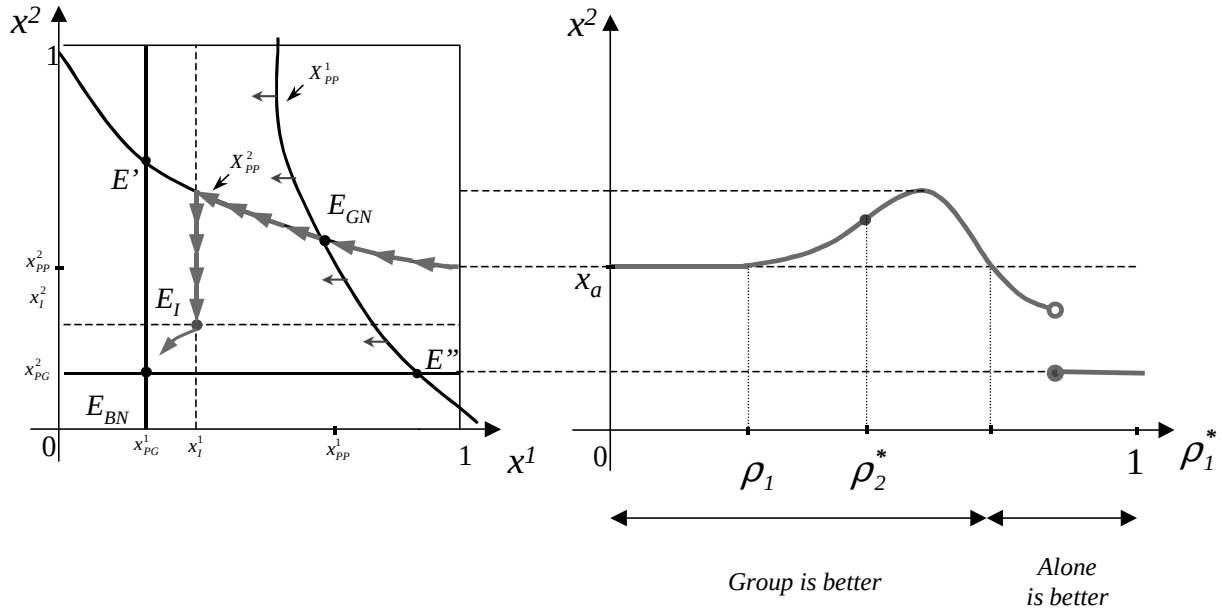

Figure 9: The effect on agent 2 of the severity of his peer's potential self-control problem. The right panel depicts both agent 2's behavior x^2 (when weak) and his ex ante welfare $W^2 = \rho^2 W^{2,s} + (1-\rho^2)W^{2,w}$. Indeed, $W^{2,w}$ strictly increases with x^2, while $W^{2,s}$ is always nondecreasing in ρ^{1*}.

ρ^{1*} informativeness is the main concern (so x^2 and W^2 increase with ρ^{1*}), whereas at higher values it is the likelihood effect that dominates (so x^2 and W^2 decline). The first case obtains as long as the Good News equilibrium can be sustained. The second case corresponds first to the Mixed equilibrium (where only agent 1 plays the good news strategy), and then to the Bad News equilibrium that necessarily prevails when one of the peers is too weak.

We now derive Proposition 6 by means of a simple graphical analysis: as ρ^1 increases from 0 to 1, the high self-restraint equilibrium set is shown to travel along the path marked by arrows on the left panel of Figure 9. The implied self-control behavior (and welfare) of agent 2 can then simply be read off the right panel of the figure. Since these graphs convey the essence of the results in Proposition 6, and the main underlying intuitions were explained above, readers who want to skip the formal analysis may move directly to the next section.

1. Good News equilibrium. Recall that when $\rho^{*1} < \rho^1$ agent 1 can always achieve complete self control on his own ($x^1 = 1$), in which case agent 2 learns nothing from observing his peer's behavior; hence $x^2 = x_a^2$, as when there is no group. For a relatively low

value of $\rho^{1*} \geq \rho^1$, we are in a configuration like that of Figure 7, with E_{GN} constituting a Good News equilibrium, located at the intersection of the two informativeness constraints $\mu^1_{PP}(x^1, x^2) = \rho^{1*}$ and $\mu^2_{PP}(x^1, x^2) = \rho^{2*}$. An increase in ρ^{1*} causes the locus X^1_{PP} to shift left, meaning that agent 1 becomes less likely to exert self-control. Indeed, in order to close the larger "self-confidence gap" $\rho^{1*} - \rho^1$ that he now faces, his perseverance must be a more credible signal of being a strong type; this requires less pooling by the weak type. Agent 2's informativeness constraint X^2_{PP}, by contrast, is unchanged. As a result, the equilibrium E_{GN} travels left and up along the X^2_{PP} locus: x^1 decreases, but x^2 increases. Thus agent 2 is actually *better off* from a (marginal) *worsening* in the severity of his peer's self-control problem.[35] The intuition for this perhaps surprising result is simple: as the extent to which agent 1's perseverance (when it occurs) is really good news increases, agent 2 (when weak) is able to engage in more pooling, which means persevering more often. In the tradeoff mentioned earlier between the likelihood and informativeness of a peer's perseverance, the latter concern dominates.

2. Mixed equilibrium. As ρ^{1*} keeps rising, agent 1 becomes less and less likely to exert self-restraint (x^1_{PP} continues to decline along the path shown on Figure 9), and we eventually reach a point where agent 2 becomes more concerned about the low likelihood of receiving good news (or high likelihood of receiving bad news) from his peer, than about their informativeness. This occurs on Figure 9 at the point where E_{GN}, in its leftward movement, encounters the vertical $x^1 = x^1_I$ locus.[36] By Proposition 5, E_{GN} then ceases to be an equilibrium, and is replaced by $E_M \equiv (x^1_I,\ (X^1_{PP})^{-1}(x^1_I))$. Further increases in ρ^{1*} cause E_M to move down along the x^1_I locus, so that $x^2 = x^2_M$ now declines, as shown on Figure 9. Thus, a weak agent 2 now *loses* self-discipline and welfare from interacting with a "worse" peer. A strong agent 2 is unaffected, since x^1 remains unchanged at x^1_I.

Putting this case together with the previous one, the fact that self-control and welfare are maximized by a match with a somewhat weaker partner (so that the peak on Figure 9 occurs to the right of ρ^{2*}) is easily seen by recalling that, in a symmetric situation, E_{GN} is an equilibrium, whereas E_M is not.

3. Bad News equilibrium. As agent 1's (potential) self-control problem becomes still more severe (ρ^{1*} continues to rise), there comes a point where the likelihood that he will

[35] This is obvious for a weak agent 2, since he gains self-restraint. The fact that the strong type is also better off is proved in the Appendix.

[36] Recall that the x^1_I locus is defined by agent 2's incentive constraints $\pi^2\left(x^1\right) = 0$, which is independent of ρ^{1*} or any other of the parameters characterizing agent 1. By contrast, as we make agent 1's self-control problem more difficult (say, decreasing B^1 increasing a^1, etc., causing ρ^{1*} to increase), the level of self-control x^2_I by agent 2 required for the indifference condition $\pi^1\left(x^2\right) = 0$ to hold rises. Thus the x^2_I locus shifts up with ρ^{1*}, and so does the point $E_M = (x^1_I, x^2_I)$.

be a source of bad news is so high that positive group externalities can no longer be sustained, and only the Bad News equilibrium survives. This occurs on Figure 9 when the Southward-moving point E_M falls below the Extreme News point E_I, which is moving North. The relevant equations from there on are $\mu^1_{PG}(x_1) = \rho^{1*}$ and $\mu^2_{GP}(x_2) = \rho^{2*}$, which correspond on Figure 9 to the lines $x^1 = x^1_{PG}$ and $x^2 = x^2_{PG}$. As ρ^{1*} continues to rise x^1_{PG} shifts left (for the same reason as the X^1_{PP} schedule did), but x^2_{PG} is unchanged. As a result, x^1 decreases, but x^2 remains unaffected. There is thus no impact on agent 2's behavior, and it is easy to see that there is no impact on his welfare either.[37]

V Conclusion

The starting point of this paper was the observation that informational spillovers are an important part of peer interactions, particularly when individuals face self-control problems. To analyze these interactions and their welfare implications, we proposed a model that combines that combines imperfect willpower, self-signaling and social learning.

Observing how others deal with impulses and temptation can be beneficial or detrimental, since these news can improve or damage the agent's self-confidence in his own prospects. One might therefore have expected that, even when learning from peers is beneficial ex ante, at the interim stage some type of agent would lose and another gain from such interactions. We showed, however, that under appropriate conditions –the main one being that everyone have some minimum level of self-confidence– *all* types can benefit from joining a group. Among individuals with really poor self-confidence, by contrast, social interactions will only aggravate the immediate-gratification problem, and lower ex-ante welfare. Furthermore, we showed that peer influences in self-control can easily give rise to multiple equilibria, even when agents' payoffs are completely independent. There is in fact often a trade-off between the potential benefits from joining a group and the underlying uncertainty about its equilibrium outcome. A higher degree of correlation between agents' types improves welfare in the best group equilibrium but lowers it in the worse one, while also widening the range of initial self-confidence levels where multiplicity occur.

We also examined the effects of heterogeneity among peers, and showed that individuals generally value the "quality" of their peers non-monotonically –in contrast to most models where social payoffs are exogenously specified. Intuitively, a person who is too weak is most likely to exhibit demoralizing behavior, while one who is too strong is one from whose likely successes there is little to be learned. Thus, there will be gains to group formation only among individuals who are not too different from one another in terms

[37] This is formally show in the Appendix, for both types.

of preferences, willpower, and external commitments. We showed furthermore that the (ex-ante) "ideal" partner is someone who is perceived to be a little weaker than oneself –reflecting the idea that "if *he* can do it, then surely I can".

Our model thus sheds light on several important aspects of the social dimension of self-control, and its premises and predictions are consistent with the available evidence from the clinical psychology literature. Nonetheless, it is still clearly oversimplistic, and could be extended in several directions. For instance, with longer horizons, what an individual learned about a peer would affect the desirability of continuing that particular relationship, leading to rich sorting dynamics through matches and quits.

A particularly interesting direction for further research would be to explore peer effects that involve *excessive*, rather than insufficient, self-regulation.[38] The social aspects of compulsive behavior seem particularly relevant with respect to work effort, and could provide a self-reputational theory of the "rat race". Finally, extending our framework to richer organizational settings should lead to a better understanding of team or employee morale.

[38] See Bodner and Prelec (1997) and Benabou and Tirole (2000) for accounts of rigid behavior and compulsive personal rules in a single-agent setting.

VI Appendix

In the proofs of Propositions 1 and 5 , and in the discussion in the text, we use certain properties of the solutions to the systems of equations $\mu^i_{PP}(x^1, x^2) = \rho^{i*}$ and $\mu^i_{PG}(x^j) = \rho^{i*}$, for $i = 1, 2$. The following Lemma establishes these properties:

Lemma 1 *For $i, j = 1, 2$ with $i \neq j$:*

i) The loci $X^i_{PP}(x^i, x^j)$ are decreasing in x^j. Furthermore $X^2_{PP}(x^1)$ cuts $(X^1_{PP})^{-1}(x^1)$ from below at most one intersection in the positive orthant.

ii) If $\rho^i < \rho^{i*}$ and the two agents are not excessively different from one another, then there is a unique interior solution for each system of equations: namely, $(x^1_{PP}, x^2_{PP}) \in [0,1] \times [0,1]$ and $(x^1_{PG}, x^2_{GP}) \in [0,1] \times [0,1]$.

Proof. i) We first verify that $X^i_{PP}(x^i, x^j)$ is decreasing in x^j. By Bayes' rule,

$$\frac{\mu^i_{PP}(x^i, x^j)}{1 - \mu^i_{PP}(x^i, x^j)} = \frac{\Pr(c^i = c_L,\, c^j = c_L) + \Pr(c^i = c_L,\, c^j = c_H)\, x^j}{\Pr(c^i = c_H,\, c^j = c_L)\, x^i + \Pr(c^i = c_H,\, c^j = c_H)\, x^i x^j}, \tag{23}$$

$$\frac{\mu^i_{PG}(x^j)}{1 - \mu^i_{PG}(x^j)} = \frac{\Pr(c^i = c_L,\, c^j = c_H)}{\Pr(c^i = c_H,\, c^j = c_H)\, x^i}. \tag{24}$$

Clearly, μ^i_{PP} and μ^i_{PG} are both decreasing in x^i. To see that μ^i_{PP} is decreasing in x^j as well, note that $\partial \mu^i_{PP}(x^i, x^j)/\partial x^j$ has the same sign as the determinant $\Pr((c_L,\, c_H))\Pr((c_H, c_L)) - \Pr((c_L, c_L))\Pr((c_H, c_H))$, which is negative by the monotone likelihood condition (17). Therefore $\partial X^i_{PP}(x^i, x^j)/\partial x^j < 0$ by the implicit function theorem. Next, to see that $X^2_{PP}(x^1)$ cuts $(X^1_{PP})^{-1}(x^1)$ from below at most a unique intersection in the positive orthant, note that $X^2_{PP}(0)$ is bounded for $x^1 \in [0,1]$. By contrast, we can easily verify that $\lim_{x^1 \to 0} (X^1_{PP})^{-1}(x^1) = +\infty$. Therefore there exists a point x^1 small enough such that $X^2_{PP}(x^1) < (X^1_{PP})^{-1}(x^1)$. To complete the argument, we now show that these two loci cross at most once in the positive orthant: so if they do intersect, it must be with $X^2_{PP}(x^1)$ crossing $(X^1_{PP})^{-1}(x^1)$ from below. Note first that any intersection must by such that $(\mu^i_{PP}(x^1, x^2)) / (\mu^2_{PP}(x^1, x^2)) = (\rho^{1*}) / (\rho^{2*})$. By (23), this implies

$$x^2 = \left[\frac{\rho^{1*} \Pr_F[(c_H, c_L)]}{\rho^{2*} \Pr_F[(c_L, c_H)]}\right] x^1 + \left[\frac{\rho^{1*}}{\rho^{2*}} - 1\right] \left[\frac{\Pr_F((c_L, c_L))}{\Pr_F((c_L, c_H))}\right].$$

This defines an upward-sloping line in the (x^1, x^2) plane, which can have at most one intersection with the decreasing curve $X^2_{PP}(x^1)$.

ii) It is straightforward to verify that if the agents are symmetric and $\rho^i < \rho^*$, then the solutions are interior in $(0,1)$. By continuity, if asymmetries are small enough, the solutions must be in $[0,1] \times [0,1]$ for both systems of equations. ¥

Proof of Proposition 1 It is easy to verify that, for any $\alpha \in (0,1)$, the two equations in ρ, $x_{PP}(\rho;\alpha) = x_I(\rho;\alpha)$ and $x_{PG}(\rho;\alpha) = x_I(\rho;\alpha)$ have a unique solution in, respectively, $(0,\rho^*)$ and $(0,\frac{\rho^*}{1-\alpha})$. We denote them as $\rho_1(\alpha)$ and $\rho_2(\alpha)$ respectively. Since $x_I(\rho;\alpha)$ is decreasing in ρ while $x_{PP}(\rho;\alpha)$ and $x_{PG}(\rho;\alpha)$ are increasing, $x_I(\rho;\alpha)$ crosses the other two loci from above. It follows that for $\rho < \rho_1(\alpha)$, $\Pi(x,1,0;\rho,\alpha) < 0$ for any $x \leq x_{PP}(\rho;\alpha)$, so one cannot have a Good News equilibrium. For $\rho \geq \rho_1(\alpha)$, $\Pi(x_{PP}(\rho;\alpha),1,0;\rho,\alpha) \geq 0 > \Pi(1,0,0;\rho,\alpha)$ so, by continuity, there is always a unique $y_{PP} \in (0,1)$ such that $\Pi(x_{PP}(\rho;\alpha),y_{PP},0;\rho,\alpha) = 0$. Clearly $x_{PP}(\rho;\alpha)$ and y_{PP} then define an equilibrium, since these values respectively make the weak type at the *interim* stage and the second-period Self willing to mix. A similar argument shows that a Bad News equilibrium exists if and only if $\rho \leq \rho_2(\alpha)$. To see that for $\rho_1(\alpha) \leq \rho \leq \rho_2(\alpha)$ we also have an Extreme News equilibrium, note that in this range $x_I(\rho;\alpha) \in [x_{PP}(\rho;\alpha), x_{PG}(\rho;\alpha)]$ and $\Pi(x_I(\rho;\alpha),1,0;\rho,\alpha) = 0$, so the weak type is willing to mix at the interim stage given the optimal reaction of the second period Self. Finally, since as $\alpha \downarrow 0$ we have $x_{PP}(\rho;\alpha) \to x_{PG}(\rho;\alpha)$, it is immediate to see that $\lim_{\alpha\to 0}|\rho_2(\alpha) - \rho_1(\alpha)| = 0$. ¥

Proof of Proposition 3 A Bad News equilibrium is, ex-ante, is strictly better then staying alone if and only if:

$$E(W_{PG} - W_a \,|\, \rho) \equiv \rho(W^s_{PG} - W^s_a) + (1-\rho)(W^w_{PG} - W^w_a) > 0. \tag{25}$$

From the informativeness constraint (23) we have $x_{PG} = (1-\alpha)(\rho/\rho^* - \rho)/(1-\rho+\alpha\rho)$; in the limiting case where the agent is alone ($\alpha = 0$) this becomes $x_a = (\rho/\rho^* - \rho)/(1-\rho)$. Substituting into conditions (15) and (16), we can then rewrite (25) as:

$$\Psi(\rho,\rho^*) \equiv (\rho^* - 1)k(\rho) + \rho^* - (1-\alpha)\rho < 0, \quad \text{where} \tag{26}$$

$$k(\rho) \equiv \frac{(1-\beta)\,c_H}{\beta\delta\,(1-y_{PG}(\rho))\,(B - c_L - a)}. \tag{27}$$

The function Ψ is increasing in ρ^* and decreasing in ρ. The first claim is obvious, and the second follows from the fact that $y_{PG}(\rho)$ is itself decreasing in ρ. Indeed, $y_{PG}(\rho)$ is defined as the solution y' to $\Pi(x_{PG}(\rho),1,y';\rho,\alpha) = 0$, or

$$B - b - \frac{c_H}{\beta} + \delta\left[(1-\alpha)\rho + (1-(1-\alpha)\rho)\,(x_{PG}(\rho) + (1-x_{PG}(\rho))y\prime)\right](b-a) = 0,$$

and $x_{PG}(\rho)$ is an increasing function into $[0,1]$. The monotonicity properties of Ψ imply that for each ρ^* there exists a unique $\hat{\rho}(\rho^*) \in [0,\rho^*]$ such that (25) holds if and only if $\rho > \hat{\rho}(\rho^*)$; furthermore, $\hat{\rho}(\rho^*)$ is non-decreasing in ρ^*. To study when this solution is interior, let us define $\underline{\rho}^*$ and $\overline{\rho}^*$ by the linear equations $\Psi(0,\underline{\rho}^*) = (\underline{\rho}^* - 1)k(0) + \underline{\rho}^* \equiv 0$

and $\Psi(1,\overline{\rho}^*) = (\overline{\rho}^* - 1)k(1) + \overline{\rho}^* - (1-\alpha)$ respectively. Then $0 < \underline{\rho}^* < \overline{\rho}^* < 1$, and for any ρ^* in $\left]\underline{\rho}^*, \overline{\rho}^*\right[$, $\hat{\rho}(\rho^*)$ lies in $(0,\rho^*)$ and is strictly increasing in ρ^*. For $\rho^* < \underline{\rho}^*$ we have $\hat{\rho}(\rho^*) = 0$, and $E\left(W_{PG} - W_a \,|\, \rho\right) > 0$ for all $\rho \geq 0$. Conversely, for $\rho^* > \overline{\rho}^*$ we have $\hat{\rho}(\rho^*) = \rho^*$, and $E\left(W_{PG} - W_a \,|\, \rho\right) < 0$ for all $\rho \leq \rho^*$. ¥

Welfare in an Extreme News equilibrium We analyze here welfare in the third type of equilibrium, where the second-period self follows a pure strategy. As one might expect, the results are intermediate between those of the Good News and Bad News cases.

As usual, we have for the weak type $W_I^w = W_a^w + (x_I - x_a)\ \frac{1-\beta}{\beta}\ c_H$. Recall from Figure 3 or 4 that $x_I(\rho;\alpha)$ declines from $x_{PP}(\rho;\alpha)$ to $x_{PG}(\rho;\alpha)$ as ρ spans the interval $[\rho_1(\alpha), \rho_2(\alpha)]$. Therefore we always have $W_{PG}^w < W_I^w < W_{PP}^w$, and there exists a threshold $\tilde{\rho}(\alpha)$ in the interval such that the weak type is better off than when alone if and only if $\rho \leq \tilde{\rho}(\alpha)$. As to the strong type, his welfare takes the same form as in the Bad News case, except that y_{PG} is replaced by 0 :

$$W_I^s = W_a^s + \delta\left[\Pr_I(P \,|\, s) - \Pr_I(P \,|w)\right](B - a - c_L) = W_a^s + \delta\alpha(1 - x_I)(B - a - c_L).$$

Since $x_I < x_{PP}$ he is better off compared not only to staying alone, but also compared to the Good News equilibrium. The comparison with his gains under the Bad News equilibrium, on the other hand, depends on the parameters. The Extreme News equilibrium is thus qualitatively similar, in terms of the value of joining a group, to a Good News equilibrium if $x_I > x_a$ (both types are better off at the interim stage), and to a Bad News equilibrium if $x_I > x_a$ (only the good type is better-off). ¥

Proof of Proposition 5 We proceed in three stages.

A. Proof of condition (18). If both posteriors were below ρ^{i*} Self 2 would never play W, therefore a weak agent i would always act myopically: $x^i = 0$. But then $\mu_{PP}^i(0, x^j) = 1$, a contradiction. Similarly, if both posteriors are above ρ^{i*}, a weak agent i will always play P, since this induces Self 2 to choose willpower with probability one. But $\mu_{PG}^i(1) = \rho^i < \rho^{i*}$, a contradiction.

B . Proof of condition (20).

1) Assume that $\pi^i(x^j) > 0$. We then cannot have $\mu_{PP}^i(x^i, x^j) > \rho^{i*}$, or else agent i's Self 2 will optimally choose $y_{PP}^i = 1$, leading to net profits of $\Pi^i\left(x^j, 1, y_{PG}^i\right) \geq \pi^i\left(x^j\right) > 0$ from choosing P rather than G in the first period. But then $x^i = 1$, so $\mu_{PP}^i(1, x^j) > \rho^{i*}$, or equivalently $x^j < X_{PP}^j(1) < 1$. Because $X_{PP}^j(x) - \left(X_{PP}^i\right)^{-1}(x)$ has the sign of $x_{PP}^i - x$ for all x (single-crossing property established by Lemma 1 and illustrated on Figure 7), this implies that $x^j < \left(X_{PP}^i\right)^{-1}(1)$, or equivalently $\mu_{PP}^j(x^j, 1) > \rho^{j*}$. As a result, agent

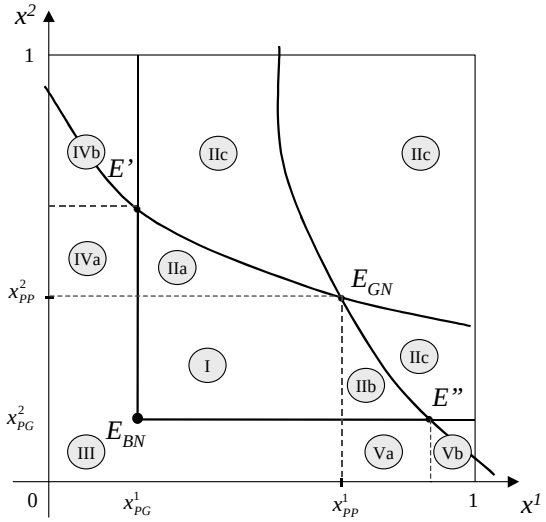

Figure 10: Equilibrium set in the general (asymmetric) model.

j's second-period self will choose $y^2_{PP} = 1$, ensuring $\Pi^j \left(1, 1, y^j_{PG}\right) = \Pi^j (1, 1, 0) > 0$. This leads to $x^j = 1$, a contradiction.

2) Assume now that $\pi^i(x^j) < 0$. We then cannot have $\mu^i_{PG}(x^i) < \rho^{i*}$, or else agent i's Self 2 will optimally choose $y^i_{PG} = 0$, leading to net profits of $\Pi^i \left(x^j, y^i_{PP}, 0\right) \leq \pi^i \left(x^j\right) < 0$ from choosing P rather than G in the first period. But then $x^i = 0$, so $\mu^i_{PG}(0) = 1 > \rho^{i*}$, a contradiction.

As shown in the text, Proposition 5 follows directly from the conjunction of these properties of the informativeness and incentive constraints.

C. Remaining Equilibria

When condition (22) hold, the intersection E_I of the two x^i_I loci lies inside the permissible region $E_{BN}E'E_{GN}E''$ of Figure 7 or 8. In Figure 10 this area is itself decomposed into areas I, IIa and IIb, which respectively correspond to cases (i), (ii) and (iii) or Proposition 5. When condition (22) does not hold, $E_I = (x^1_I,\, x^2_I)$ lies in one of the "outer areas" of Figure 10. Using (21) and the discussion which follows it in the text, it is easy to verify in each case that there is a unique equilibrium, located at a vertex or on one of the upper boundaries of the central, permissible region. Specifically, the equilibrium is, in counterclockwise order: $E\prime$ when E_I falls in IVb; $E_M = ((X^2_{PP})^{-1}(x^2_I),\, x^2_I)$ when E_I falls in IVa; E_{GN} when E_I falls in III; $E_M = (x^1_I,\, (X^1_{PP})^{-1}(x^1_I))$ when E_I falls in Va; E" when E_I falls in Vb; and E_{BN} when E_I falls in IIc. ¥

Proofs of the results in Section IV.2

Comparative statics of welfare in the Good News equilibrium. We saw that an increase in ρ^{1*} causes an decrease in x^1_{PP} and an increase in x^2_{PP}. The fact that a weak agent 2 is always better off from a (marginal) worsening of his partners' (potential) self-control problem then follows immediately from (13).[39] For a strong type, note that

$$y^2_{PP} = \frac{y^2_a}{\pi^1_{LH} + (1 - \pi^1_{LH})x^1_{PP}},$$

so the decrease in x^1_{PP} increases y^2_{PP}. Furthermore, the probability that agent 1 plays P given that agent 2 is strong is $\Pr^1_{PP}(P \mid s) = \pi^1_{LL} + (1 - \pi^1_{LL})x^1_{PP}$, whereas when agent 2 is weak it is $\Pr^1_{PP}(P \mid w) = \pi^1_{LH} + (1 - \pi^1_{LH})x^1_{PP}$. We can thus generalize (14) to:

$$W^{1,s}_{PP} = W^{1,s}_a + \delta y^1_a (B^1 - a^1 - c_L) \left(1 - x^2_{PP}\right) \left(\frac{\pi^2_{LL} - \pi^2_{LH}}{\pi^2_{LH} + (1 - \pi^2_{LH})x^2_{PP}}\right).$$

From this equation and (17), it is clear that we have an increase in $W^{2,s}_{PP}$. ¥

Comparative statics of welfare in the Bad News equilibrium. We saw that, in this case, a (marginal) increase in ρ^{1*} leaves x^2_{PG} unaffected. It immediately follows that there is no effect on the welfare of a weak agent 2. Let us now show that the same is true for a strong type. Using the indifference conditions of the weak type in a group and by himself ($\Pi^2(x^1, 1, y^2_{PG}) = 0 = \Pi^2(1, 1, y^2_a)$), we can write $y^2_{PG} = \left(y^2_a - \Pr^1_{PG}(P \mid w)\right) / \left(1 - \Pr^1_{PG}(P \mid w)\right)$. Substituting this into the expression for the welfare of the strong type, (16), and exploiting the fact that $1 - \Pr^1_{PG}(P \mid w) = (1 - x^1_{PG})(1 - \pi^1_{LH})$, we obtain

$$W^{2,s}_{PG} = W^{2,s}_a + \delta \left(1 - y^2_a\right)(B^2 - a^2 - c^2_L -) \left(\frac{\pi^1_{LL} - \pi^1_{LH}}{1 - \pi^1_{LH}}\right),$$

which is independent of any parameter of agent 1, as well as of his behavior x^1. ¥

Comparative statics of welfare in the Mixed and Extreme News equilibria. See the text and footnote 36, respectively.

[39] Equation (13) was written for the symmetric case, but directly extends to the asymmetric one if we add agent-specific superscripts $i = 1, 2$ to all functions and parameters. Similarly, the expressions below are immediate generalizations of those presented in Section III.

References

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

[2] Ainslie (2001) *Breakdown of Will,* in press. Cambridge University Press.

[3] Akerlof, G. (1991) "Procrastination and Obedience," *American Economic Review,* 81(2), 1–19.

[4] Banerjee, A. and Besley, T. (1990) "Peer Group Externalities and Learning Incentives: A Theory of Nerd Behavior," Princeton University Working Paper No. 68.

[5] Baumeister, R., Heatherton, T. and Tice, D. (1994) *Losing Control: How and Why People Fail at Self-Regulation.* Academic Press: San Diego, CA.

[6] Bem, D. (1972) "Self-Perception Theory," in L. Berkowitz ed., *Advances in Experimental Social Psychology.* New York, NY: Academic Press.

[7] Bénabou, R. (1993) "Workings of a City: Location, Education and Production," *Quarterly Journal of Economics,* 108, 619-652.

[8] Bénabou, R. and Tirole J. (2002) "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics,* 117(3), 871-915.

[9] Bénabou, R. and Tirole J. (2000), "Willpower and Personal Rules", Princeton University mimeo.

[10] Bernheim, D. (1994) "A Theory of Conformity," *Journal of Political Economy,* 102(5), 841-877.

[11] Beyth-Marom, R., Austin, L., Fischhoff, B., Palmgren, C., and Quadrel, M.J. (1993). "Perceived Consequences of Risky Behaviors: Adults and Adolescents." *Developmental Psychology*, 29, 549-563.

[12] Bodner, R. and D. Prelec (1997) "The Diagnostic Value of Actions in a Self-Signaling Model," MIT mimeo.

[13] Bodner, R. and Prelec, D. (2001) "A Neo-Calvinist Model of Conscience," in *Collected Essays in Psychology and Economics,* I. Brocas and J. Carrillo eds., Oxford University Press, forthcoming.

[14] Brocas, I. and Carrillo, J. (1999) "Entry Mistakes, Entrepreneurial Boldness and Optimism," CEPR D.P. No. 2213, August.

[15] Brocas, I. and Carrillo, J. (2001) "Rush and Procrastination under Hyperbolic Discounting and Interdependent Activities," *Journal of Risk and Uncertainty,* 22(2) 141-144.

[16] Brock, W. and Durlauf, S. (2001) "Discrete Choice with Social Interactions," *Review of Economic Studies,* 68(2), 235-260.

[17] Brueckner, J. and Lee, K. (1989) "Club Theory with a Peer-Group Effect," *Regional Science and Public Economics,* 19, 399-420.

[18] Carrillo, J., and T. Mariotti (2000) "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies,* 67(3), 529–544.

[19] Case, A. and Katz, L. (1991) The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youth," NBER Working Paper 3705.

[20] Christo, G. (1999) "Narcotics Anonymous as Aftercare," *The Center for Research on Drugs and Health Behaviour*, Imperial College London.

[21] Christo, G. and Sutton S., (1994) "Anxiety and Self-Esteem as a Function of Abstinence Time Among Recovering Addicts Attending Narcotics Anonymous", *British Journal of Clinical Psychology,* 33: 198-200.

[22] Coleman, J. (1988) "Social Capital in the Formation of Human Capital," *American Journal of Sociology,* 94, S95-S120.

[23] Crane, J. (1991) "The Epidemic Theory of Ghettos, and Neighborhood Effects on Dropping out and Teenage Childbearing," *American Journal of Sociology*, 96, 1226-1259.

[24] De Bartolome, C. (1990) "Equilibrium and Inefficiency in A Community Model with Peer Group Effects," *Journal of Political Economy,* 98, 10-133.

[25] De Soto, C. B., O'Donnell, W. E., Allred, L. J. and Lopes, C. E. (1985), "Symptomatology in Alcoholics at Various Stages of Abstinence, *Alcoholism: Clinical and Experimental Research*, 9: 505-512.

[26] Dynarski, M., Schwab, R., and Zampelli, E. (1989) "Local Characteristics and Public Production: The Case of Education," *Journal of Urban Economics*, 26, 250-263.

[27] Elster, J. (2001) "Introduction," in *Addiction: Entries and Exits,* J. Elster ed., New York: Russel Sage Foundation.

[28] Gilbert, D. and J. Cooper (1985) "Social Psychological Strategies of Self- Deception," in M. Martin, ed. *Self-Deception and Self-Understanding*, University Press of Kansas.

[29] Glaeser, E. and Scheinkman, J. (2000) "Non-Market Interactions," NBER W.P. No. 8053, December.

[30] Henderson, V., Mieszkosvsky, P. and Sauvageau, Y. (1978) "Peer Group Effects and Educational Production Functions," *Journal of Public Economics*, 10, 97-106.

[31] Hoxby, C. (2001) "Peer Effects in the Classroom: Learning from Gender and Race variation," NBER W.P. No. 7867.

[32] Kahneman D., Wakker P., Sarin R.(1997), "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics,* 112(2), 375-405.

[33] Kahneman D. (2000), "Expected Utility and Objective Happiness," in *Collected Essays in Psychology and Economics,* I. Brocas and J. Carrillo eds., Oxford University Press, forthcoming.

[34] Laibson, D. (1997) "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112: 443–478.

[35] Loewenstein, G. (1996) "Out of Control: Visceral Influences in Behavior," *Organizational Behavior and Human Decision Processes,* 65(3) March, 272–292.

[36] Loewenstein, G. and Schkade, D. (1999) "Wouldn't It Be Nice? Predicting Future Feelings," in D. Kahneman, E. Diener and N. Schwartz, eds. New York, NY: Russel Sage Foundation.

[37] Mullainathan, S. (1998) "A Memory Based Model of Bounded Rationality," mimeo, MIT.

[38] O'Donoghue, T. and Rabin, M. (1999) "Doing it Now or Later," *American Economic Review*, 89(1), 103–124.

[39] O'Donoghue T. and M. Rabin, (2000) "Risky Behavior Among Youths: Some Issues from Behavioral Economics." University of California, Berkeley, mimeo.

[40] Phelps, E. and Pollack, R. (1968) "On Second-Best National Savings and Game-Equilibrium Growth," *Review of Economic Studies, 35, 185-199.*

[41] Quattrone G., and Tversky, A. (1984) "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion," *Journal of Personality and Social Psychology,* 46, 2, 237–248.

[42] Sacerdote, B. (2001) "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics,* 116(2), 681–704.

[43] Scotchmer, S. (1994) "Public Goods and the Invisible Hand," in J. Quigley and E. Smolensky, *Modern Public Finance,* Harvard University Press, Cambridge, Massachusetts.

[44] Strotz, R. (1956) "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies,* 23, 165–180.