



Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Economic Theory 123 (2005) 105–134

JOURNAL OF
**Economic
Theory**

www.elsevier.com/locate/jet

Self-control in peer groups

Marco Battaglini^{a, b}, Roland Bénabou^{a, b, c, *}, Jean Tirole^{d, e, f}

^aDepartment of Economics, Princeton University, Princeton, NJ 08544 1013, USA

^bCentre for Economic Policy Research, 90-98 Goswell Road, London EC1V 7RR, UK

^cNational Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA

^dInstitut d'Economie Industrielle, Manufacture des Tabacs, 21 allées de Brienne, 31000 Toulouse, France

^eParis Sciences Economiques, 48 boulevard Jourdan, 75014 Paris, France

^fDepartment of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02138, USA

Received 6 April 2005

Abstract

Social influences on self-control underlie both self-help groups and many peer interactions among youths. To understand these phenomena, we analyze how observing each other's behavior affects individuals' ability to deal with their own impulses. These endogenous informational spillovers lead to either a unique "good news" equilibrium that ameliorates behavior, a unique "bad news equilibrium" that worsens it, or to the coexistence of both. A welfare analysis shows that people will find social interactions valuable only when they have enough confidence in their own and others' ability to resist temptation. The ideal partner, however, is someone with a slightly worse self-control problem than one's own: this makes his successes more encouraging, and his failures less discouraging.

© 2005 Elsevier Inc. All rights reserved.

JEL classification: C72; D82; D71; D91; J24

Keywords: Peer effects; Social interactions; Clubs; Self-control; Willpower; Addiction; Time-inconsistency; Memory; Psychology

* Corresponding author. Department of Economics, Princeton University, Princeton, NJ 08544 1013, USA.
Fax: +1 609 258 5533.

E-mail addresses: mbattagl@princeton.edu (M. Battaglini), rbenabou@princeton.edu (R. Bénabou), tirole@cict.fr (J. Tirole).

0022-0531/\$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2005.04.001

1. Introduction

The behavioral and economic implications of imperfect self-control by a single decision maker have been the focus of much recent work. Yet, people are typically immersed in social relations that exert powerful influences on their decisions. Peers and role models, for instance, play a critical part in young people's choices—particularly those that are subject to episodes of temptation like drinking, smoking, drug use, sexual activity, procrastination of effort, etc. In such settings peers may be good or bad “influences,” and the latter scenario is typically correlated with low or fragile self-esteem. At the same time, people with self-control or addiction problems often seek relief in self-help groups like Alcoholics Anonymous, Narcotics Anonymous and similar organizations that are predicated on the mutual sharing of experiences.

Psychologists and sociologists (not to mention parents) thus generally view the issues of self-control and peer effects as complementary. In economics, by contrast, they have so far been treated as largely separate areas of inquiry. In this paper we bring them together, studying how exposure to each other's behavior affects the ability of time-inconsistent individuals to deal with their own impulses.

Support groups, for instance, are an important social phenomenon. Organizations such as Alcoholics Anonymous, Narcotics Anonymous, Gamblers Anonymous, Debtors Anonymous and the like have branches in many countries, and millions of members. Economists are used to thinking about how entering contracts or binding implicit agreements with others allows agents to achieve desirable commitment. This, however, is not at all what self-help groups are about. Among the 14 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 clear that one cannot view these groups as standard commitment devices: they not only cannot, but do not even want to “control” their members. Their scope is in fact explicitly limited to fostering informational interaction (discussion) among members. Thus in “What does Alcoholics Anonymous do?” it is clearly stated that “A.A. members *share their experience* with anyone seeking help with a drinking problem” (emphasis added).

One therefore needs a theory to explain how (and when) observing the behavior of others can sometimes be beneficial for overcoming self-control problems, as with support groups, and sometimes highly detrimental, as often happens among schoolmates or neighborhood youths. Such a theory of peer effects in self-control should also be normative as well as

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/>, as can the other quotations given below.

positive. While group membership is sometimes exogenous (e.g., in public schools), it often involves of a voluntary choice, whether by the agent himself or by a “principal” invested with authority (judge ordering an addict to attend a 12-step program, parent trying to affect their child’s selection of peers).

In this paper, we take the first steps towards such a theory, by developing a 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.²

There are two fundamental assumptions in our model. First, agents have incomplete information about their ability to resist temptation and try to infer it from their 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 heavily emphasized in the psychology literature [1,2,6]. We build here on Bénabou and Tirole’s [10] formalization of these phenomena, which is based on the idea that imperfect self-knowledge gives rise to a concern for *self-reputation*. By breaking a personal rule (abstinence resolution, diet, exercise regimen, moral principle) an individual would reveal himself, in his own eyes, as weak-willed with respect to such temptations, and this reputational loss would further undermine his resolve in the future. The fear of creating precedents thus creates an incentive to maintain a clean “track record” in order to influence one’s future (selves’) morale and behavior 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 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 actions of people similar to oneself is a source of additional information about the manageability or severity of the self-control problem—or, equivalently, the

² For instance, Battaglini and Bénabou [4] study political activism by multiple interest groups or lobbies trying to influence a policymaker. While the framework differs from the present one in many key respects (no time inconsistency, imperfect recall, nor learning from peers), the techniques introduced here turn out to be applicable there as well.

³ Task-specific informational spillovers are also evident in Weightwatchers’ practice of weighing members each week and reporting to each not just his or her own loss or gain, but also the group average.

effectiveness of a particular method designed to alleviate it.⁴ The information may turn out to be good news, if the others are observed to persevere (stay “dry”, “clean,” remain in school, etc.), or bad news, if they are observed to cave in or have a relapse. When deciding whether to exercise costly self-restraint in the face of temptation, an individual will take into account the likelihood of each type of news, and how it would impact the reputational “return” on his own behavior. Therefore, 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, since it depends on their equilibrium strategies. As a result, our model, in which *peer effects are purely informational*, can give rise to amplification effects as well as multiple equilibria, where agents’ choices of self-restraint or self-indulgence are mutually reinforcing.

In the first part of the paper we focus on a symmetric situation where individuals are ex ante identical in all respects. Three main results are obtained. First, 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 membership improves self-discipline, a “bad news” equilibrium where it damages it, or to both. Second, social 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.

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 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.

⁴ Self-help groups may allow members to learn specific techniques (practical, mental or spiritual) for coping with impulses, but such “education” cannot be their sole or even main function. Techniques can be learned from a book or tape; or, if human contact is required, they are best transmitted and tailored to a person’s needs by an expert (doctor, counselor, therapist) rather than by non-chosen others who are themselves struggling, not always successfully, with their own weaknesses. Sharing experience with peers, on the other hand, is the best way to judge whether a given set of techniques can indeed work “for someone like me”. This broader interpretation, in which group membership gives access *both* to a potentially useful technique and to a pool of “experiments” where one can condition on a very fine set of variables (peers’ personal histories, etc.), is fully consistent with our model.

Whereas these typically imply monotone comparative statics, our analysis of learning-based spillovers reveals a general trade-off between the *likelihood* that someone else's behavior will be a source of encouraging or discouraging news, and the *informativeness* of this news.

The dynamics of self-confidence play a key role in our theory of peer effects. First, self-restraint by one member (e.g., abstinence) improves both his and others' self-confidence, and this in turn leads to more self-restraint by all in the future; misbehavior elicits the opposite feedbacks. Second, individuals will find self-help groups worth joining and remaining in only if they have sufficient confidence in their own and their peers' ability not to relapse. While there is no systematic literature on the subject, field studies of self-help groups consistently document correlation patterns that are in line with these results (but of course do not constitute formal tests). For instance, Christo and Sutton's [19] study of 200 Narcotics Anonymous members leads them 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.”

1.1. 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 influence of group characteristics on individual youths' behavior, whether in terms of academic achievement, school truancy, smoking, drinking and drug use, teen pregnancy, employment, criminal activity and the like [18,23,25,27,28,34].

Econometric studies are 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 they imply for the equilibrium and optimal composition of groups. Thus, De Bartolome [20] and Bénabou [8] study how peer or neighborhood effects shape the functioning of a city and its schools; Bernheim [11] examines how a concern for others' views of oneself leads to conformity; Brock and Durlauf [14] and Glaeser and Scheinkman [24] study how non-market interactions can lead to “social multipliers” and multiple equilibria. The only previous work seeking to endogenize peer effects is Banerjee and Besley's [3] model of student testing, where a benchmarking effect arising from the unknown difficulty of the test creates an informational complementarity between classmates' effort decisions.⁵

The other literature to which our paper relates is that on self-control problems, due for instance to non-exponential discounting (e.g., [29,32,35]). In particular, a recent line of research has shown how the combination of self-control and informational concerns can

⁵ That mechanism is specific to a particular setting and technology, however, and does not apply to most of the other behaviors discussed above. In particular, it has the feature that being with peers—even very bad ones—is always better than being alone.

account for many forms of “motivated cognitions” documented by psychologists. Carrillo and Mariotti [17] establish that time-inconsistent individuals may have, ex ante, a negative value for information. Bénabou and Tirole [9] develop a theory of rational self-deception through selective recall, and in [10] link personal rules to endogenous concerns for self-reputation. A related line of work by Bodner and Prelec [12] examines self-signaling in a split-self (ego-superego) model where the individual has “metapreferences” over his own tastes. Finally, our concern with interactions among time-inconsistent agents is shared with Brocas and Carrillo [13], who analyze how competition in the form of “patent races” can improve, and cooperation in joint projects worsen, individuals’ tendency to procrastinate. In our model, by contrast, *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

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 Appendix A.

2. The model

2.1. 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 [10]. The canonical example is that of an alcoholic who must decide every morning whether to try and abstain that day, or just start drinking right away. If he was sure of his ability to resist throughout the afternoon and evening, when cravings 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), $t = 1, 2$, each of which is further divided into two subperiods, I and II (e.g., morning and afternoon), see Fig. 1. At the start of each subperiod I, the individual

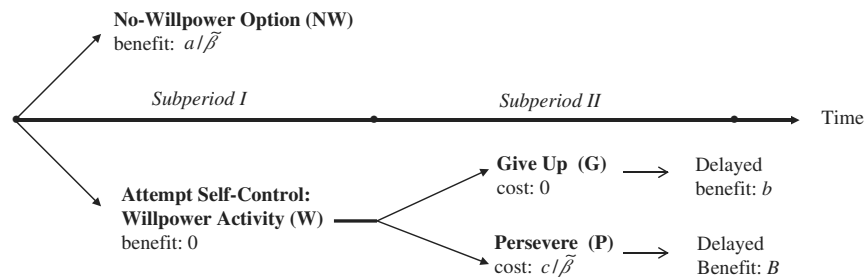


Fig. 1. Decisions and payoffs in any given period $t = 1, 2$. The parameter $\tilde{\beta}$ measures the salience of the present: for the current self $\tilde{\beta} = \beta < 1$, while for the ex ante self $\tilde{\beta} = 1$.

chooses between:

(1) A “no willpower” activity (*NW*), which yields 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.⁶

(2) 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 the intensity of temptation that he then experiences, the individual may opt, at the beginning of subperiod II, to either *persevere* until completion (*P*), or *give up* along the way (*G*).

Evaluated from an *ex ante* point of view (that of the agent’s date-zero “self”), these different courses of action result in the following payoffs. Perseverance entails a “craving” cost $c > 0$ during subperiod II, but yields delayed gratification in the form of future benefits (better health, higher consumption etc.) whose present value, starting at the end of period t , is B . As explained below, c takes values c_L or c_H for different individuals, and is only imperfectly known by the agent himself. Caving in, on the other hand, results in a painless subperiod II but yields only a delayed payoff b , where $a < b < B$. The assumption that $b > a$ means that *some* self-restraint (resisting for a while but eventually giving up) is better than none at all. 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 that may cause them to succumb to short-run impulses at the expense of their long-run interests. We thus assume that, in addition to a standard discount rate δ between periods 1 and 2, their time preferences exhibit the usual quasi-hyperbolic profile: at any decision node, the individual overestimates the gratification from an immediate payoff by a factor of $1/\beta$, or correspondingly discounts all future payoffs at a rate $\beta < 1$.

The second key assumption is that the intensity of the cravings to which an individual will be subject if he attempts self-restraint 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 introspection or memory search. As a result, the agent in period 2 will have to try and *infer* his vulnerability to temptation from his own actions (“how did I behave last night, and what kind of a person does that make me?”) and those of his peers. We discuss this assumption of imperfect self-knowledge in more detail below. First, we state it formally and show how it combines with imperfect willpower (hyperbolic preferences) to generate a self-reputational “stake” in good behavior.

We assume that agents know their general degree of present-orientation, and for simplicity we take it to be the same $\beta < 1$ for everyone. By contrast, the activity-specific cost c differs across individuals, taking values $c = c_L$ or $c = c_H$, with $c_L < c_H$. A low-cost individual will also be referred to as a “strong type”, a high-cost as a “weak type”, where “strength” is here the ability to deal with the temptation of *G*. 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 .

⁶ Note that *W* need not yield a flow payoff only in subperiod I: a could be the present value, evaluated at (t, I) , of an immediate payoff plus later ones. The important assumption is that there be *some* immediate reward to choosing *NW*. Similarly, *NW* could also lead to the *P/G* decision node but with a lower probability than *W*, without changing any of the results.

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

$$\frac{c_L}{\beta} < B - b < \frac{c_H}{\beta}. \quad (1)$$

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 his short-run impulses.⁷

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

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

Turning now to the agent's choice at the start of period 2, he will clearly only embark on a course of self-restraint when he has sufficient confidence in his ability to “follow through”. Since reputational concerns no longer operate, the expected return from attempting W exceeds the immediate (and more salient) payoff from NW only if his updated self reputation ρ' is above the threshold ρ^* defined by

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

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 of $\rho'(B - c_L) + (1 - \rho')b$ and a . A higher level of confidence ρ' in one's ability to resist temptation is then a valuable asset, because it helps offset the natural tendency to “give up without trying”. In particular, the fact that $\beta < 1$ creates an incentive for the weak type to *pool* with the strong one by persevering in the first period, so as to induce at least 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, cravings correspond to “hot,” internal, affective states, which are hard to remember later on from “cold” introspection. This intuitive idea is confirmed by experimental and field evidence on people's recollections of pain or discomfort [26] and their (mis)predictions of how they will behave under conditions of hunger, exhaustion, drug or

⁷ The precedent-setting role of lapses is emphasized by Ainslie [1]. Baumeister et al. [6] refer to it as “lapse-activated snowballing,” and Elster [22] as a “psychological domino effect”.

alcohol craving, or sexual arousal [30,31]. Second, an individual 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 ample evidence that people’s recollections are generally *self-serving*: they tend to remember (be consciously aware of) their successes more than their failures and find ways of absolving themselves of bad outcomes by attributing responsibility to others.⁸ Given imperfect or self-serving recall, introspection about one’s vulnerability to temptation is unlikely to 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., [7]). It is also supported by experimental evidence, such as Quattrone and Tversky’s [33] findings that people take actions, including painful ones, for self-signalling purposes.⁹ Such behaviors, one should again note, are conceivable only if later on the true motives and feelings behind one’s earlier actions can no longer be reliably recalled or accessed.

2.1.1. 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—whether exogenously or by choice—in a social relationship where he can observe the behavior of others. What makes such exposure 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, so that by observing *B*’s actions, *A* can learn something about himself. If *B* successfully resists temptation this news are encouraging to *A*, while if *B* caves in or has a relapse they are discouraging.

We assume that for each agent, the prior probability of being a low cost type is ρ . Moreover, types are correlated, with conditional probabilities:

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

where α is a parameter measuring correlation.¹⁰

For $\alpha = 0$ we get back the single-agent case (types are independent), while for $\alpha = 1$ correlation becomes perfect. This simple structure also has the advantage that changes in α leave the unconditional probabilities unchanged, and vice-versa. This will allow comparative statics that cleanly separate the effects of initial reputation and of correlation. Finally, we have assumed a completely symmetric situation; in particular, the two agents enter the game with the same level of self-confidence ρ , their preference structure is the same, and this is common knowledge. In this case there are only symmetric equilibria, as shown later on.

⁸ See Bénabou and Tirole [9] 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.

⁹ In their experiment, subjects were led to believe that increased tolerance, following physical exercise, for keeping one’s hand in near-freezing water was diagnostic of either a good or a bad heart condition. They reacted by, respectively, extending or shortening the amount of time they withstood that pain.

¹⁰ The probabilities that both agents are low types, high types, or of opposite types are then $\rho^2 + \alpha\rho(1 - \rho)$, $(1 - \rho)^2 + \alpha\rho(1 - \rho)$ and $(1 - \alpha)\rho(1 - \rho)$, respectively.

In Section 4 we shall extend the analysis to asymmetric initial conditions, payoffs, and equilibria.

3. Homogeneous peer groups

3.1. Main intuitions

3.1.1. The single-agent benchmark

We begin with the one-agent case, which provides a natural starting point to understand group interactions and evaluate their welfare implications. Given that the strong type always perseveres, the 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 by the dashed middle line $x_a(\rho)$ in Fig. 2; the subscript a stands for “alone”.¹¹ Complete self-restraint (perfect pooling) by Self 1 makes observing P completely uninformative for Self 2, leaving his prior unchanged; it is therefore an equilibrium only when the agent’s initial 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

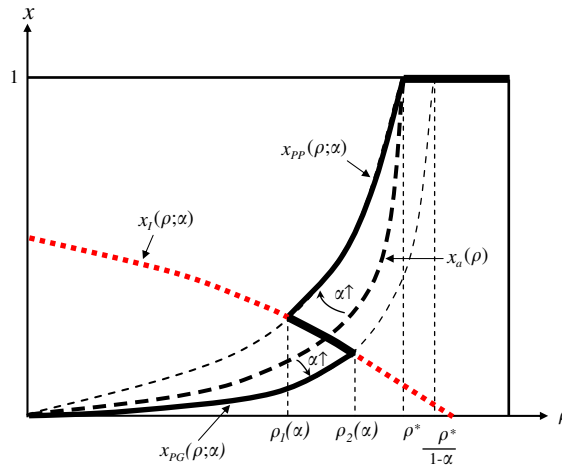


Fig. 2. Equilibrium self-restraint for a single individual (dashed line) and in a peer group (solid lines).

¹¹ The figure describes the weak type’s strategy in the (most interesting) subgame where the decision node between P and G has been reached. This confrontation with cravings could be the result of a 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.

between playing P and G).¹² Conversely, 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*, $\Pr_{x,\rho}(c = c_L | P) = \rho^*$, 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 $\rho = \rho^*$.

3.1.2. Two agents

Let us now bring together two individuals whose types are correlated as described above and examine how this affects the behavior of weak types at the temptation stage. As mentioned earlier, we focus until Section 4 on equilibria 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 that the correlation α between types is relatively low. By continuity, equilibrium behavior will not be too different from that of the single agent case; the interesting issue is 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 in Fig. 2.

(a) When initial reputation is low, j is most probably also a weak type, who will play a strategy close to $x_a(\rho) \approx 0$. Consequently, he is almost sure to be a source of "bad news" (G) that will reduce i 's hard-earned reputational gain from playing P . This *discouragement* effect naturally leads agent i to persevere with lower probability $x_{PG}(\rho; \alpha) < x_a(\rho)$, as indicated by the solid curve emanating from the origin. Intuitively, i must now 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.

(b) When initial reputation is high (just below ρ^*), j is now either a strong type or a weak type who exerts self-control with probability close to $x_a(\rho) \approx 1$. Therefore, 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, the weak type's probability of playing P increases to $x_{PP}(\rho; \alpha) > x_a(\rho)$, as illustrated by the solid curve that rises up to $(\rho^*, 1)$. In this case, the positive externality allows the agent to engage in more pooling.

(c) Where ρ is in some intermediate range, finally, if i plays P both PG and PP have non-negligible probability, and which one ends up shaping equilibrium strategies 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 . Due to this strategic complementarity (which operates purely through joint informational spillovers on

¹² The mixed-strategy nature of most equilibria in our model is, as usual, an artefact of the discreteness of the type space. As in most other dynamic games of incomplete information (e.g., bargaining games) it would disappear with a continuum of types.

the decision of Self 2), both $x_{PP}(\rho; \alpha)$ and $x_{PG}(\rho; \alpha)$ are equilibria over some range of ρ ; see Fig. 2. As usual, a third equilibrium $x_I(\rho; \alpha)$ then also exists in-between; it will be described in more detail below.

3.1.3. Increasing the correlation

As α increases the x_{PG} locus pivots down, while the x_{PP} locus pivots up (see Fig. 2): 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.

3.2. Equilibrium group behavior

3.2.1. 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:

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

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 $\rho > \rho^*$.¹³ 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. Eq. (4) thus defines two upward-sloping loci in the (ρ, x) plane, between which any equilibrium must lie:

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

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 in Fig. 2, x_{PP} increases with ρ up to $\rho = \rho^*$, 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

¹³ 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) .

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 this pivotal state, $\mu_{PP} = \rho^*$.
- (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, $\mu_{PG} = \rho^*$, that is relevant.
- (iii) *Intermediate 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 .

3.2.2. The incentive constraint

We now determine exactly when each scenario applies. In order for the weak type to be willing to mix between P and 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 use strategy x and expects his own Self 2 to choose W with probabilities y and y' following events PP and PG respectively. Since a weak type will reap payoff b under W rather than a under NW , we have

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

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

$$\pi(x; \rho, \alpha) \equiv B - b - \frac{c_H}{\beta} + \delta[(1-\alpha)\rho + (1-(1-\alpha)\rho)x](b-a) = 0. \quad (9)$$

This equation uniquely defines a downward-sloping locus $x_I(\rho; \alpha)$ in the (x, ρ) plane, which we shall refer to 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 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 and 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} curve 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 $\pi(x; \rho, \alpha) = \Pi(x, 1, 0; \rho, \alpha) \geq 0$.
- Intermediate equilibria correspond precisely to the portion of the incentive locus x_I that 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 in Fig. 2. Formally:

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

- For $\rho < \rho_1(\alpha)$ there is a unique equilibrium, which is of the “bad news” type: $x = x_{PG}(\rho; \alpha)$.
- For $\rho > \rho_2(\alpha)$ there is a unique equilibrium, which is of the “good news” type: $x = x_{PP}(\rho; \alpha)$.
- 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: $\lim_{\alpha \rightarrow 0} |\rho_2(\alpha) - \rho_1(\alpha)| = 0$.

Fig. 3 provides a convenient representation of these results in the (ρ, α) space.¹⁴ As correlation declines to zero, the area between $\rho_1(\alpha)$ and $\rho_2(\alpha)$ where multiplicity occurs

¹⁴ 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. To understand the shape of $\rho_1(\alpha)$ and $\rho_2(\alpha)$, recall that $x_{PP}(\cdot; \alpha)$ shifts up with α , while $x_{PG}(\cdot; \alpha)$ shifts down: a greater correlation magnifies both the “discouragement” and the “encouragement” effects of the other agent's choosing G or P , respectively. The incentive constraint $x_I(\cdot, \alpha)$, meanwhile, 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. Therefore, $\rho_2(\alpha)$, which is the intersection of $x_{PG}(\cdot; \alpha)$ and $x_I(\cdot; \alpha)$ is increasing in α ; $\rho_1(\alpha)$, by contrast, need not be monotonic.

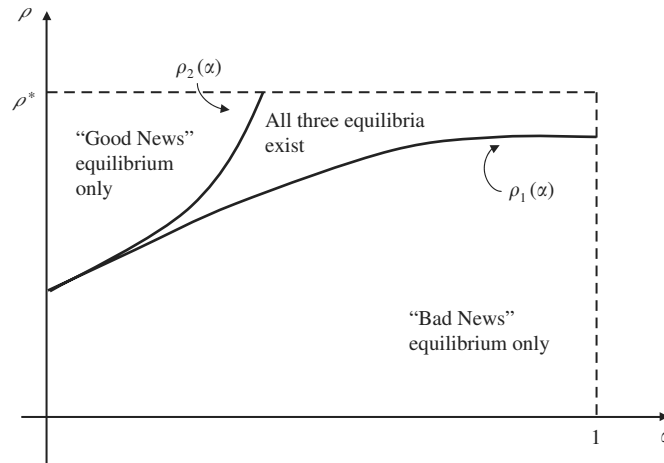


Fig. 3. Equilibrium outcomes for different levels of self-confidence and correlation.

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*.

Note that the *PG* equilibrium exists only when ρ is not too high or the degree of correlation α is large enough, while the reverse conditions are needed to sustain the *PP* equilibrium. Indeed, the first case requires a weak agent to be relatively pessimistic about his partner's type (hence about the latter's likelihood of choosing *P*), while in the second he must be sufficiently optimistic.

3.3. Welfare analysis

We shall now 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 make clear when groups do indeed provide valuable “help,” and when they actually do damage.

The question of what welfare function to use in a model where preferences change over time is a controversial one, and in our model with imperfect self-knowledge it could, a priori, be even more complicated. The results we obtain, however, are fully consistent across the different possible criteria. To understand why, consider first an agent's initial decision of whether or not to join a group. At this stage he does not yet know his type, and his temporal preferences are not yet subject to present bias. He thus makes his decision by computing the undistorted intertemporal payoffs W^s and W^w that he will reap if he turns out to be strong or weak, then examining whether the expectation $W = \rho W^s + (1 - \rho) W^w$ is higher in isolation or in a group. We shall thus be interested in *ex ante* welfare W from a positive as well as a normative point of view. The (undistorted) *interim* utility levels of each type, W^s and W^w , are essential components of this criterion; furthermore, in our model they also completely determine the value of *any* social welfare criterion that puts weight on *both* ex

ante and ex post preferences.¹⁵ This is because: (i) the strong type always perseveres, so his ex post welfare (evaluated at the time of temptation) just differs from W^s by a constant: $W^{s,\beta} = W^s - c_H(1/\beta - 1)$; (ii) the weak type always randomizes between P and G , so his welfare is the same as if he systematically chose G : $W^{w,\beta} = b + \delta a$.¹⁶ Consequently, any proposition established for W^s (respectively, W^w) immediately carries over to all (linear or nonlinear) aggregates of W^s and $W^{s,\beta}$ (respectively, W^w and $W^{w,\beta}$), i.e. to all type-specific welfare criteria. Similarly, any result on ex ante welfare $W = \rho W^s + (1 - \rho)W^w$ carries over to any weighted average of the non-tempted and tempted selves, W and $W^\beta = \rho W^{s,\beta} + (1 - \rho)W^{w,\beta}$. Finally, and perhaps most importantly, interventions that (say) raise interim welfare for both types, W^s and W^w , represent true *Pareto improvements* in any sense of the word: they make the agent better off whether his payoffs are evaluated with or without present bias, and with or without information about his type.¹⁷

We start with the natural benchmark case where the agent is alone (equivalently, $\alpha = 0$). For the weak type,

$$W_a^w = b + \delta a + x_a[B - b - c_H + \delta y_a(b - a)], \quad (10)$$

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 . Next, for the weak type to be indifferent at the temptation stage, it must be that

$$B - c_H/\beta + \delta[y_a b + (1 - y_a)a] = b + \delta a. \quad (11)$$

Substituting into (10) yields:

$$W_a^w = b + \delta a + x_a \left(\frac{1 - \beta}{\beta} \right) c_H, \quad (12)$$

in which the second term reflects the *value of the self-discipline* achieved through the reputational mechanism. Turning now to expected welfare for a strong type, we can write:

$$W_a^s = B - c_L + \delta[y_a(B - c_L) + (1 - y_a)a]. \quad (13)$$

Because $y_a < 1$ for all $\rho < \rho^*$, the strong type's average payoff is always less than $B - c_L$, which is what he would achieve under perfect information, or in a one-shot context. He is

¹⁵ Ex post utility levels, denoted as $W^{s,\beta}$ and $W^{w,\beta}$, refer here to the preferences of the second-subperiod self, which incorporate the present bias. Caplin and Leahy [15], for instance, argue that in problems with changing preferences one should aggregate the expected utilities of the different temporal selves using a Bergsonian welfare criterion, as in a standard social choice problem.

¹⁶ Throughout the welfare analysis we focus on the case where $\rho < \rho^*$ (otherwise, peers are irrelevant to self-control). We also assume that at $t = 1$ the willpower activity is undertaken (either by choice or because it cannot be avoided for sure; see footnote 11), so that he agents is indeed confronted with temptation.

¹⁷ The interim levels W^s and W^w are also of further interest in situations where the agent interacts with a better informed but altruistic principal (see [16] for such a model in the context of medical advice). Consider a parent deciding whether or not to let her child frequent certain peers, or a judge deciding whether a substance abuser should be compelled to join a "twelve-step" program. This principal (whether purely paternalistic or also concerned about externalities) will often have evidence (typically of a "soft", nonverifiable nature) on the agent's type that the latter does not, or is in denial about; she will then evaluate group membership for the agent based on her own priors over W^s and W^w . (Again, putting weight as well on ex post, salience-distorted payoffs does not change anything).

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 interactive cases discussed above: welfare in the Good News equilibrium and in the Bad News equilibrium. Since the analysis of the Intermediate equilibrium is technically very similar, it is presented in Appendix A. Readers who would like to skip the derivation of all the welfare results may go directly to Section 3.3.3, which summarizes the main insights.

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

$$W_{PP}^w = b + \delta a + x_{PP}[B - b - c_H + \delta \Pr_{PP}(P | w)y_{PP}(b - a)], \quad (14)$$

where $\Pr_{PP}(P | w) = 1 - \pi_{LL} + \pi_{LL}x_{PP}$ 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:

$$W_{PP}^w = W_a^w + (x_{PP} - x_a) \left(\frac{1 - \beta}{\beta} \right) c_H. \quad (15)$$

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:

$$W_{PP}^s = B - c_L + \delta a + \delta \Pr_{PP}(P | s)y_{PP}(B - b - c_L),$$

where $\Pr_{PP}(P | s) = \pi_{LL} + (1 - \pi_{LL})x_{PP}$ is the equilibrium probability that j will choose P , given that i is a strong type. Next, subtract (10) 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 $y_a = y_{PP}\Pr_{PP}(P | w)$. Thus

$$W_{PP}^s = W_a^s + \delta y_{PP}[\Pr_{PP}(P | s) - \Pr_{PP}(P | w)](B - a - c_L). \quad (16)$$

Thus, as long as $\alpha > 0$, the strong type is also strictly better off: $W_{PP}^s > W_a^s$. The intuition is that with two agents the payoff to i ’s playing P becomes contingent on what 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 in the single-agent case. 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 before: $y_{PP} > y_a$. This yields no extra surplus for the weak type, who remains indifferent, but generates rents for the strong type.

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

both types), and therefore also *ex ante*. The same remains true according to any social welfare criterion that puts positive weight on *ex post* as well as *ex ante* preferences.

The result that joining a group can bring about a *Pareto improvement*, rather than just transfer surplus across types or temporal selves, is somewhat surprising, since the presence of peers entails a trade-off between the positive informational spillover received when they persevere and the negative one suffered when they do 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. Thus 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.

3.3.2. Welfare in a Bad News equilibrium

Derivations similar to the previous case yield for the weak type:

$$W_{PG}^w = W_a^w + (x_{PG} - x_a) \left(\frac{1 - \beta}{\beta} \right) c_H. \quad (17)$$

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:

$$W_{PG}^s = W_a^s + \delta[\Pr_{PG}(P | s) - \Pr_{PG}(P | w)](1 - y_{PG})(B - a - c_L). \quad (18)$$

This makes clear 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.

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. The same remains true when each type's welfare is evaluated according to any welfare criterion that also puts positive weight on his *ex post* preferences.*

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 forming a group.

Proposition 4. *Assume that agents expect a Bad News equilibrium. There exist two values $0 < \underline{\rho}^* < \bar{\rho}^* < 1$ such that for all $\rho^* \in (\underline{\rho}^*, \bar{\rho}^*)$, 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^* .¹⁸*

3.3.3. The value of joining a group

We now briefly summarize the main results obtained so far. 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 (W is higher) but also at the interim stage (W^s and W^w are higher) as well as ex post ($W^{h,\beta}$ is higher, $W^{w,\beta}$ is unchanged). 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. From an ex ante point of view, forming a group is then beneficial only if self-confidence exceeds a minimal threshold.

Field studies of self-help groups for alcohol and drug abusers consistently find a strong positive correlation between self-esteem and “clean time” in the group [21,19]. The standard interpretation is that interactions with peers help individuals sustain desirable behavior, which in turn raises their self-esteem. This would be in line with our results concerning the Good News equilibrium, which is sustained by the collective building up and maintenance of self-confidence. Alternatively, the observed correlation could reflect self-selection, with low self-esteem individuals dropping out earlier. This second (non-exclusive) explanation is also consistent with our predictions: agents with very low self-confidence are always those who benefit least from group interactions, and may even prefer isolation (Bad News equilibrium).

4. Heterogeneous peer groups

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. The second and key 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

¹⁸ For $\rho^* < \underline{\rho}^*$ (resp. $\rho^* > \bar{\rho}^*$), 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.

or hurt his peers? Would anyone accept a partner whom they perceive to be weaker than themselves?

4.1. 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 $\rho^i \equiv \Pr(c^i = c_L)$, and the conditional probabilities as $\pi_{LL}^i \equiv \Pr(c^i = c_L | c^j = c_L)$ and $\pi_{HH}^i \equiv \Pr(c^i = c_H | c^j = c_H)$, for $i = 1, 2$.¹⁹ We only impose a general condition of positive correlation between agents' craving costs (monotone likelihood ratio property):

$$\frac{\Pr((c_H, c_L))}{\Pr((c_L, c_L))} < \frac{\Pr((c_H, c_H))}{\Pr((c_L, c_H))}. \quad (19)$$

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

4.1.1. Informativeness constraints

Let $\mu_{PP}^i(x^i, x^j)$ and $\mu_{PG}^i(x^i)$ 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.²⁰ The same simple reasoning as in Section 3.2 shows that, in any equilibrium, these beliefs must satisfy:

$$\mu_{PG}^i(x^i) \leq \rho^{i*} \leq \mu_{PP}^i(x^i, x^j). \quad (20)$$

As shown in the appendix and illustrated in Fig. 4, each equation $\mu_{PP}^i(x^i, x^j) = \rho^{i*}$ uniquely defines a downward-sloping function $x^i = X_{PP}^i(x^j)$, with $(X_{PP}^1)^{-1}$ steeper than X_{PP}^2 . As long as the two agents are not excessively different from one another, there is then a unique intersection $E_{GN} = (x_{PP}^1, x_{PP}^2) \in (0, 1) \times (0, 1)$, where both (weak) agents play their "good news" strategies.²¹ Similarly, each equation $\mu_{PG}^i(x^i) = \rho^{i*}$ has a unique solution

¹⁹ We shall similarly denote $\pi_{HL}^i \equiv 1 - \pi_{LL}^i$ and $\pi_{LH}^i \equiv 1 - \pi_{HH}^i$. Condition (19) below is then equivalent to $\pi_{LH}^i/\pi_{LL}^i < 1 < \pi_{HH}^i/\pi_{HL}^i$, for $i = 1, 2$.

²⁰ Clearly, $\mu_{PG}^i(x^i)$ is independent of x^j : once agent j has given in, his type is completely revealed. The functions μ_{PP}^i and μ_{PG}^i 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.

²¹ For simplicity, we shall focus on this case from here on. The case where any of the intersections x_{PP}^i occurs outside the $(0, 1) \times (0, 1)$ box is easily analyzed using the techniques developed in this section, and it yields the same intuitions.

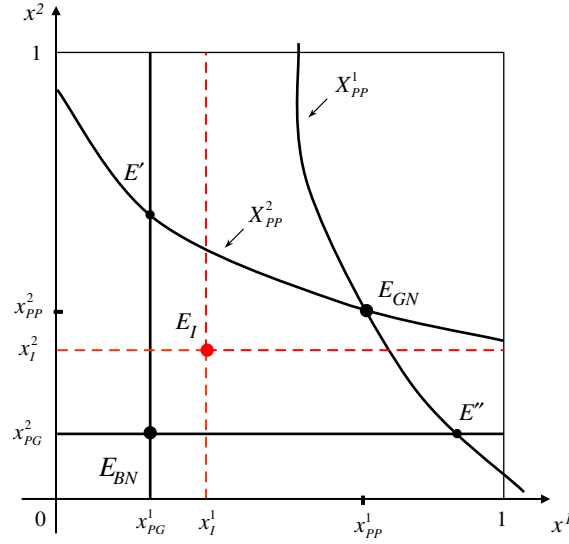


Fig. 4. Good News, Bad News and intermediate equilibria.

$x^i = x^i_{PG}$, which corresponds in Fig. 4 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 (20) define a “permissible region” $E_{BN}E'E_{GN}E''$ within which any equilibrium must lie:

$$x^i_{PG} \leq x^i \leq X^i_{PP}(x^j). \tag{21}$$

4.1.2. 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 (in \mathbb{R}) to the linear equation $\pi^i(x^j) = 0$.

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

$$\begin{cases} \text{if } \rho^{i*} < \mu^i_{PP}(x^i, x^j) \text{ then } \pi^i(x^j) \leq 0, \\ \text{if } \rho^{i*} > \mu^i_{PG}(x^i) \text{ then } \pi^i(x^j) \geq 0, \end{cases} \tag{22}$$

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

$$\begin{cases} \text{if } x^j > x^j_I \text{ then } x^i = X^i_{PP}(x^j); \\ \text{if } x^j < x^j_I \text{ then } x^i = x^i_{PG}. \end{cases} \tag{23}$$

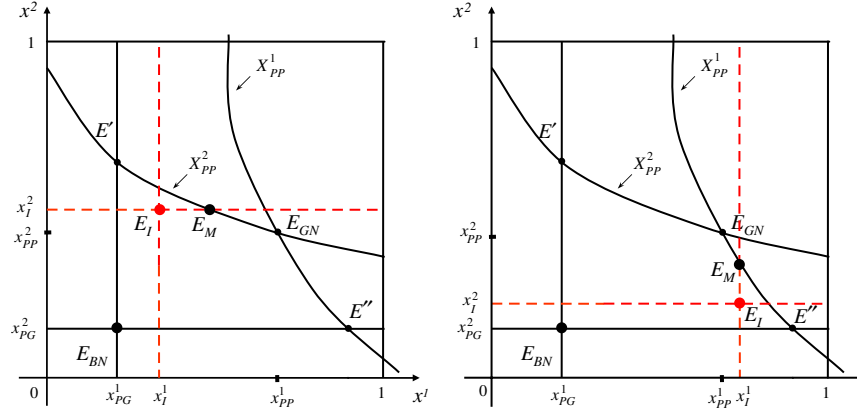


Fig. 5. Mixed, Intermediate, and Bad News equilibria.

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 (23), 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 $E_M \equiv ((X_{PP}^2)^{-1}(x_I^2), x_I^1)$ when it lies on the upper boundary of that region, as on the left panel of Fig. 5; (iii) the point $E_M \equiv (x_I^1, (X_{PP}^1)^{-1}(x_I^1))$ when it lies on the right boundary of that same region, as on the right panel of Fig. 5.

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 shall focus here on the case where all three possible types of equilibria coexist, so that we can analyze the comparative statics of each one. The complete analysis of the other possible cases is presented in [5] as well as in Appendix B, which is available through the on-line edition of this journal.

It is easily seen from (23) that a necessary and sufficient condition for such multiplicity is that the point E_I lie in the permissible region of Figs. 4 and 5, that is,

$$x_{PG}^i < x_I^i < X_{PP}^i(x_I^i), \quad \text{for } i = 1, 2. \tag{24}$$

Proposition 5. *Let condition (24) hold. 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 $x_I^2 > x_{PP}^2$ then $S = \{E_{BN}, E_I, E_M \equiv ((X_{PP}^2)^{-1}(x_I^2), x_I^1)\}$.*
- (iii) *If $x_{PG}^2 < x_I^2 < x_{PP}^2$ but $x_I^1 > x_{PP}^1$ then $S = \{E_{BN}, E_I, E_M \equiv (x_I^1, (X_{PP}^1)^{-1}(x_I^1))\}$.*

Thus, under condition (24) there is an equilibrium where both agents are in a “bad news” regime, another one where both are in an “intermediate” 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 “intermediate” 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 Fig. 5. In such a situation, the informativeness constraint $\pi^j = 0$ is binding on one agent and the incentive constraint $\mu_{PP}^i(x^i, x^j) = \rho^{i*}$ 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 thus for their respective intersections, E_{GN} and E_{BN} .

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 neither one 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).

4.2. 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

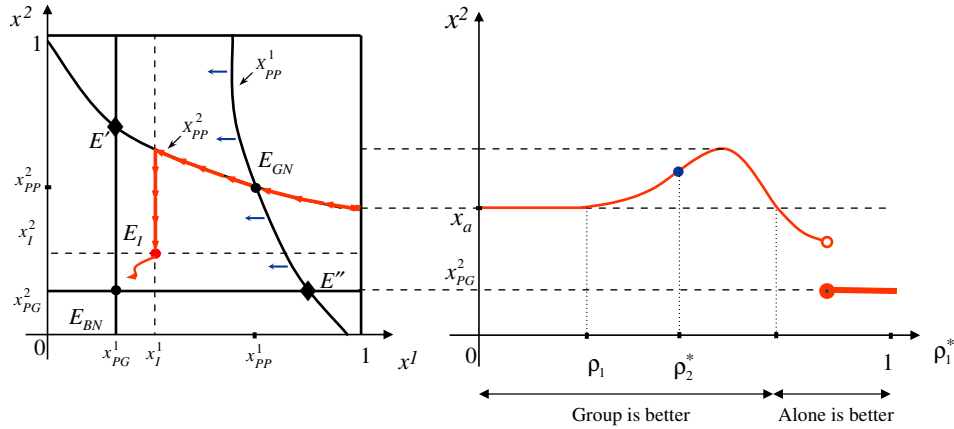


Fig. 6. The effect on agent 2 of the severity of his peer’s potential self-control problem. The right panel depicts both agent 2’s behavior x^2 (when weak) and his ex ante welfare $W^2 = \rho^2 W^{2,s} + (1 - \rho^2) W^{2,w}$. Indeed, $W^{2,w}$ strictly increases with x^2 , while $W^{2,s}$ is always nondecreasing in ρ^{1*} .

examine the effects on agent 2 of variations in ρ^{1*} , without having to specify their ultimate source.²²

Rather than examine the local comparative statics of each equilibrium separately we shall integrate them into a more interesting *global analysis* that allows 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.²³ The key results are illustrated on the right panel of Fig. 6.

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 believed to be a little weaker than him, that is, who has a ρ^{j*} somewhat above ρ^{i*} .*

²² 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, $\rho^2 = \rho^1 \pi_{LL}^2 + (1 - \rho^1) (1 - \pi_{HH}^2)$. For instance, if it is common knowledge that both agents are always of the same type ($\pi_{HH}^i = \pi_{LL}^i = 1$), then $\rho^1 \equiv \rho^2$. Conversely, for ρ^2 to remain unaffected, the conditional probabilities π_{LL}^2 and π_{HH}^2 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.

²³ The comparative statics of the Intermediate 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 the results (see Proposition 6) apply unchanged when there is a unique equilibrium that is of the Good News or Mixed type.

(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. Fig. 6 shows that for relatively low values of 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.

Proposition 6 can be derived by means of a simple graphical analysis. As ρ^{1*} increases from 0 to 1 the $X_{p,p}^1$ locus shifts left, as indicated on the left panel of Fig. 6; consequently, the high self-restraint equilibrium travels along the path marked by the thick arrows. The implied self-control behavior (and welfare) of agent 2 can then simply be read off the right panel of the figure. We omit here the complete proof for reason of space; it can be found in [5] as well as in Appendix B, which is available through the on-line edition of this journal.

5. 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 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 a person'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 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 psychology literature. Nonetheless, it is still clearly oversimplistic, and could be extended in several directions. First, 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. Second, there are a number of important aspects of peer interaction from which we abstract. Some, like learning specific techniques to deal with impulses, are quite consistent with our approach and could easily be incorporated. Others, involving a desire to “belong”, being helped by the “moral support” of others, or basic emotional mechanisms such as embarrassment at having to admit failure in front of others and deriving pride from public success, would require more substantial extensions. Another interesting direction for further research would be to explore peer effects that involve *excessive*, rather than insufficient, self-regulation.²⁴ 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.

Acknowledgments

We are grateful for helpful comments to Jess Benhabib, Leonardo Felli, Ted O’Donoghue, John Morgan, Michele Piccione, Matthew Rabin, Tom Romer and two anonymous referees, as well as to seminar participants at the 2001 Congress of the European Economic Association, the Studienzentrum Gerzensee, the Harvard–MIT theory seminar, the London School of Economics, the University of Toronto, the Wallis Institute at Rochester University and the Stanford Institute for Theoretical Economics. Battaglini gratefully acknowledges the hospitality of the Economics Department at the Massachusetts Institute of Technology during the academic year 2002–2003. Bénabou gratefully acknowledges financial support from the National Science Foundation and the John Simon Guggenheim Foundation, and the hospitality of the Institute for Advanced Study during the academic year 2002–2003.

Appendix A.

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 $\mu_{PP}^i(x^1, x^2) = \rho^{i*}$ and $\mu_{PG}^i(x^i) = \rho^{i*}$, for $i = 1, 2$. The following lemma establishes these properties:

²⁴ See Bodner and Prelec [12] and Benabou and Tirole [10] for accounts of rigid behavior and compulsive personal rules in a single-agent setting.

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

- (i) The loci $X_{PP}^i(x^i, x^j)$ are decreasing in x^j . Furthermore $X_{PP}^2(x^1)$ cuts $(X_{PP}^1)^{-1}(x^1)$ at most once in the positive orthant, and if it does the intersection is from below.
- (ii) If $\rho^i < \rho^{i*}$ and the two agents are not excessively different from one another, there is a unique interior solution for each system of equations: namely, $(x_{PP}^1, x_{PP}^2) \in (0, 1) \times (0, 1)$ and $(x_{PG}^1, x_{GP}^2) \in (0, 1) \times (0, 1)$.

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

$$\frac{\mu_{PP}^i(x^i, x^j)}{1 - \mu_{PP}^i(x^i, x^j)} = \frac{\Pr(c^i = c_L, c^j = c_L) + \Pr(c^i = c_L, c^j = c_H)x^j}{\Pr(c^i = c_H, c^j = c_L)x^i + \Pr(c^i = c_H, c^j = c_H)x^i x^j}, \quad (25)$$

$$\frac{\mu_{PG}^i(x^i)}{1 - \mu_{PG}^i(x^i)} = \frac{\Pr(c^i = c_L, c^j = c_H)}{\Pr(c^i = c_H, c^j = c_H)x^i}. \quad (26)$$

Clearly, μ_{PP}^i and μ_{PG}^i are both decreasing in x^i . To see that μ_{PP}^i is decreasing in x^j as well, note that $\partial \mu_{PP}^i(x^i, x^j) / \partial x^j$ 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 (19). Therefore $\partial X_{PP}^i(x^i, x^j) / \partial x^j < 0$ by the implicit function theorem. Next, note that $X_{PP}^2(0)$ is bounded for $x^1 \in [0, 1]$. By contrast, we can easily verify that $\lim_{x^1 \rightarrow 0} (X_{PP}^1)^{-1}(x^1) = +\infty$. Therefore, there exists a point x^1 small enough that $X_{PP}^2(x^1) < (X_{PP}^1)^{-1}(x^1)$. 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 $X_{PP}^2(x^1)$ crossing $(X_{PP}^1)^{-1}(x^1)$ from below. Note first that any intersection must be such that $\mu_{PP}^1(x^1, x^2) / \mu_{PP}^2(x^1, x^2) = \rho^{1*} / \rho^{2*}$. By (25), this implies

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

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

(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. \square

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 Intermediate 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 $x_{PP}(\rho; \alpha) \rightarrow x_{PG}(\rho; \alpha)$, it is immediate to see that $\lim_{\alpha \rightarrow 0} |\rho_2(\alpha) - \rho_1(\alpha)| = 0$. \square

Proof of Propositions 2–4. The first two were established in the text; we prove here the third one. A Bad News equilibrium is ex ante preferable to staying alone when

$$E(W_{PG} - W_a | \rho) \equiv \rho(W_{PG}^s - W_a^s) + (1 - \rho)(W_{PG}^w - W_a^w) > 0. \quad (27)$$

From the informativeness constraint (25) 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 (17) and (18), we can then rewrite (27) as:

$$\Psi(\rho, \rho^*) \equiv (\rho^* - 1)k(\rho) + \rho^* - (1 - \alpha)\rho < 0, \quad \text{where} \quad (28)$$

$$k(\rho) \equiv \frac{(1 - \beta)c_H}{\beta\delta(1 - y_{PG}(\rho))(B - c_L - a)}. \quad (29)$$

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

$$\begin{aligned} B - b - \frac{c_H}{\beta} + \delta[(1 - \alpha)\rho + (1 - (1 - \alpha)\rho)(x_{PG}(\rho) + (1 - x_{PG}(\rho))y')] & (b - a) \\ & = 0, \end{aligned}$$

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 (27) 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 $\bar{\rho}^*$ by the linear equations $\Psi(0, \underline{\rho}^*) = (\underline{\rho}^* - 1)k(0) + \underline{\rho}^* \equiv 0$ and $\Psi(1, \bar{\rho}^*) = (\bar{\rho}^* - 1)k(1) + \bar{\rho}^* - (1 - \alpha)$ respectively. Then $0 < \underline{\rho}^* < \bar{\rho}^* < 1$ and for any ρ^* in $(\underline{\rho}^*, \bar{\rho}^*)$, $\hat{\rho}(\rho^*)$ lies in $(0, \rho^*)$ and is strictly increasing in ρ^* . For $\rho^* < \underline{\rho}^*$ we have $\hat{\rho}(\rho^*) = 0$, and $E(W_{PG} - W_a | \rho) > 0$ for all $\rho \geq 0$. Conversely, for $\rho^* > \bar{\rho}^*$ we have $\hat{\rho}(\rho^*) = \rho^*$, and $E(W_{PG} - W_a | \rho) < 0$ for all $\rho \leq \rho^*$. \square

A.1. Welfare in an Intermediate equilibrium

For the weak type, we have as usual $W_I^w = W_a^w + (x_I - x_a)[(1 - \beta)/\beta]c_H$. Recall from Fig. 2 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:

$$\begin{aligned} W_I^s &= W_a^s + \delta[\Pr_I(P | s) - \Pr_I(P | w)](B - a - c_L) \\ &= W_a^s + \delta\alpha(1 - x_I)(B - a - c_L). \end{aligned}$$

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 Intermediate 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 strong type is better off).

Proof of Proposition 5. We first prove condition (22).

(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 $\Pi^i(x^j, 1, y_{PG}^i) \geq \pi^i(x^j) > 0$ 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 $X_{PP}^j(x) - (X_{PP}^j)^{-1}(x)$ has the sign of $x_{PP}^i - x$ for all x (single-crossing property established by Lemma 1 and illustrated in Fig. 4), this implies that $x^j < (X_{PP}^j)^{-1}(1)$, or equivalently $\mu_{PP}^j(x^j, 1) > \rho^{j*}$. As a result, agent j 's second-period self will choose $y_{PP}^j = 1$, ensuring $\Pi^j(1, 1, y_{PG}^j) = \Pi^j(1, 1, 0) > 0$. This leads to $x^j = 1$, a contradiction.

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

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

Proof of Proposition 6. See Appendix B.

Appendix B. Supplementary data

The on-line version of this article contain additional supplementary data. Please visit [doi:10.1016/j.jet.2005.04.001](https://doi.org/10.1016/j.jet.2005.04.001)

References

- [1] G. Ainslie, Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person (Studies in Rationality and Social Change), Cambridge University Press, Cambridge, UK, New York, 1992.
- [2] G. Ainslie, Breakdown of Will, Cambridge University Press, Cambridge, UK, 2001.
- [3] A. Banerjee, T. Besley, Peer group externalities and learning incentives: A theory of nerd behavior, Princeton University Working Paper No. 68, 1990.
- [4] M. Battaglini, R. Bénabou, Trust, coordinations, and the industrial organization of political activism, J. Europ. Econ. Assoc. 1 (4) (2003) 851–889.
- [5] M. Battaglini, R. Bénabou, J. Tirole, Self-Control in Peer Groups, CEPR Discussion Paper 3149, January 2002.
- [6] R. Baumeister, T. Heatherton, D. Tice, Losing Control: How and Why People Fail at Self-Regulation, Academic Press, San Diego, CA, 1994.
- [7] D. Bem, Self-perception theory, in: L. Berkowitz (Ed.), Advances in Experimental Social Psychology, Academic Press, New York, NY, 1972.

- [8] R. Bénabou, Workings of a city: Location, education and production, *Quart. J. Econ.* 108 (1993) 619–652.
- [9] R. Bénabou, J. Tirole, Self-confidence and personal motivation, *Quart. J. Econ.* 117 (3) (2002) 871–915.
- [10] R. Bénabou, J. Tirole, Willpower and personal rules, *J. Polit. Econ.* 112 (4) (2004) 848–887.
- [11] D. Bernheim, A theory of conformity, *J. Polit. Econ.* 102 (5) (1994) 841–877.
- [12] R. Bodner, D. Prelec, Self-signaling and diagnostic utility in everyday decision making, in: I. Brocas, J. Carrillo (Eds.), *Collected Essays in Psychology and Economics*, vol. I, Oxford University Press, Oxford, 2003.
- [13] I. Brocas, J. Carrillo, Rush and procrastination under hyperbolic discounting and interdependent activities, *J. Risk Uncertainty* 22 (2) (2001) 141–144.
- [14] W. Brock, S. Durlauf, Discrete choice with social interactions, *Rev. Econ. Stud.* 68 (2) (2001) 235–260.
- [15] A. Caplin, J. Leahy, The social discount rate, NBER Working Paper 7983, 2000.
- [16] A. Caplin, J. Leahy, The supply of information by a concerned expert, *Econ. J.* 114 (497) (2004) 487–505.
- [17] J. Carrillo, T. Mariotti, Strategic ignorance as a self-disciplining device, *Rev. Econ. Stud.* 67 (3) (2000) 529–544.
- [18] A. Case, L. Katz, The company you keep: The effects of family and neighborhood on disadvantaged youth, NBER Working Paper 3705, 1991.
- [19] G. Christo, S. Sutton, Anxiety and self-esteem as a function of abstinence time among recovering addicts attending Narcotics Anonymous, *Brit. J. Clinical Psychol.* 33 (1994) 198–200.
- [20] C. De Bartolome, Equilibrium and inefficiency in a community model with peer group effects, *J. Polit. Econ.* 98 (1990) 10–133.
- [21] C.B. De Soto, W.E. O'Donnell, L.J. Allred, C.E. Lopes, Symptomatology in alcoholics at various stages of abstinence, *Alcoholism: Clinical Exper. Res.* 9 (1985) 505–512.
- [22] J. Elster, Introduction, in: J. Elster (Ed.), *Addiction: Entries and Exits*, Russel Sage Foundation, New York, 2001.
- [23] A. Gaviria, S. Raphael, School-based peer effects and Juvenile behavior, *Rev. Econ. Statist.* 83 (2) (2001) 257–268.
- [24] E. Glaeser, J. Scheinkman, Non-market interactions, in: M. Dewatripont, L.P. Hansen, S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications*, Eight World Congress, Cambridge University Press, Cambridge, MA, 2002.
- [25] C. Hoxby, Peer effects in the classroom: Learning from gender and race variation, NBER W.P. No. 7867, 2001.
- [26] D. Kahneman, P. Wakker, R. Sarin, Back to Bentham? Explorations of experienced utility, *Quart. J. Econ.* 112 (2) (1997) 375–405.
- [27] P. Kooreman, Time, money, peers and parents: Some data and theories on teenage behavior, IZA Discussion Paper No. 931, November 2003.
- [28] M. Kremer, D. Levy, Peer effects and alcohol use among college students, NBER W.P. No. 9876, July 2003.
- [29] D. Laibson, Golden eggs and hyperbolic discounting, *Quart. J. Econ.* 112 (1997) 443–478.
- [30] G. Loewenstein, Out of control: Visceral influences in behavior, *Organ. Behav. Human Dec. Process.* 65 (3) (1996) 272–292.
- [31] G. Loewenstein, D. Schkade, Wouldn't it be nice? Predicting future feelings, in: D. Kahneman, E. Diener, N. Schwartz (Eds.), *Well-Being: Foundations of Hedonic Psychology*, Russel Sage Foundation, New York, NY, 1999.
- [32] T. O'Donoghue, M. Rabin, Doing it now or later, *Amer. Econ. Rev.* 89 (1) (1999) 103–124.
- [33] G. Quattrone, A. Tversky, Causal versus diagnostic contingencies: On self-deception and the voter's illusion, *J. Personality Soc. Psych.* 46 (2) (1984) 237–248.
- [34] B. Sacerdote, Peer effects with random assignment: Results for Dartmouth roommates, *Quart. J. Econ.* 116 (2) (2001) 681–704.
- [35] R. Strotz, Myopia and inconsistency in dynamic utility maximization, *Rev. Econ. Stud.* 23 (1956) 165–180.

[Home](#)[Search](#)[Journals](#)[Books](#)[Abstract Databases](#)[My Profile](#)[Alerts](#)[? Help](#)Quick Search:

within

[? Search Tips](#)[Journal of Economic Theory](#)[Volume 123, Issue 2](#) , August 2005, Pages 105-134[doi:10.1016/j.jet.2005.04.001](https://doi.org/10.1016/j.jet.2005.04.001) [? Cite or Link Using DOI](#)

Copyright © 2005 Elsevier Inc. All rights reserved.

Self-control in peer groups

Marco Battaglini^{a, b}, , Roland Bénabou^{a, b, c}, , ,  and Jean Tirole^{d, e, f}, ^aDepartment of Economics, Princeton University, Princeton, NJ 08544 1013, USA^bCentre for Economic Policy Research, 90-98 Goswell Road, London EC1V 7RR, UK^cNational Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA^dInstitut d'Economie Industrielle, Manufacture des Tabacs, 21 allees de Brienne, 31000 Toulouse, France^eParis Sciences Economiques, 48 boulevard Jourdan, 75014 Paris, France^fDepartment of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02138, USA

Received 6 April 2005. Available online 18 June 2005.

Abstract

Social influences on self-control underlie both self-help groups and many peer interactions among youths. To understand these phenomena, we analyze how observing each other's behavior affects individuals' ability to deal with their own impulses. These endogenous informational spillovers lead to either a unique "good news" equilibrium that ameliorates behavior, a unique "bad news equilibrium" that worsens it, or to the coexistence of both. A welfare analysis shows that people will find social interactions valuable only when they have enough confidence in their own and others' ability to resist temptation. The ideal partner, however, is someone with a slightly worse self-control problem than one's own: this makes his successes more encouraging, and his failures less discouraging.

Keywords: Peer effects; Social interactions; Clubs; Self-control; Willpower; Addiction; Time-inconsistency; Memory; Psychology

This Document

- [SummaryPlus](#)
- ▶ **Full Text + Links**
 - [Full Size Images](#)
- [PDF \(355 K\)](#)

External Links

- 

Actions

- [Cited By](#)
- [Save as Citation Alert](#)
- [E-mail Article](#)
- [Export Citation](#)

JEL classification: C72; D82; D71; D91; J24

Article Outline

1. [Introduction](#)
 - 1.1. [Related literature](#)
 2. [The model](#)
 - 2.1. [Willpower and self-reputation](#)
 - 2.1.1. [Correlation in self-control problems](#)
 3. [Homogeneous peer groups](#)
 - 3.1. [Main intuitions](#)
 - 3.1.1. [The single-agent benchmark](#)
 - 3.1.2. [Two agents](#)
 - 3.1.3. [Increasing the correlation](#)
 - 3.2. [Equilibrium group behavior](#)
 - 3.2.1. [The informativeness constraints](#)
 - 3.2.2. [The incentive constraint](#)
 - 3.3. [Welfare analysis](#)
 - 3.3.1. [Welfare in a Good News equilibrium](#)
 - 3.3.2. [Welfare in a Bad News equilibrium](#)
 - 3.3.3. [The value of joining a group](#)
 4. [Heterogeneous peer groups](#)
 - 4.1. [Equilibrium behavior](#)
 - 4.1.1. [Informativeness constraints](#)
 - 4.1.2. [Profitability constraints](#)
 - 4.2. [Comparative statics and welfare analysis](#)
 5. [Conclusion](#)
- [Acknowledgements](#)
- Appendix A. [Appendix](#)
- A.1. [Welfare in an Intermediate equilibrium](#)
- Appendix B. [Supplementary data](#)
- B.1. [Good news equilibrium](#)
- B.2. [Mixed equilibrium](#)
- B.3. [Bad news equilibrium](#)
- [References](#)
-

1. Introduction

The behavioral and economic implications of imperfect self-control by a single decision maker have been the

focus of much recent work. Yet, people are typically immersed in social relations that exert powerful influences on their decisions. Peers and role models, for instance, play a critical part in young people's choices—particularly those that are subject to episodes of temptation like drinking, smoking, drug use, sexual activity, procrastination of effort, etc. In such settings peers may be good or bad “influences,” and the latter scenario is typically correlated with low or fragile self-esteem. At the same time, people with self-control or addiction problems often seek relief in self-help groups like Alcoholics Anonymous, Narcotics Anonymous and similar organizations that are predicated on the mutual sharing of experiences.

Psychologists and sociologists (not to mention parents) thus generally view the issues of self-control and peer effects as complementary. In economics, by contrast, they have so far been treated as largely separate areas of inquiry. In this paper we bring them together, studying how exposure to each other's behavior affects the ability of time-inconsistent individuals to deal with their own impulses.

Support groups, for instance, are an important social phenomenon. Organizations such as Alcoholics Anonymous, Narcotics Anonymous, Gamblers Anonymous, Debtors Anonymous and the like have branches in many countries, and millions of members. Economists are used to thinking about how entering contracts or binding implicit agreements with others allows agents to achieve desirable commitment. This, however, is not at all what self-help groups are about. Among the 14 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 clear that one cannot view these groups as standard commitment devices: they not only cannot, but do not even want to “control” their members. Their scope is in fact explicitly limited to fostering informational interaction (discussion) among members. Thus in “What does Alcoholics Anonymous do?” it is clearly stated that “A.A. members *share their experience* with anyone seeking help with a drinking problem” (emphasis added).

One therefore needs a theory to explain how (and when) observing the behavior of others can sometimes be beneficial for overcoming self-control problems, as with support groups, and sometimes highly detrimental, as often happens among schoolmates or neighborhood youths. Such a theory of peer effects in self-control should also be normative as well as positive. While group membership is sometimes exogenous (e.g., in public schools), it often involves of a voluntary choice, whether by the agent himself or by a “principal” invested with authority (judge ordering an addict to attend a 12-step program, parent trying to affect their child's selection of peers).

In this paper, we take the first steps towards such a theory, by developing a 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.²

There are two fundamental assumptions in our model. First, agents have incomplete information about their ability to resist temptation and try to infer it from their 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 heavily emphasized in the psychology literature [1], [2] and [6]. We build here on Bénabou and Tirole's [10] formalization of these phenomena, which is based on the idea that imperfect self-knowledge gives rise to a concern for *self-reputation*. By breaking a personal rule (abstinence resolution, diet, exercise regimen, moral principle) an individual would reveal himself, in his own eyes, as weak-willed with respect to such temptations, and this reputational loss would further undermine his resolve in the future. The fear of creating precedents thus creates an incentive to maintain a clean “track record” in order to influence one's future (selves') morale and behavior 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 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 actions of people similar to oneself is a source of additional information about the manageability or severity of the self-control problem—or, equivalently, the effectiveness of a particular method designed to alleviate it.⁴The information may turn out to be good news, if the others are observed to persevere (stay “dry”, “clean,” remain in school, etc.), or bad news, if they are observed to cave in or have a relapse. When deciding whether to exercise costly self-restraint in the face of temptation, an individual will take into account the likelihood of each type of news, and how it would impact the reputational “return” on his own behavior. Therefore, 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, since it depends on their equilibrium strategies. As a result, our model, in which *peer effects are purely informational*, can give rise to amplification effects as well as multiple equilibria, where agents' choices of self-restraint or self-indulgence are mutually reinforcing.

In the first part of the paper we focus on a symmetric situation where individuals are ex ante identical in all respects. Three main results are obtained. First, 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 membership improves self-discipline, a “bad news” equilibrium where it damages it, or to both. Second, social 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.

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 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 trade-off between the *likelihood* that someone else’s behavior will be a source of encouraging or discouraging news, and the *informativeness* of this news.

The dynamics of self-confidence play a key role in our theory of peer effects. First, self-restraint by one member (e.g., abstinence) improves both his and others’ self-confidence, and this in turn leads to more self-restraint by all in the future; misbehavior elicits the opposite feedbacks. Second, individuals will find self-help groups worth joining and remaining in only if they have sufficient confidence in their own and their peers’ ability not to relapse. While there is no systematic literature on the subject, field studies of self-help groups consistently document correlation patterns that are in line with these results (but of course do not constitute formal tests). For instance, Christo and Sutton’s [19] study of 200 Narcotics Anonymous members leads them 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.”

1.1. 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 influence of group characteristics on individual youths’ behavior, whether in terms of academic achievement, school truancy, smoking, drinking and drug use, teen pregnancy, employment, criminal activity and the like [18], [23], [25], [27], [28] and [34].

Econometric studies are 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 they imply for the equilibrium and optimal composition of groups. Thus, De Bartolome [20] and Bénabou [8] study how peer or neighborhood effects shape the functioning of a city and its schools; Bernheim [11] examines how a concern for others’ views of oneself leads to conformity; Brock and Durlauf [14] and Glaeser and Scheinkman [24] study how non-market interactions can lead to “social multipliers” and multiple equilibria. The only previous work seeking to endogenize peer effects is Banerjee and Besley’s [3] model of student testing, where a benchmarking effect

arising from the unknown difficulty of the test creates an informational complementarity between classmates' effort decisions.⁵

The other literature to which our paper relates is that on self-control problems, due for instance to non-exponential discounting (e.g., [29], [32] and [35]). In particular, a recent line of research has shown how the combination of self-control and informational concerns can account for many forms of “motivated cognitions” documented by psychologists. Carrillo and Mariotti [17] establish that time-inconsistent individuals may have, ex ante, a negative value for information. Bénabou and Tirole [9] develop a theory of rational self-deception through selective recall, and in [10] link personal rules to endogenous concerns for self-reputation. A related line of work by Bodner and Prelec [12] examines self-signaling in a split-self (ego-superego) model where the individual has “metapreferences” over his own tastes. Finally, our concern with interactions among time-inconsistent agents is shared with Brocas and Carrillo [13], who analyze how competition in the form of “patent races” can improve, and cooperation in joint projects worsen, individuals' tendency to procrastinate. In our model, by contrast, *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

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 Appendix A.

2. The model

2.1. 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 [10]. The canonical example is that of an alcoholic who must decide every morning whether to try and abstain that day, or just start drinking right away. If he was sure of his ability to resist throughout the afternoon and evening, when cravings 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), $t=1,2$, each of which is further divided into two subperiods, I and II (e.g., morning and afternoon), see Fig. 1. At the start of each subperiod I, the individual chooses between:



Fig. 1. Decisions and payoffs in any given period $t=1,2$. The parameter $\tilde{\beta}$ measures the salience of the present: for the current self $\tilde{\beta} = \beta < 1$, while for the ex ante self $\tilde{\beta} = 1$.

(1) A “no willpower” activity (NW), which yields 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.⁶

(2) 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 the intensity of temptation that he then experiences, the individual may opt, at the beginning of subperiod II, to either *persevere* until completion (P), or *give up* along the way (G).

Evaluated from an *ex ante* point of view (that of the agent's date-zero “self”), these different courses of action result in the following payoffs. Perseverance entails a “craving” cost $c > 0$ during subperiod II, but yields delayed gratification in the form of future benefits (better health, higher consumption etc.) whose present value, starting at the end of period t , is B . As explained below, c takes values c_L or c_H for different individuals, and is only imperfectly known by the agent himself. Caving in, on the other hand, results in a painless subperiod II but yields only a delayed payoff b , where $a < b < B$. The assumption that $b > a$ means that *some* self-restraint (resisting for a while but eventually giving up) is better than none at all. 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 that may cause them to succumb to short-run impulses at the expense of their long-run interests. We thus assume that, in addition to a standard discount rate δ between periods 1 and 2, their time preferences exhibit the usual quasi-hyperbolic profile: at any decision node, the individual overestimates the gratification from an immediate payoff by a factor of $1/\beta$, or correspondingly discounts all future payoffs at a rate $\beta < 1$.

The second key assumption is that the intensity of the cravings to which an individual will be subject if he attempts self-restraint 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 introspection or memory search. As a result, the agent in period 2 will have to try and *infer* his vulnerability to temptation from his own actions (“how did I behave last night, and what kind of a person does that make me?”) and those of his peers. We discuss this assumption of imperfect self-knowledge in more detail below. First, we state it formally and show how it combines with imperfect willpower (hyperbolic preferences) to generate a self-reputational “stake” in good behavior.

We assume that agents know their general degree of present-orientation, and for simplicity we take it to be the same $\beta < 1$ for everyone. By contrast, the activity-specific cost c differs across individuals, taking values $c = c_L$ or $c = c_H$, with $c_L < c_H$. A low-cost individual will also be referred to as a “strong type”, a high-cost as a “weak type”, where “strength” is here the ability to deal with the temptation of G . 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 .

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 his short-run impulses:⁷

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

where the maximum reputational “stake” $b - a > 0$ reflects the fact that even partial self-restraint (choosing *W*, then later on defaulting to *G*) is better than none (choosing *NW* at the outset).

Turning now to the agent's choice at the start of period 2, he will clearly only embark on a course of self-restraint when he has sufficient confidence in his ability to “follow through”. Since reputational concerns no longer operate, the expected return from attempting *W* exceeds the immediate (and more salient) payoff from *NW* only if his 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 of $\rho'(B - c_L) + (1 - \rho')b$ and a . A higher level of confidence ρ' in one's ability to resist temptation is then a valuable asset, because it helps offset the natural tendency to “give up without trying”. In particular, the fact that $\beta < 1$ creates an incentive for the weak type to *pool* with the strong one by persevering in the first period, so as to induce at least 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, cravings correspond to “hot,” internal, affective states, which are hard to remember later on from “cold” introspection. This intuitive idea is confirmed by experimental and field evidence on people's recollections of pain or discomfort [26] and their (mis)predictions of how they will behave under conditions of hunger, exhaustion, drug or alcohol craving, or sexual arousal [30] and [31]. Second, an individual 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 ample evidence that people's recollections are generally *self-serving*: they tend to remember (be consciously aware of) their successes more than their failures and find ways of absolving themselves of bad outcomes by attributing responsibility to others.⁸ Given imperfect or 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., [7]). It is also supported by experimental evidence, such as Quattrone and Tversky's [33] findings that people take actions, including painful ones, for self-signalling purposes.⁹ Such behaviors, one should again note, are conceivable only if later on the true motives and feelings behind one's earlier actions can no longer be reliably recalled or accessed.

2.1.1. 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—whether exogenously or by choice—in a social relationship where he can observe the behavior of others. What makes such exposure 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, so that by observing *B*'s actions, *A* can learn something about himself. If *B* successfully resists temptation this news are encouraging to *A*, while if *B* caves in or has a relapse they are discouraging.

We assume that for each agent, the prior probability of being a low cost type is ρ . Moreover, types are correlated, with conditional probabilities:

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

where α is a parameter measuring correlation.¹⁰

For $\alpha=0$ we get back the single-agent case (types are independent), while for $\alpha=1$ correlation becomes perfect. This simple structure also has the advantage that changes in α leave the unconditional probabilities unchanged, and vice-versa. This will allow comparative statics that cleanly separate the effects of initial reputation and of correlation. Finally, we have assumed a completely symmetric situation; in particular, the two agents enter the game with the same level of self-confidence ρ , their preference structure is the same, and this is common knowledge. In this case there are only symmetric equilibria, as shown later on. In [Section 4](#) we shall extend the analysis to asymmetric initial conditions, payoffs, and equilibria.

3. Homogeneous peer groups

3.1. Main intuitions

3.1.1. The single-agent benchmark

We begin with the one-agent case, which provides a natural starting point to understand group interactions and evaluate their welfare implications. Given that the strong type always perseveres, the 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 by the dashed middle line $x_a(\rho)$ in Fig. 2; the subscript a stands for “alone”.¹¹ Complete self-restraint (perfect pooling) by Self 1 makes observing P completely uninformative for Self 2, leaving his prior unchanged; it is therefore an equilibrium only when the agent's initial 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).¹² Conversely, 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*, $\Pr_{x,\rho}(c = c_L | P) = \rho^*$, 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 $\rho = \rho^*$.

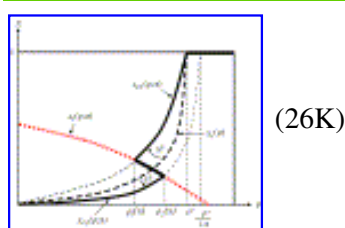


Fig. 2. Equilibrium self-restraint for a single individual (dashed line) and in a peer group (solid lines).

3.1.2. Two agents

Let us now bring together two individuals whose types are correlated as described above and examine how this affects the behavior of weak types at the temptation stage. As mentioned earlier, we focus until Section 4 on equilibria 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 that the correlation α between types is relatively low. By continuity, equilibrium behavior will not be too different from that of the single agent case; the interesting issue is 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 in Fig. 2.

- (a) When initial reputation is low, j is most probably also a weak type, who will play a strategy close to $x_a(\rho) \approx 0$. Consequently, he is almost sure to be a source of “bad news” (G) that will reduce i 's hard-earned reputational gain from playing P . This *discouragement* effect naturally leads agent i to persevere with lower probability $x_{PG}(\rho; \alpha) < x_a(\rho)$, as indicated by the solid curve emanating from the origin. Intuitively, i must now 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.

(b) When initial reputation is high (just below ρ^*), j is now either a strong type or a weak type who exerts self-control with probability close to $x_a(\rho) \approx 1$. Therefore, 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, the weak type's probability of playing P increases to $x_{PP}(\rho; \alpha) > x_a(\rho)$, as illustrated by the solid curve that rises up to $(\rho^*, 1)$. In this case, the positive externality allows the agent to engage in more pooling.

(c) Where ρ is in some intermediate range, finally, if i plays P both PG and PP have non-negligible probability, and which one ends up shaping equilibrium strategies 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 . Due to this strategic complementarity (which operates purely through joint informational spillovers on the decision of Self 2), both $x_{PP}(\rho; \alpha)$ and $x_{PG}(\rho; \alpha)$ are equilibria over some range of ρ ; see [Fig. 2](#). As usual, a third equilibrium $x_f(\rho; \alpha)$ then also exists in-between; it will be described in more detail below.

3.1.3. Increasing the correlation

As α increases the x_{PG} locus pivots down, while the x_{PP} locus pivots up (see [Fig. 2](#)): 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.

3.2. Equilibrium group behavior

3.2.1. The informativeness constraints

Let $\bullet_{PG}(x; \rho, \alpha)$ 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 $\bullet_{PP}(x; \rho, \alpha)$ be the posterior following a play of P by both agents. Since strong types always play P , we have $\bullet_{PG} < \bullet_{PP}$ for all $\rho > 0$. It is also easy to see that, in any equilibrium:

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

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 $\rho > \rho^*$.¹³ 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. Eq. (4) thus defines two upward-sloping loci in the (ρ, x) plane, between which any equilibrium must lie:

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

where

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

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

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 in Fig. 2, x_{PP} increases with ρ up to $\rho = \rho^*$, after which it equals 1.

Along the increasing part, we have $\bullet_{PP} = \rho^*$: 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 $\bullet_{PP} \geq \rho^*$ 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, $\bullet_{PG} = \rho^*$: 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 this pivotal state, $\bullet_{PP} = \rho^*$.

(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, $\bullet_{PG} = \rho^*$, that is relevant.

(iii) *Intermediate 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 .

3.2.2. The incentive constraint

We now determine exactly when each scenario applies. In order for the weak type to be willing to mix between P and 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 use strategy x and expects his own Self 2 to choose W with probabilities y and y' following events PP and PG respectively. Since a weak type will reap payoff b under W rather than a under NW , we have

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

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

$$\pi(x; \rho, \alpha) \equiv B - b - \frac{c_H}{\beta} + \delta[(1 - \alpha)\rho + (1 - (1 - \alpha)\rho)x](b - a) = 0. \tag{9}$$

This equation uniquely defines a downward-sloping locus $x_I(\rho; \alpha)$ in the (x, ρ) plane, which we shall refer to 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 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 and 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} curve 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 $\pi(x; \rho, \alpha) = \Pi(x, 1, 0; \rho, \alpha) \geq 0$.
- Intermediate equilibria correspond precisely to the portion of the incentive locus x_I that 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 in Fig. 2. Formally:

Proposition 1

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

- (i) For $\rho < \rho_1(\alpha)$ there is a unique equilibrium, which is of the “bad news” type: $x = x_{PG}(\rho; \alpha)$.
- (ii) For $\rho > \rho_2(\alpha)$ there is a unique equilibrium, which is of the “good news” type: $x = x_{PP}(\rho; \alpha)$.
- (iii) 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: $\lim_{\alpha \rightarrow 0} |\rho_2(\alpha) - \rho_1(\alpha)| = 0$.

Fig. 3 provides a convenient representation of these results in the (ρ, α) space.¹⁴ As correlation declines to zero, the area between $\rho_1(\alpha)$ and $\rho_2(\alpha)$ where multiplicity occurs 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.

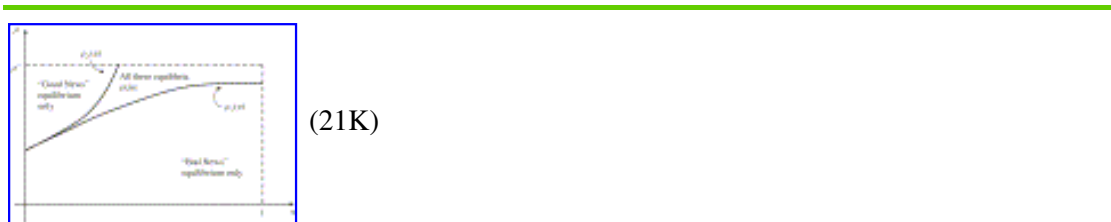


Fig. 3. Equilibrium outcomes for different levels of self-confidence and correlation.

Note that the *PG* equilibrium exists only when ρ is not too high or the degree of correlation α is large enough, while the reverse conditions are needed to sustain the *PP* equilibrium. Indeed, the first case requires a weak agent to be relatively pessimistic about his partner's type (hence about the latter's likelihood of choosing *P*), while in the second he must be sufficiently optimistic.

3.3. Welfare analysis

We shall now 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 make clear when groups do indeed provide valuable “help,” and when they actually do damage.

The question of what welfare function to use in a model where preferences change over time is a controversial one, and in our model with imperfect self-knowledge it could, a priori, be even more complicated. The results we obtain, however, are fully consistent across the different possible criteria. To understand why, consider first an agent's initial decision of whether or not to join a group. At this stage he does not yet know his type, and his temporal preferences are not yet subject to present bias. He thus makes his decision by computing the undistorted intertemporal payoffs W^s and W^w that he will reap if he turns out to be strong or weak, then examining whether the expectation $W = \rho W^s + (1-\rho)W^w$ is higher in isolation or in a group. We shall thus be interested in ex ante welfare W from a positive as well as a normative point of view. The (undistorted) *interim* utility levels of each type, W^s and W^w , are essential components of this criterion; furthermore, in our model they also completely determine the value of *any* social welfare criterion that puts weight on *both* ex ante and ex post preferences.¹⁵ This is because: (i) the strong type always perseveres, so his ex post welfare (evaluated at the time of temptation) just differs from W^s by a constant: $W^{s,\beta} = W^s - c_H(1/\beta - 1)$; (ii) the weak type always randomizes between *P* and *G*, so his welfare is the same as if he systematically chose *G*: $W^{w,\beta} = b + \delta a$.¹⁶ Consequently, any proposition established for W^s (respectively, W^w) immediately carries over to all (linear or nonlinear) aggregates of W^s and $W^{s,\beta}$ (respectively, W^w and $W^{w,\beta}$), i.e. to all type-specific welfare criteria. Similarly, any result on ex ante welfare $W = \rho W^s + (1-\rho)W^w$ carries over to any weighted average of the non-tempted and tempted selves, W and $W^\beta = \rho W^{s,\beta} + (1-\rho)W^{w,\beta}$. Finally, and perhaps most importantly, interventions that (say) raise interim welfare for both types, W^s and W^w , represent true *Pareto improvements* in any sense of the word: they make the agent better off whether his payoffs are evaluated with or without present bias, and with or without information about his type.¹⁷

We start with the natural benchmark case where the agent is alone (equivalently, $\alpha=0$). For the weak type,

$$W_a^w = b + \delta a + x_a [B - b - c_H + \delta y_a (b - a)], \quad (10)$$

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*. Next, for the weak type to be indifferent at the temptation stage, it must be that

$$B - c_H / \beta + \delta [y_a b + (1 - y_a) a] = b + \delta a. \tag{11}$$

Substituting into (10) yields:

$$W_a^w = b + \delta a + x_a \left(\frac{1 - \beta}{\beta} \right) c_H, \tag{12}$$

in which the second term reflects the *value of the self-discipline* achieved through the reputational mechanism. Turning now to expected welfare for a strong type, we can write:

$$W_a^s = B - c_L + \delta [y_a (B - c_L) + (1 - y_a) a]. \tag{13}$$

Because $y_a < 1$ for all $\rho < \rho^*$, the strong type's average payoff is always less than $B - c_L$, which is what he would achieve under perfect information, or in a one-shot context. 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 interactive cases discussed above: welfare in the Good News equilibrium and in the Bad News equilibrium. Since the analysis of the Intermediate equilibrium is technically very similar, it is presented in [Appendix A](#). Readers who would like to skip the derivation of all the welfare results may go directly to [Section 3.3.3](#), which summarizes the main insights.

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

$$W_{PP}^w = b + \delta a + x_{PP} [B - b - c_H + \delta \Pr_{PP}(P | w) y_{PP} (b - a)], \tag{14}$$

where $\Pr_{PP}(P | w) = 1 - \pi_{LL} + \pi_{LL} x_{PP}$ 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:

$$W_{PP}^w = W_a^w + (x_{PP} - x_a) \left(\frac{1 - \beta}{\beta} \right) c_H. \tag{15}$$

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:

$$W_{PP}^s = B - c_L + \delta a + \delta \Pr_{PP}(P | s) y_{PP} (B - b - c_L),$$

where $\Pr_{PP}(P | s) = \pi_{LL} + (1 - \pi_{LL}) x_{PP}$ is the equilibrium probability that j will choose P , given that i is a strong type. Next, subtract (10) 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 $y_a = y_{PP} \Pr_{PP}(P | w)$. Thus

$$W_{PP}^s = W_a^s + \delta y_{PP} [\Pr_{PP}(P | s) - \Pr_{PP}(P | w)] (B - a - c_L). \tag{16}$$

Thus, as long as $\alpha > 0$, the strong type is also strictly better off: $W_{PP}^s > W_s^a$. The intuition is that with two agents the payoff to i 's playing P becomes contingent on what 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 in the single-agent case. 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 before: $y_{PP} > y_a$. This yields no extra surplus for the weak type, who remains indifferent, but generates rents for the strong type.

Proposition 2

In the Good News equilibrium that exists for all (ρ, α) with $\rho > \rho_1(\alpha)$, joining a group is strictly better than staying alone from an interim point of view (i.e., for both types), and therefore also ex ante. The same remains true according to any social welfare criterion that puts positive weight on ex post as well as ex ante preferences.

The result that joining a group can bring about a *Pareto improvement*, rather than just transfer surplus across types or temporal selves, is somewhat surprising, since the presence of peers entails a trade-off between the positive informational spillover received when they persevere and the negative one suffered when they do 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. Thus 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.

3.3.2. Welfare in a Bad News equilibrium

Derivations similar to the previous case yield for the weak type:

$$W_{PG}^w = W_a^w + (x_{PG} - x_a) \left(\frac{1 - \beta}{\beta} \right) c_H. \tag{17}$$

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:

$$W_{PG}^s = W_a^s + \delta [\Pr_{PG}(P | s) - \Pr_{PG}(P | w)] (1 - y_{PG}) (B - a - c_L). \tag{18}$$

This makes clear 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.

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. The same remains true when each type's welfare is evaluated according to any welfare criterion that also puts positive weight on his ex post preferences.

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 forming a group.

Proposition 4

Assume that agents expect a Bad News equilibrium. There exist two values $0 < \underline{\rho}^ < \bar{\rho}^* < 1$ such that for all $\rho^* \in (\underline{\rho}^*, \bar{\rho}^*)$, 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^* .¹⁸*

3.3.3. The value of joining a group

We now briefly summarize the main results obtained so far. 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 (W is higher) but also at the interim stage (W^s and W^w are higher) as well as ex post ($W^{h,\beta}$ is higher, $W^{w,\beta}$ is unchanged). 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. From an ex ante point of view, forming a group is then beneficial only if self-confidence exceeds a minimal threshold.

Field studies of self-help groups for alcohol and drug abusers consistently find a strong positive correlation between self-esteem and “clean time” in the group [21] and [19]. The standard interpretation is that interactions with peers help individuals sustain desirable behavior, which in turn raises their self-esteem. This would be in line with our results concerning the Good News equilibrium, which is sustained by the collective building up and maintenance of self-confidence. Alternatively, the observed correlation could reflect self-selection, with low self-esteem individuals dropping out earlier. This second (non-exclusive) explanation is also consistent with our predictions: agents with very low self-confidence are always those who benefit least from group interactions, and may even prefer isolation (Bad News equilibrium).

4. Heterogeneous peer groups

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. The second and key 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?

4.1. 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 $\rho^i \equiv \Pr(c^i = c_L)$, and the conditional probabilities as $\pi_{LL}^i \equiv \Pr(c^i = c_L | c^j = c_L)$ and $\pi_{HH}^i \equiv \Pr(c^i = c_H | c^j = c_H)$, for $i=1,2$.¹⁹ We only impose a general condition of positive correlation between agents’ craving costs (monotone likelihood ratio property):

$$\frac{\Pr((c_H, c_L))}{\Pr((c_L, c_L))} < \frac{\Pr((c_H, c_H))}{\Pr((c_L, c_H))}. \quad (19)$$

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

4.1.1. Informativeness constraints

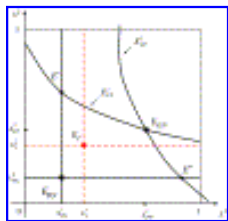
Let $\mu_{PP}^i(x^i, x^j)$ and $\mu_{PG}^i(x^i)$ denote individual i 's posteriors about his own type when both agents persevered in the previous period, and when he persevered but the other agent did not.²⁰The same simple reasoning as in [Section 3.2](#) shows that, in any equilibrium, these beliefs must satisfy:

$$\mu_{PG}^i(x^i) \leq \rho^{i*} \leq \mu_{PP}^i(x^i, x^j). \tag{20}$$

As shown in the appendix and illustrated in [Fig. 4](#), each equation $\mu_{PP}^i(x^i, x^j) = \rho^{i*}$ uniquely defines a downward-sloping function $x^i = X_{PP}^i(x^j)$, with $(X_{PP}^1)^{-1}$ steeper than X_{PP}^2 . As long as the two agents are not excessively different from one another, there is then a unique intersection

$E_{GN} = (x_{PP}^1, x_{PP}^2) \in (0, 1) \times (0, 1)$, where both (weak) agents play their “good news” strategies.²¹Similarly, each equation $\mu_{PG}^i(x^i) = \rho^{i*}$ has a unique solution $x^i = x_{PG}^i$, which corresponds in [Fig. 4](#) to a straight horizontal or vertical line. At the intersection $E_{BN} = (x_{PG}^1, x_{PG}^2)$, both (weak) agents play their “bad news” strategies. Quite intuitively, each of these lines lies closer to the origin than the corresponding X_{PP}^i curve, so that together the four constraints in [\(20\)](#) define a “permissible region” $E_{BN}E'_{GN}E''$ within which any equilibrium must lie:

$$x_{PG}^i \leq x^i \leq X_{PP}^i(x^j). \tag{21}$$



(28K)

Fig. 4. Good News, Bad News and intermediate equilibria.

4.1.2. Profitability constraints

Let $\Pi^i(x^j, y_{PP}^i, y_{PG}^i)$ 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_{PP}^i and y_{PG}^i following the events PP and PG , respectively. Let $\pi^i(x^j) \equiv \Pi^i(x^j, 1, 0)$, and denote as x_I^j the solution (in \mathbb{R}) to the linear equation $\pi^i(x^j)=0$.

Clearly, in any equilibrium it must be that $\Pi^i(x^j, y_{PP}^i, y_{PG}^i) \geq 0$, 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

$$\begin{cases} \text{if } \rho^{i*} < \mu_{PP}^i(x^i, x^j) \text{ then } \pi^i(x^j) \leq 0, \\ \text{if } \rho^{i*} > \mu_{PG}^i(x^i) \text{ then } \pi^i(x^j) \geq 0, \end{cases} \tag{22}$$

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

$$\begin{cases} \text{if } x^j > x_I^j \text{ then } x^i = X_{PP}^i(x^j); \\ \text{if } x^j < x_I^j \text{ then } x^i = x_{PG}^i. \end{cases} \tag{23}$$

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 [\(23\)](#), 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 $E_M \equiv ((X_{PP}^2)^{-1}(x_I^2), x_I^2)$ when it lies on the upper boundary of that region, as on the left panel of [Fig. 5](#); (iii) the point $E_M \equiv (x_I^1, (X_{PP}^1)^{-1}(x_I^1))$ when it lies on the right boundary of that same region, as on the right panel of [Fig. 5](#).

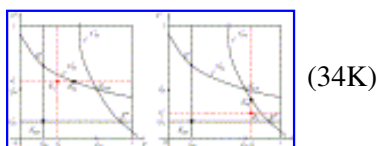


Fig. 5. Mixed, Intermediate, and Bad News equilibria.

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 shall focus here on the case where all three possible types of equilibria coexist, so that we can analyze the comparative statics of each one. The complete analysis of the other possible cases is presented in [\[5\]](#) as well as in [Appendix B](#), which is available through the on-line edition of this journal.

It is easily seen from [\(23\)](#) that a necessary and sufficient condition for such multiplicity is that the point E_I lie in the permissible region of [Figs. 4](#) and [5](#), that is,

$$x_{PG}^i < x_I^i < X_{PP}^i(x_I^i), \quad \text{for } i = 1, 2. \tag{24}$$

Proposition 5

Let condition (24) hold. 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 $x_I^2 > x_{PP}^2$ then $S = \{E_{BN}, E_I, E_M \equiv ((X_{PP}^2)^{-1}(x_I^2), x_I^2)\}$.
- (iii) If $x_{PG}^2 < x_I^2 < x_{PP}^2$ but $x_I^1 > x_{PP}^1$ then $S = \{E_{BN}, E_I, E_M \equiv (x_I^1, (X_{PP}^1)^{-1}(x_I^1))\}$.

Thus, under condition (24) there is an equilibrium where both agents are in a “bad news” regime, another one where both are in an “intermediate” 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 “intermediate” 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 Fig. 5. In such a situation, the informativeness constraint $\pi^j=0$ is binding on one agent and the incentive constraint $\mu_{PP}^i(x^i, x^j) = \rho^{i*}$ 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 thus for their respective intersections, E_{GN} and E_{BN} .

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 neither one 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).

4.2. 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.²²

Rather than examine the local comparative statics of each equilibrium separately we shall integrate them into a more interesting *global analysis* that allows 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.²³The key results are illustrated on the right panel of Fig. 6.

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 believed to be a little weaker than him, that is, who has a ρ^{j*} somewhat above ρ^{i*} .*
- (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*}).*

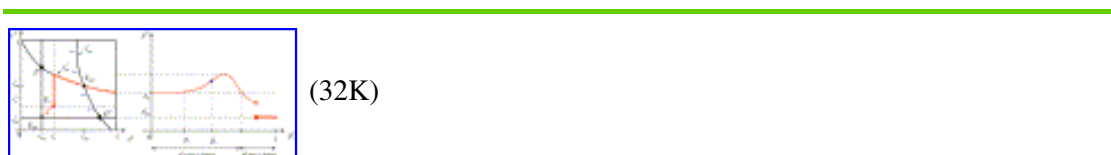


Fig. 6. The effect on agent 2 of the severity of his peer's potential self-control problem. The right panel depicts both agent 2's behavior x^2 (when weak) and his ex ante welfare $W^2 = \rho^2 W^{2,S} + (1 - \rho^2) W^{2,W}$. Indeed,

$W^{2,w}$ strictly increases with x^2 , while $W^{2,s}$ is always nondecreasing in ρ^{1*} .

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. [Fig. 6](#) shows that for relatively low values of 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.

[Proposition 6](#) can be derived by means of a simple graphical analysis. As ρ^{1*} increases from 0 to 1 the X_{PP}^1 locus shifts left, as indicated on the left panel of [Fig. 6](#); consequently, the high self-restraint equilibrium travels along the path marked by the thick arrows. The implied self-control behavior (and welfare) of agent 2 can then simply be read off the right panel of the figure. We omit here the complete proof for reason of space; it can be found in [\[5\]](#) as well as in [Appendix B](#), which is available through the on-line edition of this journal.

5. 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 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 a person'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 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 psychology literature. Nonetheless, it is still clearly oversimplistic, and could be extended in several directions. First, 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. Second, there are a number of important aspects of peer interaction from which we abstract. Some, like learning specific techniques to deal with impulses, are quite consistent with our approach and could easily be incorporated. Others, involving a desire to “belong”, being helped by the “moral support” of others, or basic emotional mechanisms such as embarrassment at having to admit failure in front of others and deriving pride from public success, would require more substantial extensions. Another interesting direction for further research would be to explore peer effects that involve *excessive*, rather than insufficient, self-regulation.²⁴ 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.

Acknowledgements

We are grateful for helpful comments to Jess Benhabib, Leonardo Felli, Ted O’Donoghue, John Morgan, Michele Piccione, Matthew Rabin, Tom Romer and two anonymous referees, as well as to seminar participants at the 2001 Congress of the European Economic Association, the Studienzentrum Gerzensee, the Harvard–MIT theory seminar, the London School of Economics, the University of Toronto, the Wallis Institute at Rochester University and the Stanford Institute for Theoretical Economics. Battaglini gratefully acknowledges the hospitality of the Economics Department at the Massachusetts Institute of Technology during the academic year 2002–2003. Bénabou gratefully acknowledges financial support from the National Science Foundation and the John Simon Guggenheim Foundation, and the hospitality of the Institute for Advanced Study during the academic year 2002–2003.

References

- [1] G. Ainslie, *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person* (Studies in Rationality and Social Change), Cambridge University Press, Cambridge, UK, New York (1992).
- [2] G. Ainslie, *Breakdown of Will*, Cambridge University Press, Cambridge, UK (2001).
- [3] A. Banerjee, T. Besley, Peer group externalities and learning incentives: A theory of nerd behavior, Princeton University Working Paper No. 68, 1990.
- [4] M. Battaglini and R. Bénabou, Trust, coordinations, and the industrial organization of political activism, *J. Europ. Econ. Assoc.* **1** (2003) (4), pp. 851–889. [Abstract-EconLit](#) | [Full Text via CrossRef](#)
- [5] M. Battaglini, R. Bénabou, J. Tirole, Self-Control in Peer Groups, CEPR Discussion Paper 3149, January 2002.

- [6] R. Baumeister, T. Heatherton and D. Tice, *Losing Control: How and Why People Fail at Self-Regulation*, Academic Press, San Diego, CA (1994).
- [7] D. Bem, Self-perception theory. In: L. Berkowitz, Editor, *Advances in Experimental Social Psychology*, Academic Press, New York, NY (1972).
- [8] R. Bénabou, Workings of a city: Location, education and production, *Quart. J. Econ.* **108** (1993), pp. 619–652.
- [9] R. Bénabou and J. Tirole, Self-confidence and personal motivation, *Quart. J. Econ.* **117** (2002) (3), pp. 871–915.
- [10] R. Bénabou and J. Tirole, Willpower and personal rules, *J. Polit. Econ.* **112** (2004) (4), pp. 848–887.
- [11] D. Bernheim, A theory of conformity, *J. Polit. Econ.* **102** (1994) (5), pp. 841–877.
- [12] R. Bodner and D. Prelec, Self-signaling and diagnostic utility in everyday decision making. In: I. Brocas and J. Carrillo, Editors, *Collected Essays in Psychology and Economics vol. I*, Oxford University Press, Oxford (2003).
- [13] I. Brocas and J. Carrillo, Rush and procrastination under hyperbolic discounting and interdependent activities, *J. Risk Uncertainty* **22** (2001) (2), pp. 141–144.
- [14] W. Brock and S. Durlauf, Discrete choice with social interactions, *Rev. Econ. Stud.* **68** (2001) (2), pp. 235–260. [Abstract-GEOBASE](#) | [Abstract-EconLit](#) | [MathSciNet](#)
- [15] A. Caplin, J. Leahy, The social discount rate, NBER Working Paper 7983, 2000.
- [16] A. Caplin and J. Leahy, The supply of information by a concerned expert, *Econ. J.* **114** (2004) (497), pp. 487–505. [Abstract-GEOBASE](#) | [Abstract-EconLit](#) | [Full Text via CrossRef](#)
- [17] J. Carrillo and T. Mariotti, Strategic ignorance as a self-disciplining device, *Rev. Econ. Stud.* **67** (2000) (3), pp. 529–544. [Abstract-EconLit](#)
- [18] A. Case, L. Katz, The company you keep: The effects of family and neighborhood on disadvantaged youth, NBER Working Paper 3705, 1991.
- [19] G. Christo and S. Sutton, Anxiety and self-esteem as a function of abstinence time among recovering addicts attending Narcotics Anonymous, *Brit. J. Clinical Psychol.* **33** (1994), pp. 198–200. [Abstract-EMBASE](#) | [Abstract-MEDLINE](#) | [Abstract-PsycINFO](#)
- [20] C. De Bartolome, Equilibrium and inefficiency in a community model with peer group effects, *J. Polit. Econ.* **98** (1990), pp. 10–133.

- [21] C.B. De Soto, W.E. O'Donnell, L.J. Allred and C.E. Lopes, Symptomatology in alcoholics at various stages of abstinence, *Alcoholism: Clinical Exper. Res.* **9** (1985), pp. 505–512. [Abstract-MEDLINE](#)
- [22] J. Elster, Introduction. In: J. Elster, Editor, *Addiction: Entries and Exits*, Russel Sage Foundation, New York (2001).
- [23] A. Gaviria and S. Raphael, School-based peer effects and Juvenile behavior, *Rev. Econ. Statist.* **83** (2001) (2), pp. 257–268. [Abstract-EconLit](#) | [Full Text via CrossRef](#)
- [24] E. Glaeser and J. Scheinkman, Non-market interactions. In: M. Dewatripont, L.P. Hansen and S. Turnovsky, Editors, *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress*, Cambridge University Press, Cambridge, MA (2002).
- [25] C. Hoxby, Peer effects in the classroom: Learning from gender and race variation, NBER W.P. No. 7867, 2001.
- [26] D. Kahneman, P. Wakker and R. Sarin, Back to Bentham? Explorations of experienced utility, *Quart. J. Econ.* **112** (1997) (2), pp. 375–405. [Abstract-EconLit](#) | [Full Text via CrossRef](#)
- [27] P. Kooreman, Time, money, peers and parents: Some data and theories on teenage behavior, IZA Discussion Paper No. 931, November 2003.
- [28] M. Kremer, D. Levy, Peer effects and alcohol use among college students, NBER W.P. No. 9876, July 2003.
- [29] D. Laibson, Golden eggs and hyperbolic discounting, *Quart. J. Econ.* **112** (1997), pp. 443–478.
- [30] G. Loewenstein, Out of control: Visceral influences in behavior, *Organ. Behav. Human Dec. Process.* **65** (1996) (3), pp. 272–292. [Abstract](#) | [PDF \(197 K\)](#)
- [31] G. Loewenstein and D. Schkade, Wouldn't it be nice? Predicting future feelings. In: D. Kahneman, E. Diener and N. Schwartz, Editors, *Well-Being: Foundations of Hedonic Psychology*, Russel Sage Foundation, New York, NY (1999).
- [32] T. O'Donoghue and M. Rabin, Doing it now or later, *Amer. Econ. Rev.* **89** (1999) (1), pp. 103–124. [Abstract-EconLit](#)
- [33] G. Quattrone and A. Tversky, Causal versus diagnostic contingencies: On self-deception and the voter's illusion, *J. Personality Soc. Psych.* **46** (1984) (2), pp. 237–248. [Abstract-PsycINFO](#)
- [34] B. Sacerdote, Peer effects with random assignment: Results for Dartmouth roommates, *Quart. J. Econ.* **116** (2001) (2), pp. 681–704. [Abstract-EconLit](#) | [Full Text via CrossRef](#)

[35] R. Strotz, Myopia and inconsistency in dynamic utility maximization, *Rev. Econ. Stud.* **23** (1956), pp. 165–180.

Appendix A.

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 $\mu_{PP}^i(x^1, x^2) = \rho^{i*}$ and $\mu_{PG}^i(x^i) = \rho^{i*}$, for $i=1,2$. The following lemma establishes these properties:

Lemma 1

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

(i) The loci $X_{PP}^i(x^i, x^j)$ are decreasing in x^j . Furthermore $X_{PP}^2(x^1)$ cuts $(X_{PP}^1)^{-1}(x^1)$ at most once in the positive orthant, and if it does the intersection is from below.

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

Proof

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

$$\frac{\mu_{PP}^i(x^i, x^j)}{1 - \mu_{PP}^i(x^i, x^j)} = \frac{\Pr(c^i = c_L, c^j = c_L) + \Pr(c^i = c_L, c^j = c_H)x^j}{\Pr(c^i = c_H, c^j = c_L)x^i + \Pr(c^i = c_H, c^j = c_H)x^i x^j}, \tag{25}$$

$$\frac{\mu_{PG}^i(x^i)}{1 - \mu_{PG}^i(x^i)} = \frac{\Pr(c^i = c_L, c^j = c_H)}{\Pr(c^i = c_H, c^j = c_H)x^i}. \tag{26}$$

Clearly, μ_{PP}^i and μ_{PG}^i are both decreasing in x^i . To see that μ_{PP}^i is decreasing in x^j as well, note that $\partial \mu_{PP}^i(x^i, x^j) / \partial x^j$ 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 [\(19\)](#). Therefore $\partial X_{PP}^i(x^i, x^j) / \partial x^j < 0$ by the implicit function theorem. Next, note that $X_{PP}^2(0)$ is bounded for $x^1 \in [0, 1]$. By contrast, we can easily verify that $\lim_{x^1 \rightarrow 0} (X_{PP}^1)^{-1}(x^1) = +\infty$. Therefore, there exists a point x^1 small enough that $X_{PP}^2(x^1) < (X_{PP}^1)^{-1}(x^1)$. 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 $X_{PP}^2(x^1)$ crossing $(X_{PP}^1)^{-1}(x^1)$ from below. Note first that any intersection must be such that $\mu_{PP}^1(x^1, x^2)/\mu_{PP}^2(x^1, x^2) = \rho^{1*}/\rho^{2*}$. By (25), this implies

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

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

(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 Intermediate 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 $x_{PP}(\rho; \alpha) \rightarrow x_{PG}(\rho; \alpha)$, it is immediate to see that $\lim_{\alpha \rightarrow 0} |\rho_2(\alpha) - \rho_1(\alpha)| = 0$. •

Proof of Propositions 2–4

The first two were established in the text; we prove here the third one. A Bad News equilibrium is ex ante preferable to staying alone when

$$E(W_{PG} - W_a | \rho) \equiv \rho(W_{PG}^s - W_a^s) + (1 - \rho)(W_{PG}^w - W_a^w) > 0. \tag{27}$$

From the informativeness constraint (25) 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 (17) and (18), we can then

rewrite (27) as:

$$\Psi(\rho, \rho^*) \equiv (\rho^* - 1)k(\rho) + \rho^* - (1 - \alpha)\rho < 0, \quad \text{where} \tag{28}$$

$$k(\rho) \equiv \frac{(1 - \beta)c_H}{\beta\delta(1 - y_{PG}(\rho))(B - c_L - a)}. \tag{29}$$

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

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

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 (27) 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 $\bar{\rho}^*$ by the linear equations $\Psi(0, \underline{\rho}^*) = (\underline{\rho}^* - 1)k(0) + \underline{\rho}^* \equiv 0$ and $\Psi(1, \bar{\rho}^*) = (\bar{\rho}^* - 1)k(1) + \bar{\rho}^* - (1 - \alpha)$ respectively. Then $0 < \underline{\rho}^* < \bar{\rho}^* < 1$ and for any ρ^* in $(\underline{\rho}^*, \bar{\rho}^*)$, $\hat{\rho}(\rho^*)$ lies in $(0, \rho^*)$ and is strictly increasing in ρ^* . For $\rho^* < \underline{\rho}^*$ we have $\hat{\rho}(\rho^*) = 0$, and $E(W_{PG} - W_a | \rho) > 0$ for all $\rho \geq 0$. Conversely, for $\rho^* > \bar{\rho}^*$ we have $\hat{\rho}(\rho^*) = \rho^*$, and $E(W_{PG} - W_a | \rho) < 0$ for all $\rho \leq \rho^*$.

A.1. Welfare in an Intermediate equilibrium

For the weak type, we have as usual $W_I^w = W_a^w + (x_I - x_a)[(1 - \beta)/\beta]c_H$. Recall from Fig. 2 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:

$$\begin{aligned} W_I^s &= W_a^s + \delta[\text{Pr}_I(P | s) - \text{Pr}_I(P | w)](B - a - c_L) \\ &= W_a^s + \delta\alpha(1 - x_I)(B - a - c_L). \end{aligned}$$

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 Intermediate 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 strong type is better off).

Proof of Proposition 5

We first prove condition (22).

(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 $\Pi^i(x^j, 1, y_{PG}^i) \geq \pi^i(x^j) > 0$ 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 $X_{PP}^j(x) - (X_{PP}^i)^{-1}(x)$ has the sign of $x_{PP}^i - x$ for all x (single-crossing property established by Lemma 1 and illustrated in Fig. 4), this implies that $x^j < (X_{PP}^i)^{-1}(1)$, or equivalently $\mu_{PP}^j(x^j, 1) > \rho^{j*}$. As a result, agent j 's second-period self will choose $y_{PP}^j = 1$, ensuring $\Pi^j(1, 1, y_{PG}^j) = \Pi^j(1, 1, 0) > 0$. This leads to $x^j = 1$, a contradiction.

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

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

Proof of Proposition 6

See Appendix B.

Appendix B. Supplementary data

We provide here the details of some proofs that were omitted from Section 4 due to space constraints.

Complement to Proposition 5 Case of a unique equilibrium

When condition (24) holds, the intersection E_I of the two x_I^i loci lies inside the permissible region $E_{BN}E'E_{GN}E''$ of Fig. 4 or Fig. 5. In Fig. 7 this area is itself decomposed into areas I, IIa and IIb, which, respectively, correspond to cases (i), (ii) and (iii) of Proposition 5. When condition (24) does not hold, $E_I = (x_I^1, x_I^2)$ lies in one of the “outer areas” of Fig. 7. Using (23) and the discussion that 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' when E_I falls in IVb; $E_M = ((X_{PP}^2)^{-1}(x_I^2), x_I^2)$ when E_I falls in IVa; E_{GN} when E_I falls in III; $E_M = (x_I^1, (X_{PP}^1)^{-1}(x_I^1))$ when

E_I falls in Va; E'' when E_I falls in Vb; and E_{BN} when E_I falls in IIc.

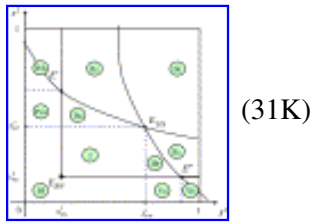


Fig. 7. Equilibrium set in the general (asymmetric) model.

Proof of Proposition 6

We derive here the path of the equilibrium with the highest level of self-control as ρ^{*1} rises (left panel of Fig. 6), and the corresponding ex ante welfare level achieved by the agent (right panel).

Recall that when $\rho^{*1} < \rho^1$ agent 1 can always achieve complete self control on his own ($x^1=1$). In this case agent 2 learns nothing from observing his peer's behavior; hence $x^2 = x_a^2$, as when there is no group: $W=W^a$. We now consider values of ρ^{*1} above ρ^1 .

B.1. Good news equilibrium

For a relatively low value of $\rho^{*1} > \rho^1$ we are in a configuration like that of Fig. 4, with E_{GN} constituting a Good News equilibrium, located at the intersection of the two informativeness constraints [Click to view the](#) and $\mu_{PP}^2(x^1, x^2) = \rho^{2*}$. As indicated by the arrows in Fig. 6, an increase in ρ^{*1} causes the locus X_{PP}^1 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_{PP}^2 , by contrast, is unchanged. As a result, the equilibrium E_{GN} travels left and up along the X_{PP}^2 locus: x^1 decreases, but x^2 increases. As a result, agent 2 is actually *better off* from a (marginal) *worsening* in the severity of his peer's self-control problem. For a weak agent 2 this follows immediately from the fact that he gains self-control: see (15).²⁵For a strong type, note that

$$y_{PP}^2 = \frac{y_a^2}{\pi_{LH}^1 + (1 - \pi_{LH}^1)x_{PP}^1},$$

so the decrease in x_{PP}^1 raises y_{PP}^2 . Furthermore, the probability that agent 1 plays P given that agent 2 is strong is $\Pr_{PP}^1(P | s) = \pi_{LL}^1 + (1 - \pi_{LL}^1)x_{PP}^1$, whereas when agent 2 is weak it is $\Pr_{PP}^1(P | w) = \pi_{LH}^1 + (1 - \pi_{LH}^1)x_{PP}^1$. We can thus generalize (16) to:

$$W_{PP}^{2.s} = W_a^{2.s} + \delta y_a^2 (B^2 - a^2 - c_L) (1 - x_{PP}^1) \left(\frac{\pi_{LL}^1 - \pi_{LH}^1}{\pi_{LH}^1 + (1 - \pi_{LH}^1) x_{PP}^1} \right).$$

From this equation and (19), which implies that $\pi_{LL}^1 > \pi$, it is clear that $W_{PP}^{2.s}$ also increases.

B.2. Mixed equilibrium

As ρ^{1*} keeps rising, agent 1 becomes less and less likely to exert self-restraint (x_{PP}^1 continues to decline along the path shown in Fig. 6), 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 in Fig. 6 at the point where E_{GN} , in its leftward movement, encounters the vertical $x^1 = x_I^1$ locus.²⁶ By Proposition 5, E_{GN} then ceases to be an equilibrium, and is replaced by $E_M \equiv (x_I^1, (X_{PP}^1)^{-1}(x_I^1))$. Further increases in ρ^{1*} cause E_M to move down along the x_I^1 locus, so that $x^2 = x_M^2$ now declines, as shown in Fig. 6. 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_I^1 .

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 in Fig. 6 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.

B.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 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 in Fig. 6 when the Southward-moving point E_M falls below the Intermediate equilibrium point E_I , which is moving North. The relevant equations from there on are $\mu_{PG}^1(x_1) = \rho^{1*}$ and $\mu_{GP}^2(x_2) = \rho^{2*}$, which correspond in Fig. 6 to the lines $x^1 = x_{PG}^1$ and $x^2 = x_{PG}^2$. As ρ^{1*} continues to rise x_{PG}^1 shifts left (for the same reason as the X_{PP}^1 schedule did), but x_{PG}^2 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. For a weak agent 2, this last implication is immediate, so let us consider a strong type. Using the indifference conditions of the weak type in a group and by himself

$$\Pi^2(x^1, 1, y_{PG}^2) = 0 = \Pi^2(1, 1, y_a^2),$$

we can write $y_{PG}^2 = (y_a^2 - \Pr_{PG}^1(P | w)) / (1 - \Pr_{PG}^1(P | w))$. Substituting this into the expression for the welfare of the strong type, (18), and exploiting the fact that $1 - \Pr_{PG}^1(P | w) = (1 - x_{PG}^1)(1 - \pi_{LH}^1)$, we obtain

$$W_{PG}^{2.s} = W_a^{2.s} + \delta(1 - y_a^2)(B^2 - a^2 - c_L^2) \left(\frac{\pi_{LL}^1 - \pi_{LH}^1}{1 - \pi_{LH}^1} \right),$$

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

 Corresponding author. Department of Economics, Princeton University, Princeton, NJ 08544 1013, USA.
Fax: +1 609 258 5533.

- 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/>, as can the other quotations given below.
- ² For instance, Battaglini and Bénabou [4] study political activism by multiple interest groups or lobbies trying to influence a policymaker. While the framework differs from the present one in many key respects (no time inconsistency, imperfect recall, nor learning from peers), the techniques introduced here turn out to be applicable there as well.
- ³ Task-specific informational spillovers are also evident in Weightwatchers' practice of weighing members each week and reporting to each not just his or her own loss or gain, but also the group average.
- ⁴ Self-help groups may allow members to learn specific techniques (practical, mental or spiritual) for coping with impulses, but such "education" cannot be their sole or even main function. Techniques can be learned from a book or tape; or, if human contact is required, they are best transmitted and tailored to a person's needs by an expert (doctor, counselor, therapist) rather than by non-chosen others who are themselves struggling, not always successfully, with their own weaknesses. Sharing experience with peers, on the other hand, is the best way to judge whether a given set of techniques can indeed work "for someone like me". This broader interpretation, in which group membership gives access *both* to a potentially useful technique and to a pool of "experiments" where one can condition on a very fine set of variables (peers' personal histories, etc.), is fully consistent with our model.
- ⁵ That mechanism is specific to a particular setting and technology, however, and does not apply to most the other behaviors discussed above. In particular, it has the feature that being with peers—even very bad ones—is always better than being alone.
- ⁶ Note that W need not yield a flow payoff only in subperiod I: a could be the present value, evaluated at (t, I) , of an immediate payoff plus later ones. The important assumption is that there be *some* immediate reward to choosing NW . Similarly, NW could also lead to the P/G decision node but with a lower probability than W , without changing any of the results.
- ⁷ The precedent-setting role of lapses is emphasized by Ainslie [1]. Baumeister et al. [6] refer to it as "lapse-activated snowballing," and Elster [22] as a "psychological domino effect".
- ⁸ See Bénabou and Tirole [9] 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.
- ⁹ In their experiment, subjects were led to believe that increased tolerance, following physical exercise, for keeping one's hand in near-freezing water was diagnostic of either a good or a bad heart condition. They reacted by, respectively, extending or shortening the amount of time they withstood that pain.
- ¹⁰ The probabilities that both agents are low types, high types, or of opposite types are then $\rho^2 + \alpha\rho(1-\rho)$, $(1-\rho)^2$

$+\alpha\rho(1-\rho)$ and $(1-\alpha)\rho(1-\rho)$, respectively.

11 The figure describes the weak type's strategy in the (most interesting) subgame where the decision node between P and G has been reached. This confrontation with cravings could be the result of a 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.

12 The mixed-strategy nature of most equilibria in our model is, as usual, an artefact of the discreteness of the type space. As in most other dynamic games of incomplete information (e.g., bargaining games) it would disappear with a continuum of types.

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

14 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. To understand the shape of $\rho_1(\alpha)$ and $\rho_2(\alpha)$, recall that $x_{PP}(\cdot;\alpha)$ shifts up with α , while $x_{PG}(\cdot;\alpha)$ shifts down: a greater correlation magnifies both the “discouragement” and the “encouragement” effects of the other agent's choosing G or P , respectively. The incentive constraint $x_I(\cdot,\alpha)$, meanwhile, 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. Therefore, $\rho_2(\alpha)$, which is the intersection of $x_{PG}(\cdot;\alpha)$ and $x_I(\cdot;\alpha)$ is increasing in α ; $\rho_1(\alpha)$, by contrast, need not be monotonic.

15 Ex post utility levels, denoted as $W^{s,\beta}$ and $W^{w,\beta}$, refer here to the preferences of the second-subperiod self, which incorporate the present bias. Caplin and Leahy [15], for instance, argue that in problems with changing preferences one should aggregate the expected utilities of the different temporal selves using a Bergsonian welfare criterion, as in a standard social choice problem.

16 Throughout the welfare analysis we focus on the case where $\rho<\rho^*$ (otherwise, peers are irrelevant to self-control). We also assume that at $t=1$ the willpower activity is undertaken (either by choice or because it cannot be avoided for sure; see footnote 11), so that the agents is indeed confronted with temptation.

17 The interim levels W^s and W^w are also of further interest in situations where the agent interacts with a better informed but altruistic principal (see [16] for such a model in the context of medical advice). Consider a parent deciding whether or not to let her child frequent certain peers, or a judge deciding whether a substance abuser should be compelled to join a “twelve-step” program. This principal (whether purely paternalistic or also concerned about externalities) will often have evidence (typically of a “soft”, nonverifiable nature) on the agent's type that the latter does not, or is in denial about; she will then evaluate group membership for the agent based on her own priors over W^s and W^w . (Again, putting weight as well on ex post, salience-distorted payoffs does not change anything).

18 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.

19 We shall similarly denote $\pi_{HL}^i \equiv 1 - \pi_{LL}^i$ and $\pi_{LH}^i \equiv 1 - \pi_{HH}^i$. Condition (19) below is then equivalent to $\pi_{LH}^i/\pi_{LL}^i < 1 < \pi_{HH}^i/\pi_{HL}^i$, for $i=1,2$.

20 Clearly, $\mu_{PG}^i(x^j)$ is independent of x^j : once agent j has given in, his type is completely revealed. The functions μ_{PP}^i and μ_{PG}^i 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.

²¹ For simplicity, we shall focus on this case from here on. The case where any of the intersections x_{PP}^i occurs outside the $(0,1) \times (0,1)$ box is easily analyzed using the techniques developed in this section, and it yields the same intuitions.

²² 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,

$\rho^2 = \rho^1 \pi_{LL}^2 + (1 - \rho^1)(1 - \pi_{HH}^2)$. For instance, if it is common knowledge that both agents are always of the same type ($\pi_{HH}^i = \pi_{LL}^i = 1$), then $\rho^1 \equiv \rho^2$. Conversely, for ρ^2 to remain unaffected, the conditional probabilities π_{LL}^2 and π_{HH}^2 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.

²³ The comparative statics of the Intermediate 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 the results (see [Proposition 6](#)) apply unchanged when there is a unique equilibrium that is of the Good News or Mixed type.

²⁴ See Bodner and Prelec [\[12\]](#) and Benabou and Tirole [\[10\]](#) for accounts of rigid behavior and compulsive personal rules in a single-agent setting.

²⁵ Eq. [\(15\)](#) 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 3](#).

²⁶ Recall that the x_I^1 locus is defined by agent 2's incentive constraints $\pi^2(x^1)=0$, 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_I^2 by agent 2 required for the indifference condition $\pi^1(x^2)=0$ to hold rises. Thus the x_I^2 locus shifts up with ρ^{1*} , and so does the point $E_M = (x_I^1, x_I^2)$.

Journal of Economic Theory

Volume 123, Issue 2 , August 2005, Pages 105-134

This Document

- [SummaryPlus](#)
- ▶ **Full Text + Links**
 - [Full Size Images](#)
- [PDF \(355 K\)](#)

External Links

- 

Actions

- [Cited By](#)
- [Save as Citation Alert](#)
- [E-mail Article](#)
- [Export Citation](#)

[Home](#)
[Search](#)
[Journals](#)
[Books](#)
[Abstract Databases](#)
[My Profile](#)
[Alerts](#)
 [Help](#)

[Feedback](#) | [Terms & Conditions](#) | [Privacy Policy](#)

Copyright © 2005 [Elsevier B.V.](#) All rights reserved. ScienceDirect® is a registered trademark of Elsevier B.V.

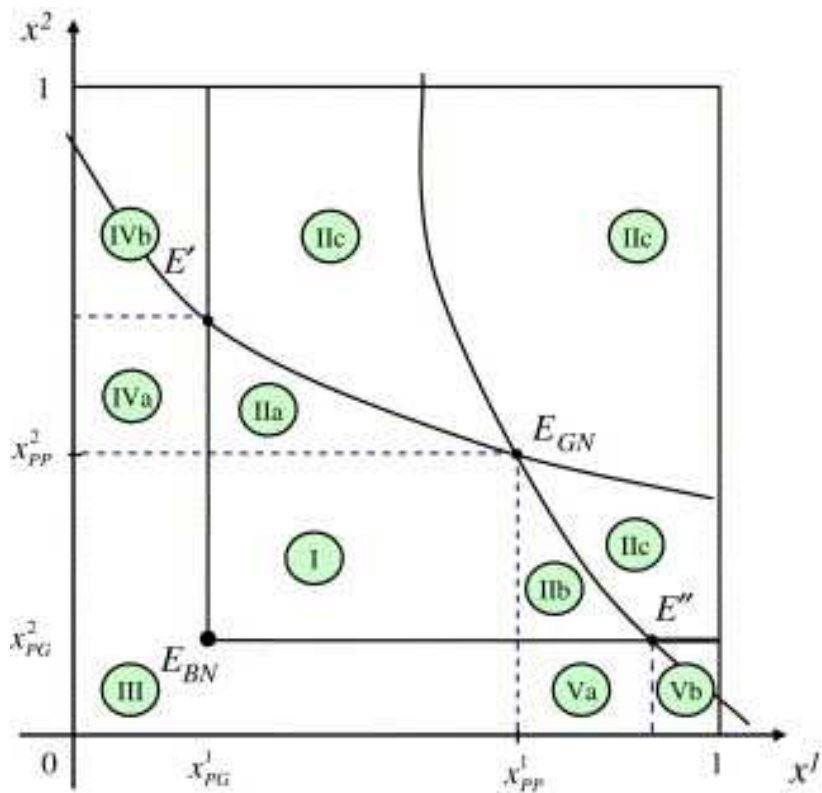


Fig. 7. Equilibrium set in the general (asymmetric) model.