

10

Social Emotions

Let's not forget that the little emotions are the great captains of our lives and we obey them without realizing it.

Vincent Van Gogh, Letter to Brother Theo (1889)

The heart has reasons that Reason knows nothing about.

Blaise Pascal, *Pensées* (1995[1670])

10.1 Introduction

Social emotions—love, guilt, shame, self-righteousness, and others—are responsible for the host of civil and caring acts that enrich our daily lives and render living, working, shopping, traveling among strangers, even the conduct of scientific research, feasible and pleasant. Adherence to social norms is underwritten not only by cognitively mediated decisions, but also by emotions (Frank, 1987, 1988; Ekman, 1992; Damasio, 1994; Elster, 1998; Boehm 2007). When Bosman and Zeelenberg (2001) assayed the feelings of respondents in an ultimatum game, they found that low offers provoked anger, contempt and sadness, that the intensity of the self-reported emotions predicted the respondents' behavior, stronger emotions inducing rejections. Interestingly, the introduction of an hour-long “cooling off” period between offer and the respondent's choice of an action had no effect on either reported emotions or on the rejection behaviors of the respondents. Recall from chapter 3 that Sanfey et al. (2003) found that those rejecting low offers in an ultimatum game led to heightened levels of activation in the brain areas associated with disgust and anger.

One of the most important emotions sustaining cooperation is shame, the feeling of discomfort at having done something wrong not only by one's own norms but also in the eyes of those whose opinions matter to you. Shame differs from guilt in that, while both involve the violation of a norm, the former but not the latter is necessarily induced by others' knowing about the violation and making their displeasure known to the violator.

We will suggest that shame, guilt, and other social emotions may function like “pain,” in providing personally beneficial guides for action that bypass the explicit cognitive optimizing process that lies at the core of the standard behavioral model in economics and decision theory. Pain is one of the six so-called ‘basic’ emotions, the others being pleasure, anger, fear, surprise, and disgust. Shame is one of the seven so-called “social” emotions, of which the others are love, guilt, embarrassment, pride, envy, and jealousy (Plutchik 1980, Ekman 1992). Basic and social emotions are expressed in all human societies, although their expression is affected by cultural conditions. For instance, in all societies one may be angered by an immoral act, or disgusted by an unusual foodstuff, but what counts as an immoral act or a disgusting foodstuff is, at least to some extent, culturally specific.

Antonio Damasio (1994):173 calls an emotion a “somatic marker,” that is, a bodily response that “forces attention on the negative outcome to which a given action may lead and functions as an automated alarm signal which says: Beware of danger ahead if you choose the option that leads to this outcome....the automated signal protects you against future losses.” Emotions thus may contribute to the decision-making process by working with, not against, reason. Damasio continues, analogizing emotions to physical pain: “suffering puts us on notice....it increases the probability that individuals will heed pain signals and act to avert their source or correct their consequences.” (p. 264)

To explore the role of guilt and shame in inducing social behaviors we will consider a particular interaction having the structure of a public goods game. In the public good setting, contributing too little to the public account may evoke shame if one feels that one has appropriated “too much” to oneself. Because shame is socially induced, being punished when one has contributed little triggers the feeling of having taken too much. In this case, the effect of punishment on behavior may not operate by changing the incentives facing the individual, that is by making it clear that his payoffs will be reduced by the expected punishments in future rounds. Rather it evokes a different evaluation by the individual of the act of taking too much, namely, shame. This is the view expressed by Jon Elster (1998):67 “material sanctions themselves are best understood as vehicles of the emotion of contempt, which is the direct trigger of shame.” Thus, self-interested actions, *per se*, may induce guilt, but not shame. If one contributes little and is not punished, one comes to consider these actions as unshameful.

If, by contrast, one is punished when one has contributed generously, the emotional reaction may be spite towards the members of one's group.

We assume individuals maximize a utility function that captures five distinct motives: one's individual material payoffs, how much one values the payoffs to others, this depending both on one's altruism and one's degree of reciprocity, and one's sense of guilt or shame in response to one's own and others' actions. To this end, we will amend and extend a utility function derived from the work of Geanakoplos, Pearce and Stacchetti (1989), Falk and Fischbacher (2006), Levine (1998), and Sethi and Somanathan (2001).

In Chapter 3, we presented experimental evidence consistent with the view that punishment not only reduces material payoffs of those who transgress norms, but also recruits emotions of shame towards the modification of behavior. In Chapter 6 we showed that the altruistic punishment of shirkers by strong reciprocators can proliferate in a population and sustain high levels of cooperation. Here, we show that social emotions and punishment of miscreants are synergistic, each enhancing the effects of the other.

In Section 10.2, we model the process by which an emotion such as shame may affect behavior in a simple public goods game. We then show that shame and guilt along with internalized ethical norms allow high levels of cooperation to be sustained with minimal levels of costly punishment, resulting in mutually beneficial interactions at limited cost. In Section 10.5, we ask how prosocial emotions such as shame might have evolved. Indeed, we show that if the return to the public good is sufficiently high and if the reciprocity motive is strong, an individual is benefitted by increasing his shame parameter, as this leads to a Nash equilibrium in which his net utility is higher.

10.2 Reciprocity, Shame, and Punishment

Consider two individuals who play a one-shot public goods game in which each has a norm concerning the appropriate amount to contribute to the public project, and each (a) values his own material payoff, (b) may prefer to punish others who contribute insufficiently, (c) feels *guilt* if he contributes less than the norm; and finally (d) experiences *shame* if he is sanctioned for having contributed less than the norm. This psychological repertoire captures some of the motives that explain cooperation in behavioral experiments. The results that follow generalize to an n -person interaction.

We assume each individual starts with a personal account equal to 1 unit. Each individual contributes to the public project an amount a_i , $0 \leq a_i \leq 1$, and each receives $\chi(a_1 + a_2)$ from the project, where $1/2 < \chi < 1$. Thus, the individuals do best when both cooperate ($a_i, a_j = 1$), but each has an incentive to defect ($a_i, a_j = 0$) no matter what the other does. In the absence of punishment, this two-person public goods game thus would be a prisoners' dilemma. But at the end of this *production period* there is a *punishment period*, in which the individuals are informed of the contribution of the other individual, and each individual may impose a penalty μ on the other individual at a cost $c(\mu) = c\mu^2/2$. This, and the other functional forms below, are chosen for expositional convenience.

In what follows, we represent the two players as i and j , where $j \neq i$. Letting μ_{ij} be the level of punishment of individual j by individual i , the material payoffs to i is then given by

$$\pi_i = 1 - a_i + \chi(a_1 + a_2) - \mu_{ji} - c(\mu_{ij}) \quad (10.1)$$

In (10.1), the first two terms give the amount remaining in i 's private account after contributing, the third term is i 's reward from the public project, the fourth term is the punishment inflicted by j upon i , and the final term is the cost to i of punishing j .

We assume that the norm is that each should contribute the entire endowment to the public project. The results generalize the case where the norm is less stringent. We represent the propensity of i to punish j for not contributing by

$$\beta_{ij} = \lambda_i(a_j - 1), \quad (10.2)$$

where we assume $0 < \lambda_i < 1$, so that unless j contributed his entire endowment, i receives a psychic benefit from lowering j 's material payoff that is proportional to j 's shortfall. The parameter λ_i , $0 < \lambda_i < 1$, is the strength of i 's *reciprocity* motive. The condition that $\lambda_i < 1$ ensures that individual i cannot value j 's payoffs negatively more than he values his own positively so that should both payoffs increase proportionally, individual i cannot be worse off. The shame experienced by i is a psychic cost proportional to the degree to which he is punished by j , and is equal to $v_i(1 - a_i)\mu_{ji}$. Thus, punishment triggers shame, which is greater the more the individual has kept for himself rather than contributing to the public project, and the larger is v_i , the susceptibility of individual i to feeling shame. Finally, i may feel guilt simply for having violated his internal standards of moral behavior. We

represent this feeling by $-\gamma_i(1 - a_i)$, which is negative for $\gamma_i > 0$ unless i contributes the full amount to the project.

The utility function i is then given by

$$u_i = \pi_i + \beta_{ij}(1 - a_j + \chi(a_1 + a_2) - \mu_{ij}) - (\gamma_i + v_i \mu_{ji})(1 - a_i) \quad (10.3)$$

We have not included the cost to j 's of punishing i , in the material payoffs of j that i takes account of when choosing his contribution level. Note that shame and punishment are complementary in the sense that an increase in the susceptibility of shame increases the marginal effect of punishment on the individual's utility, and an increase in the level of punishment similarly raises the marginal effect of a shameful action on the actor's utility. Shame thus enhances what is termed the 'punishment technology,' the effectiveness of which is measured by the ratio of the cost inflicted on the target to the marginal cost to the punisher of undertaking the punishment. This punishment effectiveness ratio for i 's punishment of j is thus

$$\frac{1 + v_j(1 - a_j)}{c\mu_{ij}}, \quad (10.4)$$

from which it is clear that the punishment of j is more effective the more susceptible to shame is j .

Because each individual's contribution will be known to the other when the levels of punishment are chosen, individual i will choose μ_{ij} to maximize u_i , given a_j . So i chooses the level of punishment of j such that $du_i/d\mu_{ij} = 0$, which gives

$$\mu_{ij} = \frac{\lambda_i(1 - a_j)}{c}, \quad (10.5)$$

and similarly for individual j , with i and j reversed. This condition requires i to adopt a level of punishment that equates the marginal cost of punishing $c\mu_{ij}$ to the marginal subjective benefit that i experiences by punishing j for his transgression, $\lambda_i(1 - a_j)$.

Individual i knows that j will punish him less, the more he contributes, the marginal effect on punishment of contributing more being $-\lambda_j/c$. Equation 10.6 is the condition that determines this utility-maximizing contribution level, namely that for which $du_i/da_i = 0$. Thus i selects a level of contribution that equates the marginal cost of contributing (the left hand side 10.6) with the marginal benefits, including this punishment reduction effect, given the contribution level a_j of player j .

$$1 = \chi - \lambda_i \chi(1 - a_j) + \gamma_i + \frac{\lambda_j}{c} + \frac{2\lambda_j v_i(1 - a_i)}{c} \quad (10.6)$$

The first term on the right hand side of (10.6) is the marginal return from the public project accruing to i . The second term is the negative effect of contributing induced by the fact that i 's contribution increases j 's payoff at rate χ , and unless j has contributed fully, j experiences this negatively. The third term is the marginal reduction in guilt from contributing, and the fourth term is the marginal reduction in punishment received. The final term is the marginal reduction in shame experienced, taking account of both the reduction in punishment and the reduction in the shame felt for a given amount of punishment.

10.3 Shame and the Economy of Punishment

The best response for individual i is just the value of a_i that solves equation (10.6) which, assuming v_1 and v_2 are positive, and upon rearranging that equation, gives:

$$\alpha_i(a_j; v_i) = 1 - \frac{\xi_i + c\lambda_i\chi(1 - a_j)}{2\lambda_j v_i}, \quad (10.7)$$

where $\xi_i \equiv c(1 - \gamma_i - \chi) - \lambda_j$. We assume $\xi_i \geq 0$, so $a_i(a_j; v_i) \leq 1$ even when $a_j = 1$, thus avoiding having to deal with corner solutions. We show below that these inequalities are necessary to ensure that the resulting Nash equilibrium is stable in the sense that a small deviation from a best response by an individual leads both individuals to react by adjusting towards equilibrium.

Note that (10.7) implies that the best response a_i is an increasing function of a_j , and the a_i schedule shifts up when v_i , γ_i or λ_j increases, corresponding to our intuitions concerning the model. There is also a minimal level of susceptibility to shame supporting positive contributions which, setting $a_i = 0$ in (10.7) and rearranging, is

$$v_i^{\min} = \frac{c}{2\lambda_j}((1 - \gamma_i - \chi(1 - \lambda_i)) - \lambda_i\chi a_j) - \frac{1}{2}, \quad (10.8)$$

from which we can see that the minimal level of shame that will induce a positive contribution is increasing in the cost of punishment, and decreasing in i 's susceptibility to guilt γ_i , j 's level of reciprocity λ_j , and the productivity χ of the public project, again confirming our intuitions.

However, we need to know more than how each individual reacts to changes in the determinants of their behavior. We need to study how the

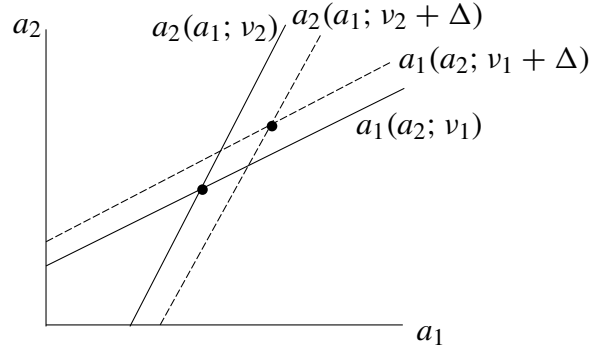


Figure 10.1. Mutual determination of contributions to a public project. Note: The functions slope upwards because the individuals are reciprocators and shift as shown when susceptibility to shame, v increases, because this enhances the effects of punishment.

interaction of the two jointly determines the levels of contribution, punishment, and payoffs. To do this, we define the Nash equilibrium outcome for the two; that is, the outcome that is a best response for each, given what the other does. The Nash equilibrium is shown in Figure 10.1. The dotted lines show the effects of increasing the shame parameter a small amount for both individuals. To determine the level of contributions in the Nash equilibrium we solve the two best response functions simultaneously for a_i and a_j . The equation for a_i^{Nash} is given

$$a_i^{\text{Nash}} = 1 - \frac{c\chi\xi_j + 2v_j\xi_i}{\lambda_j(4v_1v_2 - c^2\chi^2)}. \quad (10.9)$$

The Nash equilibrium is stable only if $c^2\chi^2 < 4v_1v_2$, which we assume. Where this condition is not fulfilled, one can see from 10.7 that $da_i/da_j > 1$, so an increase in the contribution of one is more than matched by the increase in the other, as would be the case in Figure 10.1 if the $a_1(\cdot)$ function were 'steep' and the $a_2(\cdot)$ function were 'flat'. Indeed, we will make the stronger assumption that $v_i, v_j > c/2$, so the above inequality holds for all $\chi \leq 1$. The joint satisfaction of $\xi_i, \xi_j \geq 0$, and $c^2\chi^2 < 4v_1v_2$ ensure that the equilibrium in (10.9) is stable. From (10.9) we calculate in (§A7) that $da_i^{\text{Nash}}/d\gamma_i > 0$, so an increase in i 's guilt increases i 's contribution in the Nash equilibrium. Also, we find that $da_i^{\text{Nash}}/dv_i > 0$, so an increase in i 's shame parameter increases his contribution in the Nash equilibrium. We also have (§A7) $da_i^{\text{Nash}}/d\lambda_j > 0$, so an increase in j 's reciprocity motive

leads to an increase in i 's contribution in the Nash equilibrium. We should note that this model conforms to the intuitive notion that if the efficiency of cooperation χ increases, the equilibrium contributions of both individuals increase, provided $v_i = v_j = v$ and $\lambda_i = \lambda_j = \lambda$. In this case, we can show that $da_i^{\text{Nash}}/d\chi$ has the same sign as

$$(2v + c\chi)^2(\lambda + 2v - c) \quad (10.10)$$

which is strictly positive, since we have assumed $2v > c\chi$ for all $1/2 < \chi < 1$.

10.4 The Coevolution of Shame and Punishment

Thus, we have an important result. From (10.5) we see that increasing contributions lowers punishment, so increasing v_1 and v_2 must both increase contributions, as is evident from Figure 10.1, and lower punishments in the Nash equilibrium. This is the sense in which we mean because shame enhances the effectiveness of punishment, in equilibrium it economizes on the cost of punishment. Moreover, because we can write

$$\begin{aligned} \pi_1 + \pi_2 = & 2 + (2\chi - 1)(a_1^{\text{Nash}} + a_2^{\text{Nash}}) \\ & - (\mu_{12}^{\text{Nash}} + \mu_{21}^{\text{Nash}}) - c(\mu_{12}^{\text{Nash}}) - c(\mu_{21}^{\text{Nash}}), \end{aligned} \quad (10.11)$$

it follows that $\pi_1 + \pi_2$ is an increasing function of the susceptibility of shame of the two. Thus when one individual's susceptibility to shame increases the other individual benefits and when this occurs for both, both benefit. Payoffs therefore are higher in a population that has inculcated a sense of shame in its members, as could be the case for example through the kinds of population-wide internalization studied in the previous chapter.

But could an enhanced sense of shame raise an individual's payoffs even if the other individual's sense of shame remained unchanged? Having studied the evolution of altruistic punishment in Chapter 6 and wanting to focus here on shame, we have assumed that the direct effect of increasing i 's own contribution lowers i 's payoffs, even taking account of the reduced punishment that i will receive from j as a result; namely, $1 - \chi > \lambda/c$. Thus the only way that i 's payoffs can rise as a result of an increase in his susceptibility of shame is if j 's reciprocal response to i 's increase contribution is sufficiently great. Assuming that the two have identical reciprocity levels, this requires (from 10.7) that $da_j/da_i = c\chi/2v$ be large, but not

greater than one, in which case the equilibrium would be unstable, as we have already seen that if the two ‘overrespond’ to one another, instability will result.

We can indeed show (§A7) that if the common project is sufficiently productive (χ is large) and if individuals are sufficiently reciprocal (λ is large), an enhanced level of shame may raise one’s own payoffs; that is, we may have $d\pi_i^{\text{Nash}}/dv_i > 0$. This cannot occur, however, where reciprocity is absent or where the benefits to cooperation are minimal, or where shame itself is at a very low level.

10.5 The Evolution of Social Emotions

Human behaviors systematically deviate from the model of the self-interested actor, and we think the evidence is strong that social emotions account for much of the discrepancy. But, this descriptions of behavior would be more compelling if we understood how social emotions might have evolved, culturally, genetically, or both. There are two puzzles here. First, social emotions are often altruistic, indicating actions benefiting others at a cost to oneself, so that any dynamics in which the higher payoff trait tends to increase in frequency, social emotions would eventually disappear. We have addressed this puzzle in the previous three chapters, showing that by the process of group competition and norm internalization, generically altruistic traits may evolve.

The second puzzle concerns social emotions *per se*. How could it ever be evolutionarily advantageous to bypass one’s cognitive decision making capacities and let behavior be influenced by the visceral reactions associated with one’s emotions? We have addressed a similar question in the previous chapter: internalizing norms may be a way of economizing the costs of calculating benefits and costs in each situation, and of averting costly errors when the calculations go wrong. A related argument, we think, helps explain the evolutionary viability of social emotions.

Humans tend to be present-oriented, a condition they share with other animals (Stephens et al. 2002). We tend to discount future costs and benefits *myopically*, that is, more than either a fitness-based or a lifetime welfare-based accounting would require. The mismatch between our time preference and our fitness is in part due to the payoff to patient behaviors that resulted from the extended life histories and prolonged period of learning the skills associated with the distinctive skill-intensive human feeding niche based

on hunted and extracted foods. Prior to this period in human history, the importance of the future was more limited and largely concerned the survival of one's offspring. A genetically transmitted disposition to assist one's relatives may have produced a selective degree of patience as a byproduct of inclusive fitness maximization—resisting stealing food from one's offspring, for example. Even if our genetic development in a cooperative social context has mitigated the extremes of lying, cheating, killing, stealing, and satisfying short-term bodily needs (wrath, lust, greed, gluttony, sloth), we nevertheless have a fitness-reducing bias towards behaviors that produce immediate satisfaction at the expense of our long-run well-being.

The internalization of norms and the expression of these norms in a social emotion such as guilt addresses this problem by inducing the individual to place a *contemporaneous* value on the future consequences of present behavior, rather than relying upon an accurate accounting of its probable payoffs in the distant future. One curbs one's anger today not because there may be harmful effects next month, but because one would feel guilty now if one violates the norms of respect for others and the dispassionate adjudication of differences. One punishes others for behaving anti-socially not because there are future benefits to be gained thereby, but because one is angered at the moment.

Do the social emotions thus function in a manner similar to pain? Complex organisms have the ability to learn to avoid damage. The measure of damage is pain, a highly aversive sensation the organism will attempt to avoid in the future. Yet an organism with complete information, an unlimited capacity to process information, and with an fitness-maximizing way of discounting future costs and benefits would have no use for pain. Such an individual would be able to assess the costs of any damage to itself, would calculate an optimal response to such damage, and would prepare optimally for future occurrences of this damage. The aversive stimulus—pain—would then be strongly distorting of optimal behavior, because pain will lead the individual to assuasive and avoidance behavior in addition to responding constructively to the damage. Because pain clearly does have adaptive value, it follows that modeling pain presupposes that the individual experiencing pain must have incomplete information and/or a limited capacity to process information, and/or an excessively high rate of discounting future benefits and costs. Is guilt a social analogue to pain?

If being socially devalued has fitness costs, and if the amount of guilt or shame that a given action induces is closely correlated with the level of these

fitness costs that would otherwise not be taken account of, then the answer is affirmative. The same argument will hold not only for fitness costs, but for any effect, possibly operating through cultural transmission, that reduces the number of replicas an individual will generate. Shame and guilt, like pain, are aversive stimuli that lead the individual experiencing them to repair the situation that led to the stimulus, and to avoid such situations in the future.

10.6 Conclusion

Shame and guilt, like pain, replace an involved optimization process with a simple message: whatever you did, undo it if possible, and do not do it again. Two types of selective advantage thus may account for the evolutionary success of shame and related social emotions. First, social emotions may increase the number of replicas, by either genetic or cultural transmission, of an individual who has incomplete information (e.g., as to how damaging a particular anti-social action is), limited or imperfect information-processing capacity, and/or a tendency to undervalue costs and benefit that accrue in the future. Probably all three conditions conspire to induce people to respond insufficiently to social disapprobation in the absence of social emotions. The visceral reactions associated with these emotions motivate a more adequate response, one that will avert damage to the individual. Of course the role of social emotions in alerting us to negative consequences in the future presupposes that society is organized to impose those costs on rule violators. The social emotions may thus have coevolved with the emotions motivating punishment of antisocial actions, modeled in the previous chapters.

The second selective advantage favoring the evolution of social emotions refers specifically to shame. The fact that higher levels of shame, in both individuals, raise the sum of their payoffs also suggests that shame may evolve through the effects of group competition. As we have seen, where the emotion of shame is common, punishment of antisocial actions will be particularly effective and as a result seldom used. Thus groups in which shame is common can sustain high levels of group cooperation at limited cost and will be more likely to survive environmental, military and other challenges, and thus to populate new regions or sites vacated by groups that failed. As a result, selective pressures at the group level will also favor religious practices and systems of socialization that support susceptibility to shame for failure to contribute to projects of mutual benefit of the type modeled in the previous two sections. Moreover, where the returns to cooperation and levels of

reciprocity are sufficiently great, an individual who acquires an enhanced sense of shame will increase his payoffs. This means that an individual that acquired enhanced shame by chance (a mutation, developmental accident or other) could invade a large population of individuals with lesser levels of shame. Thus, a genetic or cultural predisposition to shame could increase in a population even in the absence of groups competition.

It is quite likely then, that the 'moralistic aggression' that is involved in the altruistic punishment of miscreants and that motivated the punishment of shirkers in Chapter 6 also created a selective niche favorable to the emergence of shame and other social emotions, or what Boehm calls a conscience:

The human conscience evolved in the Middle to Late Pleistocene as a result of subsistence turning to the hunting of large game. This required...cooperative band-level sharing of meat...bands had to gang up physically against their alphas to ensure efficient meat distribution. This set the stage for morality to develop as a new, more socially-sensitive type of personal self-control became adaptive for individuals living in these punitive groups. Thus a conscience began to develop biologically. In turn...conscience transformed social control by making punitive sanctioning increasingly moral and also less lethal, as group ostracism and shaming evolved.

Combining the model of this chapter and that of Chapter 6, the emergence of shame would have reduced the costs of punishing transgressors incurred by the strong reciprocators. The reason for this is that gossip and ridicule could then suffice where physical, often violent, elimination from the group had been necessary in the absence of shame. The proliferation of strong reciprocators engaging in altruistic punishment that this cost reduction allowed would then have enhanced the adaptive advantages of shame. The groups in which this occurred initially would have enjoyed survival advantages over other groups.