Emotions, Signalling and Strategic Coordination

Don Ross

University of Cape Town
dross@commerce.uct.ac.za

Paul Dumouchel

Université du Québec à Montréal
Paul.dumouchel@uqam.ca

Abstract

This paper reviews some recent literature on emotions by economists and philosophers, especially Robert Frank and his critics. We endorse the general stance of this literature, according to which emotions evolved as devices for minimize incidence of social dilemmas. However, we reject the prevailing theory of emotions themselves, on grounds that it involves simplistic cognitive science and leads to mistakes in the associated game theory. In stead, we offer a view according to which the set of recognized emotional state-types is an evolved conventional signaling system, especially useful for conveying information about preference intensities in bargaining situations.

The emotional is often taken as the contrasting foil of the rational. As a result most economists, given their preoccupation with *rational* maximization, but simultaneously following Hume in taking reasons to be the "slaves of the passions," have seen the domain of the emotions as deeply important, but as an area that they themselves can play no role in helping to study.

This abstemiousness has been undermined over the past dozen years, however, as a result of work by Frank (1988) and Hirshleifer (1987). This Hirshleifer-Frank thesis (`HFT'), has come to widespread attention through Frank's 1988 book, *Passions Within Reason*, which used a combination of evolutionary psychology and Smithian moral philosophy to endogenize emotional influences within the model of the rational economic agent However, Frank has engendered confusion through inconsistency on his part over what, exactly, is supposed to be *rational* about rational agents. His formulation proposes what he calls a 'commitment model', which he contrasts with what he calls 'the self-interest model'. Appealing to game-theoretic work on the efficacy of threatening and promising, Frank argues that strategic commitment is necessary for the maximization of long-run interest insofar as it makes both threats and promises credible, and that evolution produced and sustains emotional responses to serve as the vehicles of such commitment. Examples from Frank's book provide the best vehicles for illustrating the logic. A person genuinely in love will have difficulty concealing this, thus signaling (contrary to what might be in her best short-run interest, construed as narrowly selfish) that the partner need not pay a higher price for fidelity than simple reciprocation and preservation of dispositions and qualities which have, presumably, elicited the love in the past. This is crucial lest couples find their attempts to settle down undermined by equilibrium-selection uncertainties. That is, partners would tend to subvert the security of their pacts (and hence, often, their long-run optimal payoffs) by being continuously unsure as to whether they were under- or over-incentivizing each other. Here, the emotion of love sustains reciprocal promise-keeping. By the same logic, a threat which its utterer would be irrational to carry out will not be regarded as cheap talk if the threatener may be disposed to uncontrollable rage in the face of defection, and if this is known to the other party. Given that this sort of reciprocal awareness is critical to the effectiveness of emotions as bargaining chips, Frank argues, it is no surprise that they are typically accompanied by visually salient facial expressions and body postures which are as difficult to willfully produce as are the underlying emotions themselves. Emotions thus constrain available strategy spaces in such a way as to foreclose the possibility of certain social dilemmas through being beyond the reach of strategic manipulation by bargainers.

The HFT finds a place in a larger literature from the past two decades on so-called `constrained maximization'. The phrase is due to Gauthier (1986), who sought to preserve the neoclassical conception of instrumental rationality while simultaneously maintaining that maximizers can sometimes rationally act against their preferences (by cooperating in one-shot prisoner's dilemmas, for example) for the sake of higher expected payoffs across ranges of related games. In Gauthier's case, the point was to justify a strategic role for morality rather than for emotions, but the logic is similar. Suppose I promise you that I will cooperate in a one-shot PD we face. This is my best strategy only if I believe that you'll believe that I'll follow through on my promise; but since it would not be rational for me to do so, it is not rational for you to believe me, or for me to expect

my assurance to convince you. But now suppose, Gauthier asks, that I am morally indoctrinated with the sanctity of promise-keeping, and that you know this; and that you are similarly indoctrinated and that I know this. Then we can both promise to cooperate and get the payoffs from mutual cooperation that are Pareto-superior to those from mutual defection. Morality on this story functions as a commitment device, just as emotions do on Frank's. Defenders of constrained maximization often contrast *myopic* with *non-myopic* agency. The unconstrained maximizer acts as if every game is his whole life, failing to observe any standing rules which, if only they were generally observed, would improve everyone's average payoffs across large classes of interaction. Of the two devices for non-myopia mentioned so far, emotions have the apparent advantage of requiring no complex psychology of social and self-indoctrination; evolution, acting so as to maximize *average* payoffs across statistical ensembles of populations (due the effects of genetic shuffling through assortative mating), can wire in the emotional responses and place them beyond the agent's control, whereas morality, being more cognitive, is prey to subversion by sophisticated ratiocination.

It is for this reason that whereas Gauthier's version of constrained maximization has been, to all intents and purposes, demolished in the subsequent critical literature, Frank's account has fared much better. However, the HFT must likewise collapse if it is committed to the logic of constrained maximization. As Binmore (1994) argues, advocates of constrained maximization seek to square circles. If agents are *effectively* non-myopic, that is, if their behavioral architecture includes devices that encourage them to cooperate where the myopic do not, then the effects of these devices must be reflected in their assigned preference structures. As a result, models of games amongst non-myopic agents must assign different payoff-sets than would feature in superficially similar games amongst the myopic. Thus such agents do not act against their preferences; they simply have different preferences from the myopic. This objection, being unrelated to the relative cognitive transparency of morals as compared to emotions, applies just as well against Frank as against Gauthier. Danielson (1991) tries to rescue Gauthier from this problem through appeal to an evolutionary mechanism for inscribing morality, and so brings the Gauthier-inspired version of constrained maximization yet closer in its logic to Frank's. However, LaCasse and Ross (1998) show that Danielson's constrained maximizers behave no differently from conventional maximizers with perfect information about the intentions of others. Thus it does not make sense to talk of agents "acting against their preferences," since if that is possible we lose all analytic grip on what a `preference' is supposed to be. Since genuine social dilemmas, such as PDs, permit no escape, the only means by which rational agents might avoid systematically undermining the stability upon which their flourishing depends is through devices *which minimize the incidence of such dilemmas*. Traditionally, only two devices have seemed plausible: the dismal remedy of Hobbesian tyranny, or reliance upon Smithian 'moral sentiments'. Frank makes clear that he conceives of emotions as roughly identical to the latter (thus stressing the `sentiments' part of Smith's phrase instead of the `moral' part), and that he is basically trying to say what he thinks Smith would have had Smith been alive to read Darwin.

Here, we also see ourselves as following Smith's lead. However, we contend that the proper neo-Smithian account as told in the light of evolutionary psychology and contemporary cognitive science is a good deal more complicated than Frank's. For reasons to be given below, we are dissatisfied with Frank's conception of emotions themselves as well as with his conception of rationality. This said, our attitude is broadly sympathetic: like Frank, we see emotions as evolved strategic devices, and we seek to explain their existence in these terms. In general, then, we aim to save both the substance and the main ambition of the HFT by quite substantially revising it.

Despite our agreement with Binmore that constrained maximization is incoherent, we endorse Nozick's (1993) claim that *being cooperatively socialized* is an aspect of being rational in a 'thick' sense. This is not equivalent to the incoherent view that it is rational to act irrationally in social dilemmas. This important distinction is a subtle one. Greenspan (2000), in her response to Frank, may be read as offering a *mechanism*, based on emotions but not equivalent to them, that could underwrite Nozick's `thick' rationality. That is, a `Nozick-rational' agent might achieve higher long-run expected utility just in case her emotional reactions and attitudes make it more difficult for her to treat her own threats and promises as cheap talk. According to Greenspan, she may achieve this state by cultivating emotional responses in such a way as to emphasize certain motivational aspects of situations – say, aspects connected with dignity and self-worth in the case of so-called `unfair offers' in surplus-division games – over others. This supplements the core idea of the HFT with the idea that agents need not just be passive recipients of emotional states and impulses, but can play cognitively motivated and active roles by encouraging dispositions in themselves to be guided by Smithian sentiments. However we *cannot* on this basis coherently say that our agent either emotes or moralizes her way out of social dilemmas. If the costs to her of reneging on threats and promises have been raised by a disposition to emotionally weight such commitments, then these costs must be factored into the agent's payoffs in any particular game. She has cultivated a set of dispositions that reduces the probability that she will find herself in social dilemmas in the first place; she has *not* performed a logical miracle in escaping from them. Because neither Nozick nor Greenspan make the mistake of confusing avoidance of dilemmas with logical miracles, their accounts are not subject to Binmore's critique of Gauthier.

Trouble threatens, however, when we turn from sequences of one-shot games to more empirically general evolutionary dynamics. The Nozick-Greenspan solution is open to the objection that *feigning* cooperative or retaliatory dispositions through emotional displays may be sound

strategy. But feigning, if adopted as policy by enough agents, or even if merely *feared* by many agents, would generate all the social dilemmas again that are brought about by unsophisticated play. Consider a PD again: if I am afraid that your signal of an emotionally grounded commitment to cooperate might be a feint, I should protect myself by defecting; and you, knowing that I may be so concerned, should therefore defect also, even if sadly. It is precisely at this point that Frank's account is supposed to come to the rescue, and to additionally earn its keep as a piece of evolutionary theory, by explaining why it is *emotion*, rather than some more intellectualized basis for sentiment, that has been selected by Mother Nature to perform the commitment job. Emotions just *are* those sorts of responses that (most) people cannot (easily) feign because their expressions are partly impenetrable to cognition.

It may be true that depressed people cannot feign happiness, or besotted people feign indifference, and this will indeed limit the strategy spaces available to them in certain games. However, many games of interest to behavioural scientists are played over extended temporal periods. In failing to distinguish between episodic emotional reactions and long-term emotional *dispositions*, Frank makes his case appear more plausible than it is once the distinction is emphasized. Frank argues that the locus of emotional responses in the lower brain region is an important aspect of what takes them beyond strategic control. However, lower brain regions are crude information processors, both in that they are relatively impenetrable to deliberative control (the property of them to which Frank's argument appeals), and in that they are responsive only to crudely discriminated data (relative to culturally and socially partitioned information spaces). Since most bargaining situations are not spontaneous, face-to-face encounters, even moderately sophisticated bargainers, in most situations, have at their disposal multiple distancing devices by which to escape the power of primitive impulses. We are thus not confident that Frank's model will work empirically for even the paradigm sorts of cases – social bargaining games – to which he intends it to apply.

Frank in effect turns Hume's distinction between reasons and passions into a chasm, in which the former apply themselves through implausibly careful calculations of utility, while the latter are equally implausibly simple reflexive, hard-wired responses to stimuli. The simplicity of this dichotomy manifests itself in a certain rhetorical schizophrenia in the way Frank tries to present the significance of the HFT. On the one hand, he frequently suggests, it is a radical alternative to the `self-interest' model because the passions of his typical agents cause them to bargain irrationally. (This is the sort of talk that rightly draws the ire of Binmore.) On the other hand, he equally often says that his model is a `friendly amendment' to the `self-interest' model, since his passionate agents *do*, on analysis, turn out to be rational maximisers after all. Building on LaCasse and Ross (1998), we take this rhetorical smoke to be clearly confused. We suggest that the radical passion / reason dichotomy posited at the psychological (as opposed to the economic) level of analysis explains the confusion: Frank's agents must be cunning economic strategists who are simultaneously emotional simpletons. They require emotions lest they cleverly reason themselves into mutual disaster, from which they are saved by being utter puppets of their genes with respect to emotional manifestations and signals. One might indeed feel torn about whether to regard such agents as rational!

What is most primitive about Frank's cognitive science is his conception of types of emotions as referring to discrete neurochemically controlled states. Emotions are thereby viewed as independent forces which `invade' agents' motivational systems; for example, rage moves agents to irrational retaliation. However, understanding emotion strategically requires attention not just to biological mechanisms but to informational dynamics that exploit them. It is not an emotion itself, the rage, but the *threat* of enraged action, that may be to an agent's advantage. Thus it is not the emotion itself, but its expression which, on Frank's account, may be strategically useful. Greenspan, by contrast, shifts attention back to agents' experiences of their own emotions. On her account, an agent cognitively appreciates that she may be able to make a credible threat or promise if she can manipulate her emotions in such a way as to `change her gestalt', that is, foreground certain motivating aspects of a strategic situation that the other party will then recognize as motivating. This is simultaneously a step forwards and a step backwards. It advances our conception by drawing attention to the relationship between emotions and *intensities of preference*. On the other hand, its focus is retrogressive to the extent that it preserves Frank's conception of emotions as fundamentally private inner states, however subject they might be to sophisticated manipulation. The underlying ontology of motivational states here is, we claim, still too simple. Our folk-psychological and cultural conceptualizations of emotion are themselves entangled with the strategic roles of emotional *expressions*; and this is fundamental to their strategic efficacy.

As almost all the recent psychological literature agrees, emotion terms do not simply refer to types of output states; their use is also sensitive to the trains of events (both external and internal) typically leading up to their expressions. According to Dumouchel (1999), emotional signaling is produced by systems of expression which are continuous, in which those events to which we give names of particular emotions are salient moments in uninterrupted processes of affective expression. (That is to say: people are not creatures who normally go about in an affectless state but are then occasionally struck by emotional meteorites. If you were a Martian and you read Frank, or Hume for that matter, that's likely how you'd imagine it was with us.) These expressive moments can be salient for different reasons, many related to the contexts of external events in which they take place, and not only to the neurochemical sources of their phenomenological interpretations by experiencers. It is therefore no surprise that emotions as understood in folk psychology do not correspond neatly to discovered

neurochemical states. The typology of emotional states is, instead, sensitive to the functions of emotional expressions as *signaling systems* in the sense of Skyrms (1996), that is, as informational conventions by which agents can coordinate their actions in games.

The proper form of the central question put by Frank thus involves asking: What interactive purposes do emotional signaling systems evolve and stabilize to serve? One simplifying assumption made in much of the signaling-system literature needs to be discharged here. Frank follows that literature in modeling dispositions to emotional commitments as digital – simply present or absent. Reality is no doubt more complicated: agents may be able to control certain responses but not others, and in some but not all types of circumstances. The agent who seeks to predict another's responses in a strategic interaction thus faces a more complex problem than mere sincerity detection. Suppose that systems of emotional categorization, either within or across cultures, are evolved (genetically, culturally, or both) bodies of signaling conventions. Then expectations in signaling games will be functions of ordered pairs matching particular assignments of emotional expressions to emotional state-types with assumptions about situationally sensitive incentives. (E.g., to interpret another's signal I may need to know which emotional modality as well as which incentive structures a behavior is set within. Her staring at me indicates strong interest; but is it love or anger or …?) These pairs of assignments would then in turn be input to non-parametric analysis. (I.e., I can only strategize within a game once I know whether it is cooperative or uncooperative, etc.) Interpretation problems confronting the design of such a system are formidable. This suggests a strategic rationale for the compression of highly variable input etiologies into constricted ranges of conventionalized emotional expressions as reported and discussed by Griffiths (1997): combinatorial explosion on the receiving ends of signals would otherwise destroy the informational efficacy of a signaling system. As Greenspan notes in considering the idea that bargainers may cognitively manipulate emotions – conceived as inner states – directly, "at a certain point, calculation would become impossibly difficult, and this yields a further reason for relying on emotions as snap interpersonal evaluations." We should add that the same challenge to computational load would arise for mapping emotional state-types onto observed information about strategic settings, given the data on input variability, unless emotional outputs were themselves products of conventionalized signaling equilibria.

Binmore (l994, 183, n. 5) is closer to Dumouchel's view than to Frank's. "I think it unlikely," he says, "that Adam Smith's moral sentiments — anger, contempt, disgust, envy, greed, shame and guilt - all have genuine physiological referents. Under certain circumstances, our bodies pump chemicals into our bloodstream. We then invent myths in seeking to explain to ourselves what we are experiencing. Such myths typically do not separate the train of events that caused the experience from the experience itself". This use of the word 'myths' suggests a policy of eliminativism — reductionism's more honest descendent (see Churchland 1979) — with respect to folk-psychological emotional categories. Dumouchel's account rejects both reductionism and eliminativism. Emotional categories, though they do not refer to neurochemical states, denote real - albeit socially malleable — patterns to which narratives about human interactions, and hence, by social pressure, human interactions themselves, are expected to conform. It is in producing such conformity that emotions as social constructs (built *from*, but not coextensive with, physiological regularities) play their crucial role in facilitating social coordination.

According to Dumouchel, what we categorize as specific emotions are latecomers in the domain of affective phenomena, in the sense that they constitute elaborated constructions resting on more basic forms of affect. (`Basic' in the sense of Panksepp 1998.)`Coordination,' here should not be taken as referring only to the focal points of well-defined conventions in particular games. Rather, this affective coordination is seen as underlying all types of social interactions. People not only play repeated games with each other, but interact with (and expect to interact with) the same individuals across ranges of different games over time. Because in most social settings we usually interact over and over again with the same individuals, we may be in need tomorrow of those with whom we are in conflict today. This leads to problems of reputation, of course, as well explored in the literature, but also to the establishment of preferences concerning *which* individuals we interact with. More generally, it leads to the fact that in a social context what has to be taken into account is not only the value of the objective pursued through the interaction but also the value of the relationship itself for the organisms involved. This is non-parametric: the value of a relationship for one individual is not independent of the value of the relationship for the other. Thus most games are embedded in meta-games, and there are few restrictions on the possible complexities in this recursion; embedding relationships may stack infinitely, may loop, and so on. This circular dependency implies uncertainty concerning the objects of analysis for which equilibria should be forecast. Reciprocal affective expression can then be seen as a means of reducing this uncertainty. Through such things as bodily posture, muscle tone, pitch of voice, and facial expression we *negotiate* reciprocal intentions into tolerably stable sets of expectations *within which* our base-level games are well-defined. At the meta-game level(s) we do not so much exchange information concerning already formed intentions as *dynamically influence and determine* each others' intentions though exchanges of affective expression. Before some negotiation of this sort takes place there is often no fact of the matter as to which game we're playing. The process of affective exchange usually evolves towards some (more or less) fixed point of coordination that `frames' or `tropes' our relationship: distaste, pleasure, love, fear, anger, confidence, disappointment, etc. – the Smithian sentiments. The objects of our emotional state labels are then simply those moments in the process of

coordination that, for one reason or another, are either salient or strongly recurrent. (Most moments of affective coordination remain unconscious and are never assigned any particular emotional state label.[1]) Thus construed, expressed emotions are signals concerning one agent's *standing policy* towards another. These signals *need not* imply anything concerning `inner states,' beliefs, or, if such things exist, the `true intentions' of agents.[2] We prefer to regard them as *proposals in meta-games*. Such proposals are `accepted' when agents agree on emotional labels with which to characterize their interactions. Such agreements then create expectations in terms of which particular games, featuring more or less stable strategic anticipation, can go on.[3]

We should be clear that we are describing this process from a `third-person analyst's' point of view. Whether or not an agent subjectively views his own emotions as something for which he is responsible is of little interest here. The important point is that public emotional-state categorization creates expectations in others, and the very existence of these expectations will then tend to constrain agents' behavior. Because these expectations guide strategic choices by other agents, observations of departures from conventions governing expressions may be punished by myopically rational ostracism. As a recent vein of analysis opening from Lewis (1969) has stressed, the rationality of ostracizing convention-breakers comes close to being an analytically necessary condition for regarding a set of practices as properly *conventional* to begin with. Since affective displays are events which are by definition public, they give signals to all, not only to their intended targets. Independently of what an agent's inner feeling may be, her emotional expression of love or hostility gives rise to expectations, and her failure to fulfill these expectations may be held against her. I may be relieved that Jane's anger was so much noise that had few lasting consequences; nonetheless her credibility will likely be reduced by her sudden change of heart, not only in my eyes, but even more so in the eyes of those who had less to lose if she had carried out her threat. Note that no issue of feigning arises so long as we are directing our attention to conventionalized emotional *expressions* instead of internal states. In systems of coordinating conventions, it must (by definition of `convention') be rational *often enough* for observers of breaches to decline further interactions with the deviant that the conventions in question police themselves. We may therefore conclude that *if* it is conventionalized emotional expressions rather than inner states that play the main strategic role assigned to emotions by Frank, then the difficulties associated with feigning disappear; agents need only be capable of detecting departures from the conventions, rather than dispositions to insincere expression. Greenspan may now be read as successfully identifying an important *secondary* aspect of this sort of process, namely, its reflexive character. As Dennett (1991) has described in detail, personalities are mainly etiologically constituted by internalized systems of expectations derived from their social histories. Thus an agent may anticipate punishing *herself* for departures from conventions relating emotional expression and action, through processes of semi-cognitive reflection that folk psychology refers to under such labels as `lowering of self-esteem' or `loss of sense of self'. Greenspan's basic point is that if agents recognize that other agents face this economy of psychic costs, they have grounds for taking emotionally expressed threats and promises seriously without having to view their utterers as irrational. This point of Greenspan's is almost exactly correct. However, she fails to extend the *scope* of the mechanism (i.e., into the kind of `distance bargaining' that interests economists and political scientists) much further than Frank does because she does not notice the conventional, as opposed to merely the psychically causal, structural dynamics that underpin it.

This first set of expectations, governing relations between emotional expressions and constraints on future social actions, generates a second set governing relations between emotional expressions and types of interactive situations. Because an agent's anger (etc.) is indicative of his attitude or intention towards another, it becomes important for all agents to be able to recognize those situations in which agents are conventionally expected to express anger. Such systems of partition, which sort the world into types of situations according to the emotional responses conventionally appropriate in them, are immediately systems of expectations. That is, to classify some situations as anger-generating is to expect agents to become angry when confronted with them. This second system of expectations can now be used at the interpersonal level as a means for infer-

---

[3] We also give emotional-state labels to consciously salient moments in the process which are not stable strategies in meta-games. Thus, someone might say both that he `loves' his wife and that he is, at some particular moment, feeling (and expressing) lots of `love' for her. The former instance is a strategy in a meta-game. The second might be a move in a particular base-level game (if it has an influence on the wife); or it might just be a bit of narrative self-reference by the lover, in which he uses his meta-game strategy as a device for labeling some experienced nervous events. This fact about our emotion-talk must be recognized and accommodated in the full account, but it complicates matters considerably and so we will pass over it here. For its full explication the reader is referred to Dumouchel (1999).

[4] Yes, we are behaviorists – sophisticated, `Dennettian' behaviorists, we hasten to add.

[5] Scientists and philosophers applying game theory to actual situations have a tendency to write as if the only informational problems that arise for game-players center on equilibrium selection. Our account here will seem very unusual to people in the grip of that over-simplification. We are emphasising that players also must overcome *game-determination* problems. That is, if I know too little about your utility function, or about your knowledge of mine, then how am I to know what game we're playing? Since our preferences over which games we'd rather be playing might often differ, game-determination can itself be the object of meta-games.

ring fine-grained information concerning the other's attitude towards me and mine towards her. If Jane does not get angry with me in a situation where she should be expected to, this may reveal something about her attitude towards me, perhaps that she loves me or that she has become indifferent. Simultaneously, it also reveals evidence about agents' preference-intensities. Charles's excessively defensive reaction at any comment about Louise indicates something about how much he cares for her. In situations which are at a higher remove from immediate personal contact, these expectations become guidelines that constrain behaviour. We expect agents to react in certain normal ways in given situations and deviations from these standard reactions can then serve as significant sources of information.

Our basic claim, then, is that conventions governing emotional expression can constrain interactive behaviour by creating expectations to which agents hold one another responsible in systems of self-enforcing equilibria. They can play this role equally well not only at the ground level of face to face interactions, but also in more abstract and temporally extended bargaining contexts where agents never actually meet. Emotions can also fulfill their strategic function independently of any hypotheses concerning agents' internal states, or of epistemological mysteries about their putative `true feelings'.

According to his thesis, emotional kinds, though constructed on *the basis* of physiological regularities and functional roles ascribed on the basis of behaviour, should not be *identified* with either aspect of their constructive base. As *socially* constructed kinds, they are not identical to purely internal states, either motivationally or behaviourally. Rather, they are – like Smithian sentiments! – conventionalized norms to which people at least generally conform their narratives of themselves and others. As publicly constructed and reinforced conventions, they set exogenous limits on both actions and motivations. Thus, consider one of Frank's examples. As Frank (1988, 243) correctly objects, game-theorists were hasty in regarding the nuclear policy of Mutually Assured Destruction, MAD, as rightly deserving of its acronym. The view that MAD was mad was based on the following well-known reasoning. If one side launches a first strike, no one is better off, and millions are vastly worse off, should the other side retaliate. The threat to retaliate is thus cheap talk. But if both sides know this, and prefer their own security and/or supremacy above all else, then both have an incentive to launch a first strike. Thus, MAD set up a Prisoner's Dilemma, which by some wondrous failure of logic the world escaped during the Cold War. Frank criticizes this conventional wisdom by appeal to his `commitment' model. Suppose that the leaders of the nuclear powers were both, like most people, disposed in such a way that the unprovoked destruction of their nations would have filled either with an uncontrollable wrath leading to an effective revenge-impulse. Furthermore, suppose that both leaders were aware of this human frailty in one another. In that case, their reciprocal threats were not empty, and MAD was not mad; hence, we are still here to say so. Now, on our view, Frank has got the diagnosis *half*-right. This implies, of course, that he has also got it half-wrong. To see where he is wrong, imagine that one of the leaders falls into a deep depression, such that his affectual responses are diminished to vanishing,[4] and that his opponent's intelligence service acquires and relays this information. On Frank's account, it would then be incumbent on the rival leader, if she is rational, to seize the day and let fly her arsenal, knowing that the (irrational) anger which would otherwise bring about retaliation will not arise (unless she knows that her opponent is both *suicidally* depressed and monstrously selfish.) But *is* the rational attacker in fact being sensible here? Our answer is "no". The depressed victim of the first strike may feel nothing beyond further grounds to be depressed. Nevertheless, it is expected of him — both by himself and his compatriots — that he must, given his social role, manifest the responses and actions associated with vengeful anger. No internal state associated with anger need be operative to provoke the retaliatory strike, merely sound appreciation, based on appreciation of social convention, that actions falling under the rubric 'angry response' are expected of him.[5] In due time, however, the clenched jaw and stern stare which his conventional expression encourages are likely to give rise to the other physiological properties with which they are more typically associated.

This view, according to which conventionalized emotions are conveyors of strategically relevant information, invites a question. Why would biological and cultural evolutionary processes support such a roundabout set of mechanisms for transmission of information? Note first that the phenomenon of indirect transmission of social information is not, in fact, unusual. Relatively few differences between people over preferred outcomes are settled by formal bargaining processes; elections, referenda, lawsuits and so forth are generally means of last resort, or employed only

---

[6] 'Depressed,' for the purpose of this example, should be understood consistently with Frank's concept of an emotion rather than ours. That is, assume that the President suffers from a pharmacological condition that prevents him from being able to naively (one might say `sincerely') experience himself as attaching emotional labels to any felt affective states. This will suffice for the purposes of following our story in the example. In fact, we think that phenomena involving depression typically involve conventional signaling dimensions just like other publicly labelled emotional states; we will note the potential significance of this to our story in the next footnote.

[7] The reader is likely to find this part of the scenario much more plausible if she now reads our understanding of an emotional state, rather than Frank's, into the President's depression. To the extent that the President is using public conventions about emotions associated with depression to interpret his own pharmacological affliction, then we need not imagine him playing his vengeful role robotically or in a self-consciously insincere way. He need merely see his identity as "the President" trumping his identity as "a depressed person".

where the number of parties whose immediate interests are impinged is too large for settlement by merely implicit negotiation. More importantly, the most crucial information which must be conveyed in a negotiation is usually not what each party's preference-ordering is, but with what intensities particular outcomes are desired. This is because preference-orderings can often be inferred without any direct communication. This is not so often true of preference intensities. We suggest that one of the sustaining functions of emotions is to serve as devices for signalling information about cardinal utility.

We will build an hypothetical example with which to illustrate and elaborate our claim. Suppose I would rather that Smith order anchovies on his pizza than not, simply because I like anchovies, and if no one but me ever ordered them, pizzerias would stop making them available. However, since I know that Smith's behaviour is of infinitesimal importance to the anchovy market, if I happen to express my preference to him, it would likely be in the tone of a joke — a signal that I do *not* expect him to change his pizza orders on account of my preference. By contrast, if Smith is in the habit of leaving his uneaten pizza residue on my driveway, my preference that he cease doing so will likely be expressed in some anger, in order to signal that failure by him to take my preferences with respect to *this* outcome into account will result in a dispute. Notice that I am unlikely to present Smith with an argument, or with suggestions as to alternative policies he could adopt with respect to his pizza remains, or offer to pay him some small sum in order to change his ways (in which case, the size of the sum offered would signal my preference-intensity). Rather, I am likely to simply angrily tell him something he already knows, namely, that he has left pizza on my property. The entire content of the information-signal in this example consists in the anger, not the proposition reported.

This imagined case is useful for addressing the question about the roundaboutness of emotional signalling. Why do I not simply ignore Smith's aversion to anchovies, while threatening some specific sort of retaliation if Smith does not reform himself in the domain of garbage disposal? After all, Smith has no use for the first bit of information, since it is, *ex hypothesi,* irrelevant to our relationship as bargainers, and he could infer the intensity of my preference in the second case by calculating the cost to me of carrying out my threat. (Though if he does not know the intensity of my preference Smith will have a hard time doing this, since he will not be able to distinguish between a real and an empty threat until it's too late for his estimate to help him.) A first answer appeals to economy. A threat must be calibrated: one must ensure that it is sufficient to achieve its desired effect, but not so severe as to constitute cheap talk and be taken as such (e. g. 'Smith, one more pizza on my driveway and I will blow up your house!'). Attempting to find overly fine calibrations in these sorts of cases is likely to produce excessively costly haggling amongst multiple equilibria, all of which would lie approximately on our contact curves in an Edgeworth box, but none of which may be selected if both parties attach utility to having the last word (presumably as a result of anticipated reputation effects). Thus it may be a more effective policy on my part simply to convey to Smith that his behaviour matters enough to me that he will face consequences for continued pizza-dumping which *might* bring costs likely to be in excess of those he will incur by finding an alternative site for his waste. Resorting to indirect communicative conventions to select general *regions* of exchange prices near which equilibria lie is, in one sense, an opposite approach to that of using conventions to derive logically arbitrary focal points (Schelling 1960) to solve equilibrium selection problems. Where reputations for toughness are at stake, however, conventions that facilitate approximate solutions may serve essentially the same purpose, permitting each of us to avoid appearing to give in to the other. Instead, we both defer to social norms whose weight is greater than our fine stakes in the particular interaction at hand.

While this answer to the roundaboutness problem seems to us to be an important part of the story, it says little about the specific strategic role of the emotions *per se*. Here, then, is a more interesting way in which emotions so conceived can perform a unique sort of task. When I express anger to Smith over his choice of pizza-disposal sites, he should hardly be surprised; I am more likely to be the surprised party, given such odd behaviour. I *can*, however, convey a great deal of information to Smith by way of the extent to which *my degree* of anger *departs from* what conventions would lead us to expect in such situations. Suppose, on the one hand, that I have become aware that Smith is a mafia don, and so express my anger in very diffident tones. Given the relevant social conventions governing emotional expression, I have (very likely) conveyed all of the following information to Smith: (i) that I know that there is reason to fear and respect him; (ii) that I in fact *do* fear and respect him; (iii) but the extent on my fear and respect is not so great that I will suffer the indignity of his throwing pizza onto my property in abject silence; so although I might in fact do nothing further if he continues depositing his food where he does, he should perhaps not remove bodies from the boot of his car in plain view of my window. To say all of this to Smith, under the imagined circumstances, would risk explicit bargaining with someone I would prefer, at all costs, not to involve in a PD or other social dilemma (at least with me). So I reduce the risk by simply conveying that I have a certain *approximate* preference-intensity, and let Smith infer the rest on the basis of his knowledge of our mutual knowledge that my expression of anger has been unusually muted.

In actual historical cases of political bargaining, correct understanding of the relevant conventions governing emotional expression has often been crucial. Churchill would have emboldened Hitler, and harmed the morale of the British public, had he said in 1940 (truly) that he preferred not to sue for peace, instead of filling his addresses as he did with Old Testament moral vehemence. It is, of course, almost certain that Churchill experienced some powerful visceral responses while the *Luftwaffe* pounded British cit-

ies, but, pace Frank, even if he didn't this was irrelevant once his speeches were on the public record. Furthermore, Churchill's literal words "We shall never surrender" were not (on the face of things) altogether to the point, since (a) Hitler was not then demanding British surrender, and (b) no doubt some logically possible German action, if pursued with sufficient means, could have brought the British to the table. What Churchill's choice of emotional tone did was set contours around the bargaining situation; they ensured that only offers of a level of generosity inconsistent with Hitler's own levels of public bombast were worth pursuing seriously. As a result, no negotiations were undertaken and Churchill avoided the risk of being made to seem more uncompromisingly bellicose than his less confident colleagues might have supported.

On our interpretation of the role of the emotions in bargaining, their status as social conventions enables their expression to be used as early moves in games, ruling out certain outcomes which might otherwise be thought by other parties to be possible equilibria. This can be expected to influence the other party's choice of strategy so long as the structure of the game is such that the other party has a choice at all. Most importantly, by revealing information about the *cardinality* of preferences, deployment of one sort of emotional expression or another can reveal to bargainers that some paths through a game-tree at hand terminate in PDs or other insoluble dilemmas. (Of course, to reiterate, if the game is *simply* a PD to begin with, then the parties are trapped.) Since these are always best avoided if possible (i.e., if strategy vectors leading to alternative regions of the game-space are available), then conveying such information is of potentially critical importance. However, whereas gathering information about ordinal preferences is frequently straightforward, bargainers are often in uncertain positions for inferring preference-intensities merely from propositionally encoded information and situation-types. No adequate set of conventions requiring explicit reference to any and all possible bargaining situations could arise, since the set of such situation-types is, for practical purposes, infinite. However, a finite set of conventions sorting human responses into the types we call 'emotional states' can do the trick, at least up to certain very useful limits. In particular, public departures from these conventions are powerful signals indeed. While commitment devices are *one* means of avoiding social dilemmas, they have a serious flaw: they must be erected in advance of bargaining, and in light of quite detailed foreknowledge concerning the types of games which will likely be played. Perhaps, as Frank argues, evolution endowed us with such devices. However, if we are granted our claim that individual control over emotional states is much more complex than Frank supposes, then it seems to us unlikely that establishing commitment, at least directly (that is, by self-binding, in Elster's sense), is the principal strategic role of the emotions. Furthermore, deliberate signalling of information is a more powerful instrument in complex games than is commitment, because it can be more flexibly brought to bear on games which do not fit familiar patterns. Finally, our account, unlike Frank's, claims no incoherent quarrel (not even a "friendly" one) with the technical apparatus of orthodox game theory. Nothing in our account suggests the technically oxymoronic concept of 'constrained maximization'; neither, however, does it conflict with the sound intuition that the world is better fit for human flourishing to the extent that we have devices which help us to stay out of (as opposed to *get out of*) PDs and their relatives, that is, that we are rational in Nozick's `thick' sense.[6]

Binmore, K. (1994). *Game Theory and the Social Contract, v. 1: Playing Fair.* Cambridge, MA: MIT Press.

Churchland, P.M. *(1979). Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

Danielson, P. (1991). *Artificial Morality.* London: Routledge.

Dennett, D. (1991). *Consciousness Explained*. Boston: Little Brown.

Dumouchel, *P.* (1999). *Emotions: essai sur le corps et le social*. Paris: Synthelabo.

Frank, *R.* (1988). *Passions Within Reason.* New York: Norton.

Gauthier, D. (1986). *Morals By Agreement.* Oxford: Oxford University Press.

Greenspan, P. (2000). `Emotional Strategies and Rationality'. *Ethics* 110: 469-487.

Griffiths, P (1997). *What Emotions Really Are.* Chicago: University of Chicago Press.

Hirshleifer, J. (1987). `On the Emotions as Guarantors of Threats and Promises.' In J. Dupre, ed., *The Latest on the Best*. Cambridge, MA: MIT Press / Bradford, 307-326.

LaCasse, C., and Ross, D. (1998). 'Morality's Last Chance'. In P. Danielson, ed., *Modelling Rationality, Morality and Evolution*. Oxford: Oxford University Press, 340-375.

Lewis, D. (1969). *Convention.* Cambridge, MA: Harvard University Press.

Nozick, R. (1993). *The Nature of Rationality.* Cambridge, MA: Harvard University Press.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotion.* Oxford: Oxford University Press.

Schelling, T. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Skyrms, B. (1996). *The Evolution of the Social Contract.* Cambridge: Cambridge University Press.

[9] We thank Dan Dennett for his helpful comments.