UNIVERSITÉ DU QUÉBEC À MONTRÉAL

"IT'S LIKE I TOLD YOU":
ESSAYS ON THE ECONOMIC ANALYSIS OF INTERPERSONAL COMMUNICATION

DISSERTATION
PRESENTED IN PARTIAL REQUIREMENT
FOR THE DEGREE OF
DOCTORATE IN ECONOMICS

BY

ELI SPIEGELMAN

AUGUST 2012

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

« C'EST CE QUE JE VOUS DIS » : ESSAIS SUR L'ANALYSE ÉCONOMIQUE DE LA COMMUNICATION INTERPERSONNELLE

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN ÉCONOMIQUE

PAR

ELI SPIEGELMAN

AOÛT 2012

To Prisca

ACKNOWLEDGEMENTS:

I gratefully acknowledge the guidance, encouragement and support of my supervisor, Claude Fluet. Many improvements to the work presented here are also due to Raúl López-Pérez, whose opinions and insight extend well beyond the chapter that bears his name. Discussions with Charles Bellemare, Matthieu Chemin, Sabine Kröger, Stephane Pallage, Christian Traxler and members of several conferences have also helped clarify ideas. Jacob Goeree and two anonymous reviewers gave useful insight on Chapter II. Especially helpful were the comments from the members of the defense jury: Bruno Deffains, Claude Fluet, Sean Horan and Pierre Laserre. I thank Vanier College for its policies which allowed me to take time off work to pursue professional development – but keep the office in the meanwhile. The VCTA's generous provision of the staff lounge with fair trade coffee has been instrumental in achieving this work. I also benefitted from financial and material support from CIRPÉE, the latter including statistical software used in the thesis. I have been blessed with three generations of amazing family support: my children, who, mostly uncomplaining, saw so little of me (and piloted the experiment in chapter 2); my parents, in whose reflection I judge all my ideas, even on the rare occasions where we disagree; and my wife, touchstone for what is truly valuable, always ready to meet theory with good sense, who bore the weight of the project in equal measure to myself, and carried it forward with me. Thank you.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Les trois chapitres principaux de cette thèse ont pour point commun l'analyse économique de la communication interpersonnelle en tant que choix sous contrainte. Or, comme on dit, *talk is cheap*. Pourtant, les trois chapitres élaborent trois contraintes différentes qui pourraient s'imposer à la communication interpersonnelle. Dans le premier chapitre, la contrainte est de nature stratégique : en équilibre, les communications différentes suscitent des réactions différentes chez les interlocuteurs. Donc les individus choisissent leur communication en fonction de la réaction qu'ils espèrent susciter. Dans le deuxième chapitre, la contrainte est plus intrinsèque. On y recherche de manière expérimentale dans quelle mesure les gens se contraignent à dire la vérité pour la vérité. Dans le troisième chapitre, la contrainte est encore différente car elle est due cette fois au fait que le communicateur n'a pas pour but de communiquer quoi que ce soit. Il veut, en fait, se donner une idée faussée du sujet. Ainsi, un interlocuteur qui essaierait d'apprendre de la communication doit tenir en comte ces distorsions éventuelles, qui deviennent une contrainte sur la transmission de l'information.

Les trois chapitres s'inscrivent dans un courant d'économie comportementale. Cette mouvance essaie d'ouvrir un peu la « boîte noire » des préférences, en soulevant des questions sur les facteurs qui motivent les choix réels. Ces questions sont dangereuses car elles risquent de supposer ce qu'il fallait démontrer, en réduisant l'explication d'un comportement à la simple volonté de le faire. Pour éviter ce piège, il faut se pencher sur une vision des motivations qu'on croit raisonnables avant de travailler sur les comportements spécifiques. Les prolégomènes de la thèse bâtissent donc un cadre d'analyse qui cherche à soutenir une réflexion précise sur la nature du comportement social qu'on essaie d'interpréter. Ce cadre est appliqué par la suite dans l'élaboration des modèles dans les trois chapitres.

Mots-clés : incitations normatives; communication; jeux; comportement; honnêteté

SUMMARY

The three main chapters of this thesis share the common characteristic that each analyses interpersonal communication as an economic act; that is, as a problem of constrained choice. This raises an immediate problem: if, as they say, talk is cheap, what can the constraints on communication be? The three chapters develop three different kinds of constraint. In the first chapter, the constraint is strategic: in equilibrium, different communications elicit different reactions from the people who receive them. Thus, individuals choose their communication based on the reaction that they hope to elicit. In the second chapter, the constraint is more intrinsic. This paper studies the extent to which people constrain themselves to speak the truth for the truth's sake. The paper describes an experiment that drains a communicative act of nearly all its significance save truth value, and asks subjects to forego monetary gain to preserve their truthfulness. In the third chapter, the constraint is again different, due in this case to the fact that the communicator's goal is not actually to transmit information, but rather to distort her own idea of the truth. Thus, the interlocutor must try to disentangle the speaker's self-deception from the unobservable truth she may have seen. This becomes a constraint on the extent to which information passes between the two individuals.

The three chapters fall into the domain of behavioral economics, construed widely as an attempt to open up somewhat the "black box" of economic preferences, by raising questions about what factors motivate actual choice. These questions can be dangerous, as they risk begging their own question, "explaining" behavior by assuming a preference to engage in it. To avoid this trap, it is important to give some attention to a vision of what might constitute reasonable motivations generally, before working on specific behaviors. The prolegomenon of the thesis outlines a conceptual framework to support more precise reflection on the nature of the social behavior under study. This framework provides a structure which is then applied in the elaboration of the models within the three chapters proper.

Keywords: normative incentives; honesty; communication; behavior; games

PROLEGOMENON:

NORMATIVE INCENTIVES IN ECONOMIC CHOICE: A GENETIC TAXONOMY[1]

Abstract: This essay introduces the concept of a normative incentive, a component of the economic choice process that explicitly reflects what people feel they "should", as opposed to what they "want to" do. It elaborates an analytic framework for normative incentives based on two dimensions of categorization, and illustrates how important normative concepts such as reciprocity and morality seem to cut across the categories defined, suggesting that they may not be monolithic phenomena.

Résumé: Ce papier développe une taxonomie des « motivations normatives ». Il propose que les individus fassent un arbitrage entre les coûts et avantages habituellement pris en compte dans le modèle de l'*homo economicus* et leurs motivations normatives. Celles-ci renvoient à ce que l'individu « pense qu'il devrait faire » par opposition ce qu'il « aurait envie de faire ». La taxonomie proposée est bidimensionnelle. Le premier axe fait référence aux intentions, aux actions ou aux conséquences; le second à la distinction entre motivations intrinsèques et extrinsèques, selon que le sentiment qui les soutient est la culpabilité ou la crainte de la désapprobation sociale. On montre que la plupart des modèles de motivations informelles peuvent être décrits par cette taxonomie.

0.1 Introduction

Economics has a famously impoverished view of human nature. Just what makes the science seem so thoroughly dismal may be the view it takes of the human objects of

---

[1] This essay is an expanded form of Spiegelman (2011).

its investigations: the notorious *Homo economicus*[2]. In its most primal version, this strange species is portrayed as something of a comic book supervillain: the brilliant sociopath, with infinite intellectual capacity aimed only at satisfaction of his own desires[3]. In both aspects (limitless capacity and selfish intent), the caricature is, obviously, not just an illegitimate portrait of human experience, but also an unnecessarily simplistic interpretation of the theory that bore it. Justifications of the model have evolved from Mill's (1874) restrictive argument that economic analysis is limited to domains where people are (a) mostly selfish and (b) capable of determining the optimal way to go about their business. The more current generalizing formulation, as expressed by Friedman (1966), is that (a) the agent's "own desires" can be taken to include whatever richer aspects of human nature the modeler – or more importantly, the economic actor in question – deems relevant, and (b) as long as people behave "as if" they maximized an objective function, it does not matter (to the theory) whether they are actually capable of doing so. It is therefore argued that what the *H. economicus* model provides is not content describing individual behavior, but rather an abstract framework for analysis.

To be at all useful in organizing discussion of social phenomena, the framework must be dressed in content. For instance, it is generally assumed that people prefer more consumption of marketable goods to less, *ceteris paribus*; that they dislike effort; and that future costs and benefits are discounted. It is often also assumed that risky prospects are evaluated differently from sure ones, for given material outcomes. These elements of the content of preferences are probably justified on the grounds of their apparent universal relevance; they are considered to hold nearly as generally as the preference relation itself, though in a less "analogical" fashion. But another

---

[2] Persky (1995) has noted that from its inception, this term has been one of criticism. It was apparently originally coined to differentiate the agents of John Stuart Mill's (1836) economic analysis from actual human beings.

[3] No accident, perhaps, that Oskar Morgenstern (1935) illustrates what we would today call an anti-coordination game with Dr Moriarty following Sherlock Holmes on the train to Dover!

common feature of the "standard" content of preferences emerges: it conforms curiously well to the *Homo economicus* carticature! Due, I believe, to the history of the exclusionary interpretation of *H. econ.*, many economists would agree to a statement along the lines "we are on safer ground assuming that people are basically self-interested than assuming the contrary." Fortunately, this (undeniably dismal) position is often patently false. Perhaps as universal as labor disutility are, for example, the tendency to seek *distinction* (a desire for rank, rather than level, of benefit), the tendency for *reciprocity* (a desire to respond to others in kind to their behavior), the tendency for *altruism* (a desire to help others) and the tendency for *rule-following* (a desire to do the "right" thing). Each of these tendencies has been the subject of a long and deep literature in economics, and it would be beyond the scope of any paper to survey them all. Rather, my point in this essay is to provide a rough taxonomy of the motivations that serve as the mechanisms by which these tendencies arise. The four tendencies above can perhaps be compared to the phenotype of social behavior, based on a "genetic code" of underlying mechanisms. Thus two instances of reciprocity, for instance, may be based on quite different choice mechanisms. Conversely, there may be some underlying similarity of mechanism in the source of several different "phenotypic" regularities listed. My goal is to provide a classification for the "genetic building blocks of" types of mechanisms, which I call normative incentives. The specific variety of mechanism that operates in a specific instance is important because it may determine the empirical predictions the model generates. If so, it will be of significance for testing the theories, as several examples will show.

The point of departure is a basic proposition: *When human actors are aware that their decisions are inter-related with those of other human actors, they trade material costs and benefits off against normative incentives.* In this context, a *normative incentive* is a component of the choice process that makes a certain choice more appealing from a moral, personal or socially constructed perspective. If, in the

standard heuristic, human nature is to calculate and carry out the lowest-cost method of getting what you want, then in modest contrast the heuristic of normative incentives describes people factoring what they feel they should do into the equation.

The proposition that normative incentives are somehow primed by the awareness of the decision-makers that their choices interact with those of other human beings receives substantial support from the empirical literature on social distance, which attempts to experimentally manipulate the degree of this awareness (Leonard 1968; Charness, Haruvy et al. 2001; Dufwenberg and Muren 2006; Rankin 2006; Ahmed 2007; Charness and Gneezy 2008; Hoffman, McCabe et al. 2008; Fiedler, Haruvy et al. 2011). It also implies that the appropriate theoretical construct for their analysis will usually take the form of a *game* of some kind, with the normative incentives providing some structure to specify the payoffs. The structure is chosen to represent social context effects in the interaction. These contextual effects are the "genes" in the biological analogy above, which combine into the unique DNA of a particular interaction. In the following, I propose a framework of sub-categories based on two aspects, or "axes" of the situation: first, the kind of social object to which they apply – intentions, actions or outcomes – and second, the social nature of the incentive – extrinsic versus intrinsic. These classifications correspond to ideas that have emerged in various places in the literature. As I go, I will illustrate the classification with references to some of the major work.

0.2 Axis I: The moments of an interaction

Analysis of the social object to which normative incentives are attached has focused generally on three chronological "moments," or phases, of an interaction. In the first phase, prior to the interaction, the individuals in the interaction all have various *intentions*. The second phase commences once the interaction has begun. Some individuals act (through choices the make), and those *actions* constitute the second phase of the interaction. The third phase is the *outcome* for each individual, which is

produced by the choices made by all the acting individuals coming together. Each phase may be the object of a normative incentive, and together they represent the first axis of variation I will consider. I will say that normative incentives may be outcome-based, action-based or intention-based.

Let us begin with the end: the outcome. Rationality in economics is often characterized as instrumental in the sense that decisions are made so as to guide the actor towards some preferred end. This is certainly the case for *H. economicus*, for instance, who cares only for his own material ends. In general, however, the ends that the actor seeks need be neither material, nor his own. The simplest manner in which normative incentives might enter into consideration of the outcome of an interaction is what has been termed "benevolence", formalised in economic models as long ago as Edgeworth (1880). For instance, suppose two players, i and j, (denoted by subscripts) are interacting. The utility of a benevolent player i might be represented as

$$U_i = x_i + ax_j$$

where x is the material payoff and a is the altruism term showing the strength of the benevolence. Ordinarily, one assumes that $a < 1$ (otherwise i would give all his money to j). Notice also that if $a < 0$, then i can be interpreted as being "spiteful" to j.[4] This utility formulation means that i's preferences over any two values of x_i will generally depend on the vector $x = (x_i, x_j)$. Whether i prefers outcome vector x or outcome y depends, perhaps crucially, on how much j gets from the deal. Notice that this does not give one any grounds to assume that i would not behave as a *Homo*

[4] Throughout, the "natural language" labels I give to various incentives will be, necessarily, as vague as the concepts behind them. Indeed, one finds that in the (economics) literature, the same incentive is often labeled differently in different papers, and different incentives often receive the same label. But these differences are semantic, not essential. As long as the meaning behind the term is clear, the disconcerted reader may, with apologies to Wittgenstein, substitute "bububu" for any term which seems misused.

economicus with regards to the function *U*. However, it does supply more descriptive structure for how people might make decisions in real situations.

Benevolence does not exhaust the possibilities for outcome-based normative incentives. Martin Dufwenberg and co-authors (Dufwenberg and Kirchsteiger 2004; Battigalli and Dufwenberg 2007; Battigalli and Dufwenberg 2009) adapt the basic motivation into an outcome-based model of guilt, in which people care about the material payoffs of those with whom they interact only to the extent that they think those others are disappointed[5]. These are special cases of reference-dependent utility, since the perceived subjective benefit of a given payoff depends on how it is "framed" by the second-order expectations. Such guilt aversion can perhaps be considered to approximate moral codes of appropriate conduct regarding other people. Another model of outcome-based moral behavior is that of Deffains and Fluet (2009), in which the moral code of "do no harm" only restricts the behavior or intentions of the actors inasmuch as these influence the probability of the harmful result.

Models of inequity aversion (Bolton and Ockenfels 2000; Fehr, Klein et al. 2001; Demougin and Fluet 2003) offer the related insight that often what people care about is not (only) the levels of the payoff they receive, but also their *relative* payoff, compared to other players. More generally, rank-based or positional utility (Frank 1985; Clark and Oswald 1998) suggests that individuals care about their standing overall among a potentially large group. Numerous laboratory tests of this kind of preference confirm the effect. Frank (ibid.) also adduces empirical field data to support this claim, suggesting that this positional utility accounts for shallower pay scales than would be otherwise predicted in many industries, as those low in the pay

[5] The payoff that player i thinks player j expected to receive is called i's "second-order expectation" for j's payoff. Of course, one could also define third- fourth- or any higher order of expectation, and indeed these are implicit in Nash equilibria. The difference in models of guilt is that these expectations enter explicitly into the utility function.

scale are "compensated" for their position, and those high in the pay scale "pay for the privilege" with lower material wages. It has also been noted since Keynes (1936) that this kind of comparison-based utility can lead to "arms races" in which people overexert themselves in order to "keep up with the Joneses." This, of course, was one of Veblen's (1899) main insights.

Notice that when the motive is "keeping up with the Jonses," the payoff is no longer material. As Veblen says (p. 75)

> ...it is only when taken in a sense far removed from its naïve meaning that consumption of goods can be said to afford the incentive from which accumulation invariably proceeds. The motive that lies at the root of accumulation is emulation... The possession of wealth confers honor.

The implication of the argument is that the institution of property itself is primarily founded on normative, rather than purely instrumental, preferences. Once again, this is pedigree economics. Alfred Marshall (1994, p. 73) remarks with Nassau Senior that

> Strong as is the desire for [goods consumption], it is weak compared with the desire for distinction: a feeling which if I consider its universality, and its constancy, that it affects all men and at all times, that it comes with us from the cradle and never leaves us till I go into the grave, may be pronounced to be the most powerful of human passions.

Gary Becker (1974) included preferences for the good opinion of one's peers into a formal economic model. Schelling (1974), working in a sociological framework more ideologically amenable to the idea, also proposed an alternate conception of social influence. Akerlof (1980), Bernheim (1994) and more recently Bénabou and Tirole (2006), and Deffains and Fluet (2007) have formulated

asymmetric information models in which the reputation attendant on a certain choice is an endogenous feature. Such models are based on the different equilibrium actions of individuals with different, unobservable characteristics. Individuals tend to "shade" their actions to resemble those of people with more favorable characteristics. An interesting wrinkle on these reputational models can be found in the work on self-esteem by Bénabou and Tirole (2010) and self-signaling by Botond Koszegi (2006). In these models one meets the surprisingly intuitive idea that people don't have perfect access to their own character. As a result, they make their choices partly in order to give themselves evidence that they have some favorable characteristics. Notice that actions motivated by this "taste for reputation" are conceptually distinct from the usual "signaling" models (Spence 1973). In signalling models, the reputation is purely instrumental, whereas here, as Veblen points out, it represents a kind of "consumption" utility all its own.

In summary, "genotypically" outcome-based normative incentives can produce the "phenotypes" of altruism, reciprocity, moral rule-following or distinction. The outcomes of interactions that appear to exhibit these phenotypes can either be material payoffs or beliefs. Among outcome-relevant beliefs, we distinguish between posterior beliefs that are part of the outcome itself, and prior beliefs which serve to frame the evaluation of the outcome which eventually occurs. The uniting feature of all outcome-based normative incentives is that they are *consequentialist* in a strong sense that the process by which a result occurs does not directly matter to the evaluation it receives. It may indirectly matter, as when a person can generate different reputational effects by achieving the same material result in different ways. But all different procedures which generate the same reputational result will, *ipso facto*, be evaluated as equivalent by a person who cares only for reputational outcomes. In this respect, outcome-based normative incentives do not represent a large divergence from the standard models of *H. economicus*. Actions, for instance, are still entirely instrumental in their value.

However, experimental evidence tends to agree with intuitive experience that such strongly consequentalist models are inherently insufficient to completely describe people's preferences in an interaction. People don't only care what happens, they also, for various non-instrumental reasons, care how (action) and for what reason (intention) it happens. That is, they may have preference over changes in the two other moments of an interaction; the acts themselves, and the intentions of the actors leading into the interaction.

Given the importance of consequences in economics, the proposal that *actions* have costs and benefits, independently of their consequences, is surprisingly common. The disutility of labor, for instance, is based on an action, not a consequence. Even consumption benefit, in fact, is not really an outcome, but rather an action.[6]. To build it into a consequentialist model of behavior requires a weaker kind of consequentialism, in which the "consequences" are expanded to include the process by which they are attained. In other words the consideration of the disutility of work requires an implicit formulation along the lines of "I prefer to eat a lot without having worked to get the food."[7] Formally, this is identical to the procedure by which we can consider act-based, non-consequentialist normative incentives. It is merely the source of the utility which changes.

If altruism seems particularly linked to outcome-based models, the idea of rule-following seems well suited to models of act-based normative incentives. This relates to the idea of social norms, which have been the subject of a rather extensive literature (Elster 1989; Bicchieri 2002; López-Pérez 2008; Tammi 2008; Adena 2011). One of the earliest formal introductions of act-based normative incentives was

---

[6] Mill, , p. 321, recognized this in his formulation of the original *Homo economicus*, arguing that political economy is concerned with man "solely as a being who desires to possess wealth, ... [and] makes entire abstraction of every other human passion or motive; *except ... aversion to labour, and desire of the present enjoyment of costly indulgences*" (emphasis added).

[7] Notice that part of the problem with the analysis comes from modeling a fundamentally dynamic phenomenon (action) with a static design.

the work of Andreoni (1989; 1990), who described a "warm glow" of extra utility that agents feel in addition to material consequences of performing some "good" act. López-Pérez (2008) introduces an elegant mechanism in extensive games, by which the final payoff is adjusted to take account for the normative impact of the actions taken in the history of that terminal node. As with outcomes, it may be the case that beliefs can frame the evaluation of actions. López-Pérez and Spiegelman (forthcoming) consider an application of the guilt aversion addressed in an outcome-based model by Charness and Dufwenberg (2006) to act-based models. They consider two players A and B, where A has an incentive to lie to B. The model focuses on lies, and predicts that A will refrain from lying only if A thinks that B expects the truth. Generalized to any action, act-based guilt aversion yields a model of conditional norm-following, much like that described in Bicchieri (2006).

One might have expected that a primary normative incentive based purely on the act itself, with no regard for the consequences, would be morality. Indeed, that is usually what is meant by "deontological" concerns. Considering morality as a non-consequentialist incentive corresponds to Sen's (1977) argument that principled behavior is counter-preferential: the outcomes involved are not the deciding factor. White (2004) elaborates general arguments about the form that Kantian morality should take in preferences, and Karni and Safra (2002) characterize a utility representation of justice. Deffains and Fluet (2009) present a model where agents suffer disutility when they transgress a moral code comparable to Kant's categorical imperative. Brekke, Kverndokk et al. (2003) have agents suffering disutility as their actions diverge from a social-welfare-maximizing level. Kaplow and Shavell (2007) develop a model of moral guilt and pride that keeps agents from taking some specified harmful acts. More on the mechanics of integrating morality into the utility function can be found in Spiegelman (2011).

One difficulty with treating morality from an economic perspective arises from the tension it creates with economics' normative positions. It is not clear that the common strategy of equating moral strictures with utilitarian social benefit is justified. Problems such as the footbridge-and-trolley dilemma (Thomson 1976; Foot 1978) show that moral intuition goes beyond material outcomes. As Sen (1977) recognized, counterpreferential choice drives a potential wedge between the concepts – identical in the standard framework – of goal-oriented behaviour and welfare maximization. This wedge opens the "serious" questions that have to be answered before normative implications can be teased out of theoretical predictions (Hausman and McPherson 1993). The theorist and policy maker must make decisions about their own ethical positions, decisions which are implicit in the structure of the utility-based rules in the models above. For instance, Shiell and Rush (2003) find evidence suggesting that stated willingness to pay is influenced by "commitment" as well as by the more "consequentialist" considerations. The extent to which these should be considered in the cost-benefit analysis is an open question. Is it legitimate to maintain a utilitarian social welfare function when individuals are constrained by moral rules? The answer depends on the metaphysical nature of the rule, and the nature of the constraint. If the "true" ethic is deontological, then social welfare must be recast in terms of violations of the rule. If the "true" ethic is utilitarian, and the rule is mostly a "just" method of achieving it, then the basic social welfare principles of standard economic modeling remain justified.

Economic models have, for the most part, assumed the latter. It seems clear that moral behavior is at the very least not completely independent from social welfare. Kant's (2005 [1785]) Categorical Imperative (CI) requires choosing a "maxim" or rule of action that one could, at the same time, wish to become a universal law. This suggests socially optimal behaviour, and has been taken in several studies to imply it. For instance, Brekke, Kverndokk et al. (2003) identify the Kantian ideal with the efficient production of a public good in their model of moral behaviour, which leads

to an equivalence between Kantian and Benthamite morality! Similarly, Kaplow and Shavell (2007) assume that a "policy maker" chooses the "guilt" or "pride" associated with a certain action to minimize the harm those actions cause, subject to various constraints[8], and suggest that, in fact, existent moral codes seem to behave "as if" they were so constructed. Bilodeau and Gravel (2004) show the equivalence between social optimum and categorical imperative-driven outcomes is not general. It does hold in public goods games, and other cases of similar structure, but in general there may not even exist any Categorical Imperative. For instance, in a "matching pennies" game there is no rule that everyone could follow, while at the same time wishing that everyone else would do the same.

I have focused here on the rule-following phenotype. However, it is quite apparent that reciprocity could also be sparked by act-based incentives. Indeed, the Chapter I of this thesis develops just such a model. Summarizing act-based normative incentives, we see they can stem from (a) "pure" aversion to the act in question (potentially based on constraint to follow some "moralistic" rule), (b) the actor's interpretation of other people's prior expectations (guilt aversion) or (c) from observers' approval or disapproval of the act itself, in which case, as in reputational concerns, the posterior beliefs of other people are the source of the utility. Indeed, the difference between stigma and disapproval is subtle. For disapproval to be operative, the source of the approval must be the act itself, and not the resultant inference. The admonition to "love the sinner, hate the sin" reflects the difference. Although the theoretical difference is clear, in many cases it may be difficult to distinguish empirically between stigma and disapproval as motivations. One empirical foothold may be that disapproval of the action itself, as an impersonal effect, should be relatively invariant to social distance. Stigma, on the other hand, may be much more

---

[8] They note that this harm can, in principle, refer to non-material outcomes such as rights violations (an example of the generalizing solution to the *H. economicus*).

keenly felt when the interacting people are less anonymous. For instance, the effects of experimental treatments that alter social distance might be interpreted as identifying disapproval (base effect) and stigma ("slope").

The final moment of the interaction is the intention of the interacting people going in. Informally, there can be no question that intentions matter. To give a few examples:

- Someone cuts ahead of you in line at the cinema. Your reaction will be different, depending on whether you think the person didn't see you, or whether they intentionally ignored you.
- A new acquaintance doesn't return a phone call. Are they busy, or are they avoiding you?
- One of the key requirements in labor negotiations is often that the parties feel they are negotiating "in good faith". This comes down to whether they really intend to find middle ground, or just to push through their preconceived expectations.

In all of the above examples, the actions and outcomes are the same, and yet a "reasonable" evaluation of the behaviour varies widely with the perceived intention. A significant difference between intention-based incentives and act- or outcome-based incentives is that while outcomes are observable more or less by definition, and actions may well be observable, unless they are hidden by the actor, intentions are, as a general rule, unobservable, and so will usually have to be inferred. This inference can be extracted mostly through observable signs, i.e., through actions or outcomes. As a result, models of intention-based normative incentives must specify how people use observable outcomes and/or actions to infer intentions. Indeed, they must specify exactly what an "intention" is. In general, it seems that a person's intention is closely related to the goal they are trying to attain. In other words, to some extent a person's *intentions* may be the same thing as their preferences. Following this line of thought, a model of intention-based normative incentives would involve people who

intrinsically cared what other people's preferences were. This is the approach developed by Levine (1998) and Rotemberg (2008). The specific intention is a generalized altruism, in which (in Levine) the a parameter noted above can be positive (altruism) or negative (spite). Notice that this means that the perceived intention in Levine (1998) is essentially same thing as a reputation, a posterior belief about an unobservable type based on the equilibrium distribution of actions. The model nevertheless is *not* an outcome-based reputational model, because the player whose reputation is established does not get any direct benefit from it. The reputation provides them with instrumental benefit, because the normative incentive acting on the others (who assess the reputation) leads these others to act favorably towards those whom they perceive to have good intentions, and unfavorably towards those whom they perceive to be spiteful. Rotemberg's formulation formalizes an emotional response (anger) which is triggered when the perceived type of the interacting agent is below a certain threshold. Actions suggesting "good" (altruistic) characteristics lead to esteem, and this esteem generates a material benefit. Actions suggesting "bad" (spiteful) characteristics lead to stigmatization, and generate a material harm. However, the normative incentive is the component of the system that leads people to engage in this "reciprocal" behavior, and this depends on the perceived intentions of the actor.

This tendency to model the *character* of intention-based normative incentives as a kind of reciprocity is pervasive (Rabin 1993; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006)[9]. As a general term, reciprocity is usually defined as the tendency of people to do unto others as they have had done unto themselves

---

[9] One exception is Battigalli and Dufwenberg (2009), which includes an intention-based theory of guilt. In this model, player A will be more generous with player B if A's second-order expectations are that she believes player B thinks player A had bad intentions, regardless of what A thinks B's intentions are. Thus there is no reciprocity, but (second-order) perceived intentions do enter the deliberative process.

(Dohmen, Falk et al. 2009)[10]. As shown above, this definition does not imply that reciprocity is intention-based. In principle, one could define a reciprocal tendency over outcomes (so that A will try to give B about the same outcome that B gave A, regardless of how that comes about, or why A believes B did what she did), or over acts themselves (in which case, if B did x to A, then A will tend to want to do x back, regardless of the consequences or of the interpreted intentions). However, as an empirical matter, it seems that intention-based models of reciprocity offer significantly more explanatory power than other models. In empirical settings, perceived intentions have been identified by comparing reactions to human players with reactions to computerized players in a dictator game (Falk, Fehr et al. 2003), by direct elicitation of beliefs (Falk and Fischbacher 2006), and by modifying the action set possible (McCabe, Rigdon et al. 2003; Cox and Deck 2006; Falk, Fischbacher et al. 2008; Hoffman, McCabe et al. 2008). It has been robustly suggested that people's preferences include consideration of the intentions of the other people with whom they interact.

The models described above are all normative because they explicitly include considerations of interests that go beyond personal preferences. Outcome-based normative incentives dictate how we "should" respond to different posterior beliefs about payoffs or unobservable types, or what kinds of posterior beliefs we "should" try to instil. Act-based normative incentives dictate how we "should" respond to certain actions, or what actions we "should" take. Finally, intention-based incentives dictate how people "should" think of or act towards each other, and how we "should" respond when they do or don't.

---

[10] The literature on reciprocity has developed into a very large field of its own, and in the process distinctions have arisen. For instance, positive reciprocity – rewards for "good" behavior – has been distinguished from negative reciprocity – punishment for "bad" behavior. In a similar vein, one can distinguish between weak reciprocity – which is restricted to good or bad behavior directed at the reciprocator herself – and strong reciprocity, in which the reciprocator may reward or (more often) punish a third party

To illustrate, one subject that has received particular attention is dishonesty.[11] It is frequently observed that people do not lie as much as a naive interpretation of *H. economicus* would predict. Why not? One can distinguish, in principle, between several different reasons based on the axis of moments elaborated above. For instance, lies have an outcome, which is sowing false beliefs in others. If people feel "bad" about this, or about the subsequent decisions that the deceived other might make, then such outcomes will attenuate any gain that might be had from lying. On the other hand, "pure" lie aversion posits a disutility experienced merely by uttering an untruth. (Lundquist, Ellingsen et al. 2007; Kartik 2009; Lundquist, Ellingsen et al. 2009). One might say that people suffer from some sort of "cognitive dissonance" when they make statements they believe to be untrue, or that it is inherently displeasing to them to violate a social or moral rule against lies. López-Pérez and Spiegelman (2011) test a theory that players feel bad for lying only if they believe that the person being lied to expected the truth. This theory, for instance, explains the exculpable nature of the "bluff" in poker, and is an application of models of guilt (Battigalli and Dufwenberg 2007; Battigalli and Dufwenberg 2009) to act-based, rather than outcome-based incentives. Finally, it should be noted that an inadvertent lie – in which person *A* may say something to person *B* that he believes (in error) to be true – does not have nearly the normative force of an intentional one. Thus the intention to cause others to believe things that the speaker does not believe seems to be a key part of what lying means. This is also consistent with the acceptability of the poker bluff. These different models of lie-aversion have quite different empirical implications, which could be susceptible to experimental manipulation.

---

[11] See Gneezy (2005), and attendant literature, including the modest contribution in the second chapter of this thesis (López-Pérez and Spiegelman, 2011).

0.3 Axis 2: Intrinsic versus extrinsic

There are other alternative classifications for incentives which can be made. For instance, several authors distinguish *intrinsic* from *extrinsic* incentives (Kreps 1997). It should be noted that, all incentives are fundamentally intrinsic. Even *Homo economicus* has an intrinsic desire for consumption and leisure, which mediates the way he interacts with his environment. On the other hand, no motivation is wholly an island. People always seek information from their environment to help them determine how to apply their standards of behavior. In practice, many authors refer to extrinsic incentives as material incentives, and intrinsic incentives as a one of a variety of normative incentives. I will categorise normative incentives as intrinsic or extrinsic in the following way. If a person requires information about the outcome, action or intention to be transmitted to others in order for the incentive to bind, then it is extrinsic. If the incentive binds even when no information is transmitted to others, then it is intrinsic. The definition that the dichotomy rests on the importance of information to other people reflects the fundamentally social aspect of normative incentives (and human behaviour). It has the advantage of separating incentives which can be manipulated through dissimulation or exaggeration from those which cannot, which in turn has useful empirical implications.

The main extrinsic normative incentives are esteem versus stigmatization (outcome), approval versus disapproval (act) and some kinds of reciprocity (intention). Intrinsic incentives include personal and moral opinions, and social norms that have been internalized. Guilt and pride, moral outrage or duty, benevolence, self-esteem and ego-utility and aversion to norm-breaking are intrinsic incentives. Models of inequity aversion and positional or rank-based utility appear to pose something of a challenge to these definitions, since in these cases the incentive is defined only in terms of the outside relationships. Such considerations show again the empirical usefulness of the classification scheme. The question, which can be experimentally identified, turns upon a simple point: must information pass to others in order for the

incentives to bind? To the extent that players cannot diminish their rank-based utility or disutility (which is an internal opinion) by hiding their rank from others, the incentives are intrinsic. The distinction therefore permits a closer analysis of the phenomenon, both theoretically and in its empirical predictions. The table below classifies some of the more common normative incentives along the two axes I propose.

Table 0.1 Normative incentives in the taxonomy

	Intrinsic	Extrinsic
Intention (I)	Guilt-from-blame	Intention-based Reciprocity; Spite
Action (A)	Cost of lying; moral concerns; act-based guilt aversion	Act-based reciprocity, disapproval
Outcome (O)	Simple guilt; benevolence; self-esteem; inequity aversion; rank-based utility	Esteem, reputation

0.4 Conclusion

The classification system that I propose in this paper diverges from the standard interpretation of "economic man", but more importantly it diverges from the main sources of normative feeling that are observed in social interactions. I argue that familiar goals such as distinction, reciprocity, altruism and rule-following are analogous to phenotypes, outward expressions of motivational mechanisms that can be profitably explained at a "genotypic" level. The genotypic and phenotypic classification systems do not have a one-to-one relationship, in general. Distinction may be intrinsic or extrinsic, but is largely outcome-based. An intention-based distinction might be conceived as a motivation to, for example, moral one upmanship. This could be empirically discriminated from the (outcome-based) incentive to have the greatest moral reputation by testing to see whether it was sensitive to privacy.

Considering altruism, outcome-based motivations could be empirically disentangled from non-consequentialist motivations by cutting any sure link between altruistic actions and their expected results. If altruism is intention-based, for example, "it's the thought that counts", so a costly action that ended up having no impact would fulfill the obligation that the normative incentive generates. The distinction between extrinsic and intrinsic altruism can be considered as a parallel to Sen's (1977) discussion of commitment and sympathy. In the latter, decision-maker *A* chooses to help decision-maker *B* because *B*'s welfare gives benefit to *A*. Such sympathy is described by the altruism model that goes back to Edgeworth, cited earlier. In cases of commitment, by contrast, *A* helps *B* even though there is no personal benefit from doing so. In Sen's example, appeals to send aid for a famine would be more effective on the sympathetic if they contain information about the suffering of the hunger-stricken. The committed, on the other hand, would give anyway.

Reciprocity and rule-following, for their parts, can bind at intentions (I), actions (A) or outcomes (O). Examples of familiar moral rules of the three kinds include: do no harm (O); do not kill (A); and do not lie (I). The classification of the last two can be justified with the claim that even killing accidentally is traumatic; however, the culpability from lying is tied to the intention to instil a false belief – lying by accident carries no blame. Of course, the relevance of these motivations in any particular case is an empirical issue. Indeed, the principal use of the taxonomy may be to generate empirical predictions such as those above. To take the example of reciprocity, suppose some process generates an allocation in which *A* receives more than *B*, and then *B* has the opportunity to generate allocations that are equal (no reciprocity) or that favor *B* (reciprocity). If the initial process were a lottery, it would incite reciprocity in case (O), but less in case (I) and (A). If the initial process were one which *A* chose over another, the availability of such alternative courses of action might affect reciprocity in case (A) or (I) but not in case (O).

I have found the "genetic taxonomy" elaborated in this paper to be a useful analytical tool for developing models of normative incentives. Its use for identifying empirical predictions has also been demonstrated, for instance, in Chapter III of this thesis. Economics has resisted the explicit introduction of normative incentives in large part because of the well known fact that any observed behavior can be explained by assuming the "right" utility function. My goal here has been to try to establish some guidelines for determining whether the "right" utility function is really right. The development of experimental methods in economics has from the start held promise of this sort of identification, but for that to work, the experiment should be based on clear predictions. The taxonomy I propose supplies such predictions, and will, I hope, therefore be useful in the further work of identifying a more behaviorally descriptive content to the kind preference relation that guide economic choice in *Homo sapiens*.

Bibliography

Adena, M. (2011). Accounting for norms in game theoretical models. Berlin, Department of Economics, Free University of Berlin.

Ahmed, A. M. (2007). "Group identity, social distance and intergroup bias." *Journal of Economic Psychology* **28**(3): 324-337.

Akerlof, G. (1980). "A theory of social custom, of which involuntary unemployment may be one consequence." *Quarterly Journal of Economics* **94**(4): 749-775.

Andreoni, J. (1989). "Giving with impure altruism: applications to charity and Ricardian equivalence." *Journal of Political Economy* **97**(6): 1447-1458.

Andreoni, J. (1990). "Impure altruism and donations to public goods: a theory of warm-glow giving?" *Economic Journal* **100**(401): 464-477.

Battigalli, P. and M. Dufwenberg (2007). "Guilt in games." *American Economic Review* **97**(2): 170-176.

Battigalli, P. and M. Dufwenberg (2009). "Dynamic psychological games." *Journal of Economic Theory* **144**(1): 1-35.

Becker, G. (1974). "A theory of social interactions." *Journal of Political Economy* **82**(6): 1063-1093.

Bénabou, R. and J. Tirole (2006). "Incentives and prosocial behavior." *American Economic Review* **96**(5): 1652-1678.

Bénabou, R. and J. Tirole (2010). Identity, Morals and Taboos: Beliefs as Assets. *IDEI Working Papers*, Institut d'Économie Industrielle (IDEI), Toulouse: 49 p.

Bernheim, B. D. (1994). "A theory of conformity." *Journal of Political Economy* **102**(5): 841-877.

Bicchieri, C. (2002). "Covenants without swords: Group identity, norms and competition." *Rationality and Society* **14**(2): 192-228.

Bicchieri, C. (2006). *The Grammar of Society* New York, Cambridge University Press.

Bikhchandani, S., D. Hirshleifer, et al. (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades." *Journal of Political Economy* **100**(5): 992-1026.

Bilodeau, M. and N. Gravel (2004). "Voluntary provision of a public good and individual morality." *Journal of Public Economics* **88**(2004): 645-666.

Bolton, G. and A. Ockenfels (2000). "ERC: A theory of equity, reciprocity and competition." *The American Economic Review* **90**(1): 166-193.

Brekke, K. A., S. Kverndokk, et al. (2003). "An economic model of moral motivation." *Journal of Public Economics* **87**(9-10): 1967-1983.

Charness, G. and U. Gneezy (2008). "What's in a name? Anonymity and social distance in dictator and ultimatum games." *Journal of Economic Behavior & Organization* **68**(1): 29-35.

Charness, G., E. Haruvy, et al. (2001). Social Distance and Reciprocity: The Internet vs. the Laboratory. *University of California at Santa Barbara, Economics Working Paper Series*, Department of Economics, UC Santa Barbara.

Clark, A. E. and A. J. Oswald (1998). "Comparison-concave utility and following behavior in social and economic settings." *Journal of Public Economics* **70**(1998): 133-155.

Cox, J. C. and C. A. Deck (2006). "Assigning intentions when actions are unobservable: the impact of trembling in the trust game." *The Southern Economic Journal* **73**(2): 307-314.

Deffains, B. and C. Fluet (2007). Legal versus normative incentives under judicial error. *CIRPEE Working Paper 07-18*: 32 pgs.

Deffains, B. and C. Fluet (2009). Legal liability when individuals have moral concerns. *CIRPEE Working Paper 09-51*: 42 pgs.

Demougin, D. and C.-D. Fluet (2003). Inequity Aversion in Tournaments, SSRN.

Dohmen, T., A. Falk, et al. (2009). "Homo reciprocans: Survey of evidence on behavioral outcomes." *The Economic Journal* **19**(March , 2009): 592-612.

Dufwenberg, M. and G. Kirchsteiger (2004). "A theory of sequential reciprocity." *Games and Economic Behavior* **47**: 268-298.

Dufwenberg, M. and A. Muren (2006). "Generosity, anonymity, gender." *Journal of Economic Behavior & Organization* **61**(1): 42-49.

Edgeworth, F. Y. (1880). *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences* London.

Elster, J. (1989). "Social norms and economic theory." *Journal of Economic Perspectives* **3**(4): 99-117.

Falk, A., E. Fehr, et al. (2003). "On the nature of fair behavior." *Economic Inquiry* **41**(1): 20-26.

Falk, A. and U. Fischbacher (2006). "A theory of reciprocity." *Games and Economic Behavior* **54**: 293-315.

Falk, A., U. Fischbacher, et al. (2008). "Testing theories of fairness -- Intentions matter." *Games and Economic Behavior* **62**(1): 328-303.

Fehr, E., A. Klein, et al. (2001). Theories of fairness and reciprocity: evidence and economic applications. *Working Paper*, Institute for Empirical Research in Economics, University of Zurich.

Fiedler, M., E. Haruvy, et al. (2011). "Social distance in a virtual world experiment." *Games and Economic Behavior* **72**(2): 400-426.

Foot, P. (1978). The Problem of Abortion and the Doctrine of the Double Effect. *Virtues and Vices*. Oxford, Basil Blackwell.

Frank, R. H. (1985). *Choosing the Right Pond: Human behavior and the quest for status*. Toronto, Oxford University Press.

Friedman, M. (1966). The methodology of positive economics. *Essays in Positive Economics*. Chicago, University of Chicago Press: pp. 3-16, 30-43.

Hausman, D. M. and M. S. McPherson (1993). "Taking Ethics Seriously: Economics and Contemporary Moral Philosophy." *Journal of Economic Literature* **31**(2): 671-731.

Hoffman, E., K. McCabe, et al. (2008). Chapter 49 Social Distance and Reciprocity in Dictator Games. *Handbook of Experimental Economics Results*, Elsevier. **Volume 1:** 429-435.

Kant, I. (2005). *Groundwork for the Metaphysics of Morals*. Peterborough, ON, Broadview Press.

Kaplow, L. and S. Shavell (2007). "Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System." *Journal of Political Economy* **115**(3): 494-514.

Karni, E. and Z. Safra (2002). "Individual sense of justice: a utility representation." *Econometrica* **70**(1): 263-274.

Kartik, N. (2009). "Strategic Communication with Lying Costs." *Review of Economic Studies* **76**(4): 1359-1395.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. Cambridge, Macmillan Cambridge University Press.

Koszegi, B. (2006). "Ego utility, overconfidence and task choice." *Journal of the European Economics Association* **4**(4): 673-707.

Kreps, D. M. (1997). "Intrinsic motivation and extrinsic incentives." *American Economic Review* **87**(2): 359-364.

Leonard, B. (1968). "Responsibility, reciprocity, and social distance in help-giving: An experimental investigation of english social class differences." *Journal of Experimental Social Psychology* **4**(1): 46-63.

Levine, D. (1998). "Modeling altruism and spite in experiments." *Review of Economic Studies* **1**(3): 593-622.

López-Pérez, R. (2008). Introducing social norms in game theory. *Games, Rationality and Behavior*. A. Innocenti and P. Sbriglia, Palgrave McMillian: 26-46.

López-Pérez, R. and E. Spiegelman (2011). Why do people tell the truth? Experimental evidence for pure lie aversion, Universidad Autonoma de Madrid.

Lundquist, T., T. Ellingsen, et al. (2007). The cost of Lying. *SSE/EFI Working Paper Series in Economics and Finance*. Stockholm, Stockholm School of Economics: 40.

Lundquist, T., T. Ellingsen, et al. (2009). "The aversion to lying." *Journal of Economic Behavior & Organization* **70**(1-2): 81-92.

Marshall, A. (1994). *Principles of Economics*. Philadelphia, Porcupine Press.

McCabe, K., M. Rigdon, et al. (2003). "Positive reciprocity and intentions in trust games." *Journal of Economic Behavior & Organization* **52**(2003): 267-275.

Mill, J. S. (1874). *Essays on Some Unsettled Questions of Political Economy*, Library of Economics and Liberty.

Persky, Joseph. "Retrospectives: The Ethology of Homo Economicus." The Journal of Economic Perspectives, Vol. 9, No. 2 (Spring, 1995), 221-231

Rabin, M. (1993). "Incorporating fairness into game theory and economics." *American Economic Review* **83**(5): 1281-1302.

Rankin, F. W. (2006). "Requests and social distance in dictator games." *Journal of Economic Behavior & Organization* **60**(1): 27-36.

Rotemberg, J. J. (2008). "Minimally acceptable altruism." *Journal of Economic Behavior & Organization* **66**(3-4): 457-476.

Schelling, T. (1974). *Micromotives and macrobehavior*. New York, W.W. Norton & Co.

Sen, A. K. (1977). "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* **6**(4): 317-344.

Shiell, A. and B. Rush (2003). "Can willingness to pay capture the value of altruism? An exploration of Sen's notion of commitment." *Journal of Socio-Economics* **32**(2003): 647-660.

Spence, M. (1973). "Job Market Signaling." *Quarterly Journal of Economics* **87**(3): 355-374.

Spiegelman, E. (2011a). Moral decision-making and self-deception: theoretical premises and grounding observations. *Mimeo*.

Spiegelman, E. (2011b). "The (Human) Nature of economic choice." *Ithaque* Special issue: Colloque de ADEPUM. http://www.revueithaque.org/colloque-de-ladepum/

Tammi, T. (2008). Do norms and procedures speak louder than outcomes? An explorative analysis of an exclusion game. *Discussion Papers*, Department of Economics, University of Joensuu.

Thomson, J. J. (1976). "Killing, letting die, and the trolley problem." *The Monist* **59**(April, 1976): 204-217.

Vanberg, C. (2008). "Why do people keep their promises? An experimental test of two explanations." *Econometrica* **76**(6): 1467-1480.

White, M. D. (2004). "Can *homo economicus* follow Kant's categorial imperative?" *The Journal of Socio-Economics* **33**(2004): 89-106.

CHAPITRE I

PRIDE, PREJUDICE, AND THE CREDIBILITY OF CHEAP TALK: THEORY AND EXPERIMENTAL EVIDENCE.

Abstract: This paper presents a simple application and experimental test of a theoretical model of discriminatory social preferences. First, it establishes conditions under which a combination of "guilt and pride" can support an equilibrium in which costless messages engender cooperative strategies in a one-shot Prisoners' Dilemma (PD) game. The main result is a signaling equilibrium with full separation of signals, and cooperation whenever the social distance is less than a certain threshold level.

The second part of the paper is an experimental application and test of a discrete version of the theory. Using the minimal group paradigm, subjects are put into a set of interlocking groups. 116 subjects played 4 games each (within-subjects) under different informational (treatment) contexts. Results are broadly consistent with the theory: people cooperate more when faced with group members. They are also less "honest" with their messages when messages are sequential than when they are simultaneous, which suggests a strategic motivation to the messages.

Résumé: Ce chapitre présente un modèle et un test expérimental d'une théorie de préférences sociales discriminatoires. Le premier consiste en un jeu de dilemme de prisonnier (DP) où l'avantage de faire défection est affecté par le sentiment de culpabilité que donne au joueur le fait de faire du mal à autrui. Si l'on ajoute que ce sentiment diminue avec la distance sociale perçue entre les joueurs, on peut définir des groupes sociaux comportementaux comme les individus suffisamment proches en distance sociale pour qu'ils coopèrent dans le DP. Le résultat principal théorique consiste à un équilibre de signaux où les individus révèlent leur véritable distance sociale en « cheap talk » simultanée, avant de jouer.

La deuxième partie du papier construit un test expérimental de la théorie, en appliquant le paradigme des groupes minimaux (PGM). 116 sujets ont été assignés à des groupes qui reflètent la structure du modèle théorique et ont joué au DP sous 4 traitements, qui varient dans l'information donnée sur l'identité de l'opposant. Les prédictions du modèle sont globalement confirmées.

1.1 Introduction

Ich bin ein Berliner!

- John F. Kennedy, June 26, 1963

The door-to-door salesman, having noticed the tricycle in your front yard, pulls a well-thumbed snapshot of his own children from his wallet for your admiration. What makes him think this increases the chance of a sale? At a conference overseas, you unexpectedly meet someone from your own university. Although you hardly speak when you work two floors apart, at the conference you go out to dinner together. Why? The "human library" allows members of the public "borrow" people – gays or blacks or single parents or the homeless – and have a conversation with them for a certain period of time in an attempt to fight prejudice against these groups Human library web site, (2010). Why does this attempt seem intuitively plausible? The unifying thread that runs through all these examples is "common ground" – shared personal characteristics. When two people realize that they both are parents, or that they are both from the same city, it generates or amplifies a sense of solidarity between them. In the case of the human library, one might think that any two people, if they spend time talking, will discover some basis for identification, which will engender a kind of sympathy born of solidarity.

If this solidarity is strong enough, then it can induce sacrifices by individuals in the group for the benefit of other group members. Thus communication can, *by revealing common characteristics*, engender solidarity and promote pro-social behaviour. In order for this process to work, individuals must have some way to credibly communicate their group membership[12]. Modern humans use a variety of methods to signal the groups to which they feel some affiliation: hairstyle, clothing

---

[12] They must also in fact share some common characteristics – if communication reveals no commonalities, all bets are off! Some implications of this will be discussed below.

choice, tattoos, even the operating system you run on your computer can be seen, and in some way intended, as signals of affiliation with some social groups. (Are you a Mac person, or a PC person? Or are you a Linux person?) Many of these signals are costly in some way, so give evidence of a vested interest in the group. But many, including probably the richest, simple verbal communication, are also fraught with a certain inherent "cheapness".

The theoretical part of this paper establishes conditions under which cheap signals are credible enough to engender cooperative strategies in a one-shot Prisoners' Dilemma (PD) game. It is well documented that pre-game communication increases cooperativeness in this context. Overviews of the literature include those of Ledyard (1995), Sally (1995), and Crawford (1998). This paper runs somewhat counter to most of this literature in focusing on communication about *type*, rather than communication about *intentions*.[13] It thus works at the intersection of cheap talk and another venerable literature, that of in-group bias. Empirical results from social psychology robustly show that people behave more favorably towards those with whom they can identify a common group membership, although the mechanism by which this works is not entirely clear. This paper was not designed to elucidate this mechanism. Rather, I focus on two other limitations to this empirical literature as it currently stands. First, all the studies of which I am aware let the group membership information remain exogenous since subjects in the experiment, or agents in the model, always know for sure the group membership of the other people with whom they interact. One contribution of the model presented here is to relax this assumption, and let the credibility of the group membership information be an endogenous (equilibrium) choice by the interacting agents. Groups are unobservable

---

[13] It might be noted in this context, along the lines of Rabin (1990) that in many cases type implies behavior, and expectations depend on perceived type. Thus type and expectations cannot always be truly separated.

in this paper, and identifiers are cheap talk. Second, previous models tend to be characterized by exclusive group membership (Chen and Li 2009). A player who belongs to one group cannot simultaneously belong to other groups, as well. On the other hand, casual introspection and models such as (Wichardt 2008) highlight the insight that often people have many group affiliations simultaneously, which may place different or conflicting demands on behavior, causing choices to be highly dependent on social context. (Spiegelman 2009) showed that altruism could not be an effective force to ensure that cheap messages were credible in the presence of such exclusive unobservable groups. The current study instead develops the concept of social "neighborhoods," overlapping regions within any one of which (the theory predicts) that individuals will cooperate. These suggest interlocking sets of groups such that the in-group bias behavior occurs between any individuals who come from the same neighborhood.

While the mechanism behind in-group bias remains obscure (Guth, Ploner et al. 2008), what does seem clear is that, if one wishes to maintain the rational choice framework to explain behavior, then considerations beyond material self-interest must be involved. This literature, and extensions of it by experimental economists, shows that people are willing to sacrifice some expected material well-being for the benefit of others when faced with so-called "in-group" opponents than when faced with "out-group" others. Squaring such behavior with an assumption of expected *utility* maximization has implicated a broad class of considerations known as normative incentives. Normative incentives are factors that affect a player's decisions, but not her payoff. They are often interpretable as what she feels she "should" do from a personal, moral or socially constructed perspective, which is the source of the name. Spiegelman (2011) develops a conceptual framework for such normative incentives according to whether they bind at the level of *outcomes* (including posterior beliefs), *actions* themselves, or *intentions* (prior beliefs about types or actions). As an example of the difference, simple benevolence is an outcome-

based normative incentive: one player gains utility from co-players' payoffs. A "warm glow," (Andreoni, 1989) or moral utility by contrast, comes from the *act* of choosing cooperative behavior, regardless of the consequences, and represents another kind of incentive.

It should be a familiar idea to economists that many different competing impulses live simultaneously within the human heart. This is, after all, the heart of the idea of a tradeoff. Robinson Crusoe wants both to eat and to sleep, in many formulations independently from each other. A monopolist wants to sell many units, but also to sell them at a higher price. The principal wants to reduce the agent's payment, but also to give him an incentive to work. An indifference curve "exists" because the agent whom it describes wants both more x_1 and more x_2. A central tenet of my work – and I am of course far from alone in this – is that normative incentives can be analysed in a very similar way to material ones. In the current model, I consider a continuum of agents occupying the perimeter of a unit circle. The arc distance between them is a measure of "social distance". In-group bias is explained as the effect of two complementary normative incentives that operate in the "social landscape" of an arc. I dub the two normative incentives "guilt" and "pride". Guilt is the disutility that agents feel when defecting in a PD game. This can be envisioned as negative affect due to causing harm to another person for material gain. I assume that the negative affect decrease with the social distance between the actor and the object of the harm. The second, competing normative incentive is closer to the "sinful" *orgueil* than to *fierté*. Rather than a positive feeling coming from some exemplary act, it is modeled as the negative response to a perceived affront. I adopt faithfully the interpretation of the PD outcomes (Temptation, Sucker, Reward, Punishment), assuming that when faced with the "sucker" payoff (cooperation in the face of defection) players suffer an additional utility penalty. The result is that players faced with an opponent who they expect will defect are placed in a classic bind. If they cooperate, they will feel no guilt, but will have to "swallow their pride". On the other

hand, if they "stand up for themselves" and defect in turn, they keep their pride intact, but suffer the pangs of guilt for harming another. In terms of the intensity of the emotion, there may well be some crowding out between these two competing incentives, but I assume that they are both simultaneously operative. As, I believe, much introspective evidence corroborates, I will assume that in cases of conflict, the pride will generally win out.

Characterizations of the strength of the normative incentive required to ensure cooperation under (a) known neighbor status and (b) no information about neighbor status serve as benchmarks for the usefulness of signals as coordination devices. In the full model, the only source of group information in the model is costless, unverifiable messages.[14] These messages are modeled as locations on the circle; players share a "language" consisting of a conventional zero point. The central research question is: under what conditions can such messages transmit credible information about group membership? The theoretical answer to this question is formalized as a separating equilibrium, where each social type chooses a different message to send, and players cooperate if and only if each believes the other comes from their "neighbourhood," defined as any distance less than some maximum arc length on the circle.

This model yields several empirical predictions. These predictions are tested in the second part of the paper, which outlines a discrete version of the model for empirical application, and reports on an experiment comprising four within-subjects treatments, administered to a sample of 116 subjects. The treatments all impose exogenous groups on subjects, and involve a one-shot prisoner's dilemma-type

[14] In this paper, I focus on "strategic" motivations for communication. In doing so, I abstract from the expressive role of communication. This may appear a rather surprising choice, particularly to non-economists. To the extent that messages are credible because people simply want to express their group membership, the results in this paper will only be strengthened. I will note several points at which this kind of effect might be visible. Also notice that the experimental treatment described in this paper, based on a minimal group paradigm of random, exogenous groups, represents a setting where expressive motivations will be minimized.

decision. They differ in the informational context. In Treatment 1 (T1), groups are observable; in Treatment 2 (T2), they are unobservable; in Treatment 3 (T3), subjects can exchange simultaneous messages which may or may not represent their groups; Treatment 4 (T4) is like T3, except that the signals are sequential – one player is randomly chosen to send a message first, and the second observes the first message before sending a reply. Thus we have a short (4-period, 116-subject) panel design.

This design permits several kinds of insight. First, it replicates the result of in-group bias in a minimal group paradigm, without several of the characteristic restrictions on groups, thereby generalizing those results. In particular, players who have indication that their opponent comes from the same "team" (social neighborhood) cooperate significantly more often than either those who have information that the opponent does not come from their team or those who have no information at all. Further, the latter proportions are statistically identical. The panel nature of the data allows control for personal characteristics and order effects in the evaluation of this question. Second, and more central to the research question, it allows investigation of the extent to which cheap group affiliation signals can be used strategically, which requires that they contain some information. Specifically, we see that when they are emitted simultaneously, these cheap signals are taken at face value, at least by some subjects; the in-group bias survives cheap information. On the other hand, we see that when the messages are sequential, players try to engage in some strategic signaling. Thus "lie" rates, while always high, are significantly higher in the sequential treatment than in the simultaneous treatment, and significantly higher among second-message senders, who arguably have a clearer way to lie, than among first-message senders.

1.2 Previous Literature

The current model is one where *cheap talk coveys information* about *normative incentives* relating to *group affiliation*, and in particular *cooperative in-group bias*. As such, the model straddles several strands of literature. First, it investigates situations in which players have an opportunity to communicate without any exogenous costs – that is, where talk is cheap. Notice that, in order for communication to have any real meaning, there must be some informational asymmetries *ex ante*. Foundational articles in the study of how these asymmetries may or may not be resolved are Green and Stokey (2007) and Crawford and Sobel (1982). These develop an endogenous cost of information transmission, based on the effect of the receiver's equilibrium reaction to the message on the sender. The most important feature of these models for the current purposes concerns the relationship between the theoretical possibility of sending a credible, yet cheap, message, and the extent to which players' interests are aligned. Interests are aligned to the extent that, for any given realization of the informed player's private information, the optimal response by the uninformed player is also preferred by the informed player. Any strategic interaction means that each agent has instrumental preferences over the others' beliefs: when my payoff depends on your behaviour, and your behaviour depends on your beliefs, I would prefer that you have beliefs that lead you to do make the choices beneficial to me. Interests are aligned, then, to the extent that the preferred beliefs coincide with the truth. It is thus intuitive that, if the interests are sufficiently aligned, then the sender has an incentive to reveal private information truthfully. The receiver, knowing this, can rely on the credibility of the message. The sequential equilibrium message, as succinctly expressed by Crawford (1998) means: "Given the realization of my private information variable, I like what you will do when I send this message at least as much as anything I could get you to do by sending a different message." When it is known to both parties that "this message" coincides with "my private information" for every realization, then messages are, in Farrell and Rabin's (1996) terminology "self-revealing" (if the information concerns unobservable types) or "self-committing" (if the information concerns future actions).

However, when the interests of the two players are insufficiently aligned, then this credibility evaporates. In the case of perfectly opposed interests, any message interpretation that resulted in an action that benefited the sender (which is the only kind a rational sender would emit) would at the same time result in a detriment to the receiver (an action a rational receiver would never commit). Thus when players' interests are opposed the only equilibria are "babbling" equilibria, in which the Sender's message is uninformative and is ignored by the Receiver (Crawford 1998).[15]

In the current context, the cheap talk opportunity does something of a double duty in this regard. The essence of the dilemma posed to the players in this paper is that incentives are aligned (so cheap talk should be effective) in some cases, and misaligned (so it should not) in other cases. The cheap talk itself must both allow players to identify the case, and by extension to coordinate on a subsequent action. This setup can be compared to Sally (2005). For example, his argument that talk is potentially conflict-dampening can be captured within the kind of normative incentive that I model explicitly.

The presence of normative incentives removes the current paper somewhat from the more standard theoretical models of cheap talk. In this sense, the current model has more in common with the continuing "challenge" to "neoclassical orthodoxy" which comes from behavioral studies in experimental economics. The scare quotes are used because much of this challenge is aimed not at the neoclassical model itself, but rather at an almost straw-man mental shortcut of equating "utility" with "payoffs" or, worse, with one's own money. Money can only be a measure of the relative worth of various options, and hence an index for utility. It is the

---

[15] Moreover, babbling equilibria exist even when interests are aligned. If messages are ignored, then there is no reason to emit a meaningful message. If messages are not meaningful, then they will be ignored. As several authors (e.g. Sally, 2005; Farrell and Rabin, 1996) have pointed out, it is quite likely that the theoretical possibility of such equilibria overstates their empirical relevance. In many cases (particularly those which might be characterized as self-revealing or self-committing), babbling equilibria seem to require a willfully perverse interpretation of communication as meaningless.

measurement, not the object itself. The utility associated with the outcome of any social – including any economic – interaction is a complex affair of which material benefits are only one of many determining factors (Guala, 2005). The actual challenge is to attempt to identify what those other factors are in any given situation. This challenge has been enthusiastically undertaken. Experimental tests of the "self-interest hypothesis" are numerous, and largely consistent. For example, 15 years ago David Sally (Sally 1995) conducted a meta-analysis of 130 PD game treatments in 37 published studies, finding a mean cooperation rate of just over 47 percent, with a minimum of 5 percent and a maximum of 96.9 percent. It appears very difficult to get (all) people to behave selfishly.

The natural response to this observation is to attempt to figure out why it is so hard. In this vein, theories of "other-regarding", or "social" preferences, similar in spirit to the normative incentives outlined above, constitute several different literatures by now. The unifying feature of all such work is the attempt to explicitly model non-pecuniary costs and benefits associated with different alternatives.

The normative incentives I consider in the current paper are action-based. The idea that they should be action based has several antecedents in the literature. For instance, it is related to lie-aversion, the idea that people suffer a utility cost when they tell a lie (e.g., Kartik 2009; Lundquist, Ellingsen et al. 2009). In this view, people are affected by social norms or ethical principles that forbid lying. For instance, most religions have some proscription against dishonesty.[16] More generally, it can be seen as a simple formulation of the kind of "commitment-based" incentives that were proposed by Sen (1976). The effect is also similar to that in Andreoni (1989), where in addition to the benefit of a public good, individuals gain some utility

---

[16] Examples of Christian, Jewish, Hindu and Islamic pronouncements can be found in Leviticus (19:11), Talmud (Shabbat 55), Taittiriya Upanishad (1.11.1), and Qur'an (4.135), respectively. Gneezy (2000) and Ellingsen and Johannesson (2004) review some psychological literature on lie-aversion. Gneezy (2000) also considers the views of some classical philosophers on the morality of deception.

from the act of giving (a "warm glow"), although it should be noted that the interpretation is quite different. Andreoni (1989) frames the glow as an "egoistic" motive, relating to the direct pleasure experienced from engaging in what one considers to be "ethical" actions. It might also be noted that the "guilt" which I discuss is different from that elaborated in Battigalli and Dufwenberg (2007, 2009). Their models assume that people suffer a utility cost if a co-player does not get the payoff that they think she expects. That is, a player's *second-order* beliefs (beliefs about beliefs) appear in the utility function. In the psychological literature (e.g., Gore and Harvey,1995; Tangney & Dearing , 2002; Tilghman-Osborne et al., 2010), guilt is caused by various factors including: impersonal transgressions, harming another person, and trust/oath violation, and more generally by the acknowledgement of a wrongful commission or omission of acts. This last is closest to the spirit of the use of guilt in my model.

In my model, guilt is taken to diminish with social distance. This is a venerable conjecture in the social sciences (see Charness and Gneezy 2008 for some early references), and has proven to be fertile ground for economic research in the past couple of decades. There have been several different strategies to operationalise this idea in experimental treatments. One has been to identify social distance with the degree of anonymity, under the idea that the social proximity that identification engenders between any individuals will trigger pro-social norms (Ali M 2007; Charness, Haruvy et al. 2007; Hoffman, McCabe et al. 2008; Fiedler, Haruvy et al.; Wu, Leliveld et al.). In a similar vein Catherine Eckel has done extensive work on the effect of seeing the face, smiling or otherwise, of one's co-player in experimental games (Eckel and Grossman 1996; Eckel, Kacelnik et al. 2001; Eckel 2007; Eckel and Petrie 2008).

Given a level of identification, however, social distance may not constant between any individuals. If a player's behaviours are different with those who are "close" from her behaviours with those who are "far", incentives sensitive to social

distance will result in so-called in-group bias. Economic research specifically in group identification was pioneered by Akerlof and Kranton (2000, 2005). In this work, social identities were chosen (2000), or fostered in agents by a principal's investment (2005). Thus the distribution of group identification was endogenous. By assumption, affiliation with a given group induces a specific kind of behavior in these models. Identification with a give group "automatically" causes agents to behave in the manner associated with members of that group. Wichardt (2008) extended this analysis to include the possibility of multiple, conflicting group affiliations. For instance, behavior appropriate among the members of the soccer team may become inappropriate in the context of dinner with his family. He also (2007) constructed a mechanism in which adherence to an identity whose "appropriate behavior" comprises conforming to a cooperative social norm becomes an evolutionarily advantageous trait.

In the abovementioned models, the group behavior is characteristic, but not necessarily favorable to other group members. The model in this paper is based more on an idea from social psychology, which attributes cooperative behavior to a "we-feeling" that comes from shared social identity. Simpson (2006, p. 444) describes this so-called social identity theory "social identification increases cooperation by reducing actors' tendency to draw distinctions between their own and others' welfare." At the same time, it has been emphasized that recognition of identity mismatches can generate significant discriminatory behaviour. Smith (2007) models this phenomenon theoretically in a large population with two well-defined groups, in which "behavioral" players always discriminate against the other group, while "rational" players face no inherent need to do so. He shows that in equilibrium rational players may discriminate.

However, empirical work seems to suggest that, at least in the kind of setting studied in the laboratory, in-group favoritism is often a stronger force than out-group discrimination. This preponderance seems quite robust, replicated, for instance, in

(Yamagishi and Mifune 2008); Koopmans Rebers (2009); Ben-Ner, McCall, Stephane, Wang (2009); Ahmed (2007); Lemyre Smith (1985); Mackie and Cooper (1984); Brewer (1979); and Ferguson and Kelley (1964). This in-group favoritism can manifest itself in many ways. For instance, Waldzus et al. (2005) show that people assign in-group members higher rankings in various characteristics than out-group members; Ben-Ner, McCall, Stephane, and Wang (2009) show greater conditional cooperation with in-group members; Kremer and Brewer (1984) find that in-group public goods receive greater contributions; Mackie and Cooper (1984) that hearing in-group members espouse a particular position makes it more appealing; and Ferguson and Kelley (1964) that people find products made by group members to be of superior quality. In a recent paper, (Chen and Li 2009) run a wide battery of tests allowing them to single out several motivations. They find evidence of in-group favoritism, but also find that in-group matching results in less envy, and a greater tendency to choose options that maximize social welfare. On the other hand, it seems that out-group discrimination can be primed, particularly in men (Yuki, Yokota, 2009); and some studies have found group status to be less relevant to certain results. For instance, Guth, Levatti, Ploner (2009) find that group status has little effect on trust. Unsurprisingly for a phenomenon based centrally on interpretation, it has been found that group-based prefrerences are susceptible to priming and framing effects (Yuki, Yokota,2009, Hertel and Kerr, 2001, Kramer and Brewer 1984). In a related vein, group status is a stronger predictor of behavior when subjects are members of real-life minorities (Espinoza Garza, 1985), and when they are "self-uncertain" (Hogg 2001; Hogg et al. 2007). (Hargreaves Heap and Zizzo 2009) find that groups may make people worse off, even though we like cooperating. In their experimental results they find that average trust falls when there are salient groups, although psychological benefits more or less compensate for the loss in material welfare. It is worth noting as well that many of these cited studies use the MPG framework (Tajfel, 1971), in which individuals are randomly assigned to groups on the basis of some bogus test. Pinter (2006) surveys this and other methods of inducing minimal groups, determining that

memorization of arbitrary identifiers both increases group identification and incidentally, eliminates the deception involved in some other methods. That is the methodology used in the experiment described in this paper.

To sum up the place of this model in the literature, then, the theoretical literature on cheap talk largely seeks instrumental reasons for information transmission, while this work looks at the relationship to normative incentives. Experimental work on cheap talk and lying usually has one informed party and one potential liar, and usually focuses on talk about intentions, rather than types. Here I posit bilateral cheap talk as a coordination mechanism on types. This comes from the interpretation of talk as a method of signalling group affiliation. In fact, the literature on normative incentives has so far dealt little with group affiliation. Psychologists investigate the strength and robustness of the phenomenon, but always with objectively observable (not self-identified) groups. The literature on group affiliation rarely seeks to investigate under what conditions groups can be identified, which is the main focus of the current work.

1.3 Model

1.3.1 Introduction

A continuum of agents lives on the perimeter of a unit circle[17]. The point occupied by a particular agent will be considered her *type*. Types are not ordered, but they are different. Assume that the players share a language, which consists of a conventional zero-point on the circle, so that each can make an understandable indication of any given type as a point in the interval $[0, 2\pi)$. The distance between two points x and y can be measured as the minimum arc length between them, denoted $\theta(x, y)$, or simply

[17] Or a more general symmetric structure. The objective is that every point be surrounded by an equal measure of other points.

θ where possible. The maximum distance between any two points is therefore π. Agents are distributed around the circle according to a positive, continuous and "not-too-steep" density function f. For expositional clarity, I will make the stronger assumption

Simplification 1: $f = 1/(2\pi)$; that is, that the distribution is uniform.

Agents are drawn by some "cosmopolitan" procedure and paired (i and j) for an interaction. By cosmopolitan, I mean that they have a good chance of meeting people from all over the circle. Again for exposition, I will make a stronger assumption:

Simplification 2: for any individual i of type x and arc of the circle $[a, b]$,

$$\Pr\left[a < t_i < b \middle| t_i = x\right] = \int_a^b f(z)\,dz \tag{1.1}$$

That is, draws are entirely independent. Combining this with simplification 1 means that the probability of being paired with another agent of some type between any point a and b is $\theta(a,b)/(2\pi)$. The selection process hides the agents' types, so these types are effectively private information in what follows.

The interaction consists of two stages. The last is a PD game, whose payoffs I will denote by the convention (x(C,C), x(D,C), x(C,D), x(D,D)) = (R[eward], T[emptation], S[ucker], P[unishment]), where x(A,B) is the monetary payoff from choosing action A when the opponent chooses action B. This is preceded by a

message stage, where the messages can take any value in $[0, 2\pi) \cap \varnothing$[18]. Thus, players can send a message consisting of some point on the circle, or send no message at all. Messages will be governed by – and interpreted according to – conventional (equilibrium) behaviour. Although costless, they are important because players' utility depends on their perception of the opponent's type. Specifically, the closer a player feels her opponent to be on the circle, the worse she feels for choosing the "selfish" option D in the PD game. On the other hand, people are proud, and no one wants to be taken for a "sucker." The specific form which I will use for illustrative purposes is simple:

$$V_i = x_i - \phi\left[A - gq\right] I_i^D - \rho\left(1 - I_i^D\right) I_j^D \tag{1.2}$$

where $x \in \{R, T, S, P\}$ is the monetary payoff from the PD game; ϕ is a scalar measurement of the intensity of "guilt", which diminishes linearly from some maximum value A as θ, the arc length between the opponents, measured in radians, increases; I_p^D is an indicator for choosing D in the PD game for players $p = i, j$; and ρ is a "pride" parameter, a utility-based penalty to getting the "sucker" payoff in the PD game. The linear form is not strictly necessary, and is adopted only for expositional convenience. We will see that certain conditions are required for the results to hold with this formulation; other formulations would, naturally, impose different conditions.

This general form of the model is similar to a "cold prickle" as formulated by (Andreoni 1989), the moral cost of doing harm in (Deffains and Fluet 2009), or the cost of lying in (Kartik 2009). It differs from the altruism-based conception that is at the base of interpretations of social identity theory such as (Chen and Li 2009), but yields broadly similar results, and has several advantages. First, it can be interpreted

---

[18] Technically, the messages could come from anywhere on the real line, and simply be mapped to the $[0, 2\pi]$ interval in the obvious way.

as a social norm, a behavioral rule that the player applies, and which the cost-benefit analysis must overcome. Much research seems to confirm that this kind of rule describes decision-making fairly well (López-Pérez 2008). Second, it is somewhat more flexible, as altruism parameters are bound below unity. Third, it can easily be interpreted as a "reduced form" version of other motivations such as, for instance, altruism, the guilt aversion elaborated in (Battigalli and Dufwenberg 2007; Battigalli and Dufwenberg 2009), or the "presumption of reciprocity" alluded to in (Yamagishi and Kiyonari 2000). Fourth, it does not rely on the opponent's payoff (it is *action-based* or *deontological* as opposed to *outcome-based* or *utilitarian*); since the experimental treatment does not vary payoffs, this model seems a better fit. I will assume that the PD (money-based) lab game structure exhibits the property $Q \equiv R - S - T + P = 0$[19], which implies that a player gets the same benefit from defecting, independently of the opponent's choice. I will denote this benefit Δ.

1.3.2 Benchmark cases

As a first benchmark, consider the model's prediction when types are perfectly observable. Denote the perceived probability that the opponent defects as α. Then as a function of this probability, (1.2) implies that a players have a "zone of cooperation;" they will decide to cooperate with an opponent who is less than distance

$$\theta^*(\alpha) \equiv A\phi - (\Delta + \alpha\rho) \tag{1.3}$$

---

[19] Q is a measure of the strategic "quasi-complementarity" of the PD (money-denominated) lab game. If $Q > 0$, then, while cooperation is always dominated, it is less harmful when the other also cooperates, which pushes players "towards" strategic complementarity. If $Q < 0$, then cooperation becomes more harmful when the other cooperates, which pushes players "towards" strategic substitution.

away. Consider the extreme cases of $\theta^*(0) = A\phi - \Delta$ and $\theta^*(1) = A\phi - (\Delta + \rho)$, which correspond, respectively, to the decision when playing against someone who is sure to cooperate, or who is sure to defect. Notice first that the maximum social distance required for cooperation increases in the perceived probability that this opponent will defect: the zone of cooperation contracts as the belief of defection grows. Thus the model generates some reciprocity, since for any player i, there will be types j against whom i will cooperate when α is less than some threshold, and defect when it is higher. To take a specific – and important – example, if $\rho > A\phi - \Delta$, then $\theta^*(1) < 0$, and no one will cooperate with a sure defector, regardless of how close they are. This primacy of "pride" over "guilt" seems to be an intuitive characteristic of human interaction. For example, Fernandez-Dols, Aguilar et al. (2010) find that people rate non-contribution to a public good as morally acceptable, following non-contribution from others. I will therefore maintain the assumption that it holds in the following.

To take the opposite example, the very premise of the model is that people are more likely to cooperate with those towards whom they feel some social proximity. This implies that $\theta^*(0) = A\phi - \Delta \in (0, \pi)$. The lower bound implies that people will cooperate with those to whom they feel close; the upper bound implies that they will stop cooperating as the distance crosses some threshold. Because distance is reflexive, moreover, a player type who finds her opponent is less (more) than distance $\theta^*(0)$ away can be sure that the opponent will reach the same conclusion and cooperate (defect). Thus in the perfectly observable case, these two extreme examples exhaust the possibilities for play except in marginal cases of equality, which are of measure zero. The cooperation zone can therefore be interpreted as a neighborhood, such that all agents of type t will cooperate with opponents from within their zone (which will often be referred to as "neighbors", and defect against opponents from outside their zone ("strangers"). The result is summarized below.

Observation 1: With observable types, the existence of in-group bias implies that $0 < A\phi - \Delta < \rho$, *and that* $A\phi - \Delta < \pi$. *Groups are defined by the "cooperation zone," which is calculated for any type t as* $t \pm \theta^*(0)$.

When there is uncertainty about the opponent's type, I will assume players perceive utility based on the expected value that type has; agents are risk-neutral over social distance. As an example of particular importance, consider the case in which all types give the same signal. Thus players have no more information about their opponents' type after receiving the message than before. This therefore represents the converse benchmark to the perfect information setting described above. Given the assumption of complete symmetry among the agents, this means that either everyone will cooperate, or no one will. In neither case will the "pride" component of (1.2) be operative. In the case where $f = 1/(2\pi)$, this implies any type will choose to defect if

$$\Delta > 2\int_0^{\pi} \phi(A - z)\frac{1}{2\pi}dz$$

or

$$\Delta > \phi\left[A - \frac{\pi}{2}\right] \tag{1.4}$$

which places an upper bound on the difference between $A\phi$ and Δ, but does not violate the conditions set out in Observation 1, above. It does have an implication in terms of the size of the "cooperation zone", however. Recalling that $\theta^*(0) = A\phi - \Delta \in (0, \pi)$, (1.4) implies that even under conditions of perfect information, agents who defect without information will reciprocate cooperation only with those who are closer than

$$\theta^*(0) \equiv A\phi - \Delta < \phi\frac{\pi}{2} \tag{1.5}$$

Expression (1.5) implies that a type t's "neighborhood" of t $\pm\theta^*(0)$ depends positively on the strength of the guilt felt at defection, but cannot exceed some maximum arc length of $\phi\pi$. These points are summarized below.

> *Observation 2: In contexts of no information, agents will defect when the material benefit is high enough. This implies that in such contexts, the maximum social distance compatible with cooperation with perfect information is bounded above at* $\phi\pi/2$.

In the following, I will assume a context satisfying both expression (1.4) and the conditions from Observation 1.

1.3.3 Equilibrium

The central idea that I wish to explore in this paper is that communication fosters cooperation by letting interacting individuals credibly signal areas of common social interest, which engenders a kind of conditional solidarity. In the terms of game theory, this can be interpreted as a separating equilibrium in an appropriately specified signaling game. The conditions for such an equilibrium are outlined in this section. It is well known that these games rarely have unique equilibria. Games where the signal is cheap talk have the additional irritation of babbling equilibria, in which, as described above, everyone ignores the signals, and (therefore) nobody tries to send any worthwhile information with them. The current model is no exception to these rules. However, I will focus on one in particular, which corresponds to the hypothesized phenomenon that people can use cheap-talk messages to credibly transmit unobservable group information, and thus to coordinate cooperative behavior, conditional on matching types. Specifically, I will consider the following equilibrium candidate

1. For each type $x \in [0, 2\pi)$, the chosen message is $m = x$.
2. Beliefs μ, interpreted as subjective conditional probabilities, are such that for any x and y in $[0, 2\pi)$,

$$\mu\left[t = x \middle| m = x\right] = 1$$

$$\mu\left[x < t < y \middle| m = \varnothing\right] = \int_x^y f(z)\,dz$$

3. Cooperate iff $\theta(m_i, m_j) < \theta^*(0)$

These conditions describe action both on and off the equilibrium path. Condition (1) says messages are honest, and the first part of condition (2) says they are taken as such. The second part of condition (2) says that silence (the only non-equilibrium message) will be met with uniform beliefs around the circle, an interpretation that is supported by the symmetry of the agents *a priori* when $f = 1/(2\pi)$. Hence, by the assumption that (1.4) holds, silence will meet with universal defection in the PD game, and is a dominated strategy. The third condition implies that players will defect with those they believe to be neighbors in the (off-path) cases where their own deceptive signal means those neighbors do not recognize them, and thus will defect. This reflects the idea that pride is stronger than guilt. However, it may bear more explanation.

Consider the case of a deviation in which a certain type t sends a message m different from t, and then meets a neighbour k (of t) who is not neighbours with the announced type m – that is to say, $\theta(t, k) < \theta^*(0) < \theta(m, k)$. Naturally, will k not cooperate, since she (incorrectly) believes t not to be a neighbour. This means that t is placed in the bind described at the outset. If she cooperates, she avoids the guilt associated with cheating on a neighbour. This incentive would, if the neighbour in question were also cooperating, be enough to get her to do so. However, if she cooperates, she will get the "sucker" payoff, which stings her pride. The assumption

that $\theta^*(1) < 0$ says that the second consideration wins. Thus the only sequentially rational strategy is for t to defect, even though she (correctly) believes k to be a neighbour.

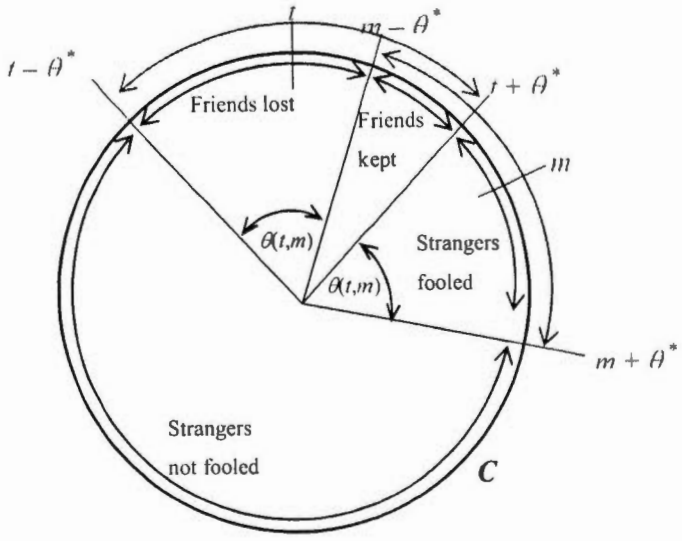


Figure 1.1 Type t deviates from the equilibrium to send message m.

The remaining deviations from this equilibrium concern the advantages of sending "deceptive" messages. To determine whether agents will deviate from this equilibrium, let us consider the various effects deviation will have. The illustration below shows the possible results of a deviation in which an individual of type t sends a message of type m, with $\theta(t, m) \neq 0$. Those in neither t nor m's neighbourhood are met with defection, which they reciprocate. If $\theta^*(0) < \pi/2$, then each type's neighborhood contains strictly less than half the circle, then there are guaranteed to be such types. Those in the intersection of the neighbourhoods (which, like the previous category, may or may not exist) are still treated as friends (although deceptively so). In any event, these two segments of the circle are the same in the equilibrium as in any deviation, so they will wash out of the decision.

The determining factor for equilibrium play will be the difference in benefit and cost of the remaining two regions. To take the cost first, those in the t's neighbourhood who are not in m's are "lost". The result is mutual defection, resulting in the P payoff from the PD game, instead of R, as well as significant guilt from harming a "close friend". However, there is no damage to pride. Following message $m > t$, these people are met with probability

$$\Pr\left(t_i - \theta^*(0) < t_j < m_i - \theta^*(0) \middle| \theta(t_i, m_i) = x > 0\right) = \int_{t-\theta^*}^{m-\theta^*} f(z)\,dz \tag{1.6}$$

Therefore, the expected net cost of a deviation is

$$\int_{t-\theta^*}^{m-\theta^*} \left[P - \phi(z) - R\right] f(z)\,dz$$

Under the assumption of a uniform density of agents and the formulation from (1.2), this gives

$$\int_0^{\theta(t,m)} \left[(P-R) - \phi(A-z)\right] f(z)\,dz$$

$$\frac{1}{2\pi}\left[\theta(t,m)(P-R) - \phi\int_0^{\theta(t,m)} (A-z)\,dz\right]$$

$$\frac{1}{2\pi}\left[\theta(t,m)(P-R) - \phi\left(A\theta(t,m) - \frac{1}{2}\theta(t,m)^2\right)\right]$$

$$\frac{\theta(t,m)}{2\pi}\left[(P-R) - \phi\left(A - \frac{\phi(t,m)}{2}\right)\right] \tag{1.7}$$

Inspection shows that (1.7) is negative and decreasing for all non-negative values of θ. Thus deviation has a cost, which increases with the magnitude of the deviation envisaged.

Turning to the benefit of deviation from the equilibrium, those in m's neighbourhood but not in t's are "tricked". The guilt from defecting on these is not great enough to restrain t, and since these people (erroneously) believe t to be a neighbour, they will cooperate. The result is that t receives monetary payoff T, the highest possible. Further, the (modest) guilt cost of doing so is paid in the equilibrium action, as well as in any deviation, and so will not temper the net benefit of deviation. And again, the deviator feels no attack on his pride, although the violated party will. Focusing on the case with $m > t$, these people will be met with probability $\int_{t_i+\theta}^{m_i+\theta} f(z)dz$. This is clearly an arc of the same size as that in (1.6). Thus, the probabilities will be the same when the distribution is uniform. The expected net benefit of a deviation is

$$(T-P)\frac{\theta(t,m)}{2\pi} \tag{1.8}$$

which increases linearly with the magnitude of the deviation. The overall effect of a deviation of size θ is therefore

$$\left((T-R-\phi A)\theta-\frac{\phi}{2}\theta^2\right)\frac{1}{2\pi}. \tag{1.9}$$

One can take the derivative of (1.9) with respect to deviation angle θ, to find the optimum deviation. The FOC is $\theta = (T-R)/\phi - A$, subject to $0 < \theta < \pi$. The SOC is satisfied, since the form of (1.9) is a downward-opening parabola. This in itself is interesting, since it means that under any parametric restrictions (maintaining the simplifications on the functional forms) there is an optimum deviation. In this case,

that optimum is zero. Notice that $T - R < T - P = \Delta < \phi A$ by the parametric restrictions above, so that (1.9) is always negative, and no deviation can ever give greater utility than the equilibrium. Thus conditions 1-3 above constitute a sequential equilibrium of the model. This is summarized below.

> *Proposition 1: In contexts where (a) agents defect against unknown opponents, and (b) agents cooperate with identifiable opponents conditional on believing they come from within the social neighborhood defined by $\theta^*(0)$, there exists a Perfect Bayesian Equilibrium satisfying conditions (1)-(3), above.*

It is worth noting that there are other equilibiria. For instance, as in most cheap talk contexts, if it is expected that messages will hold no information, then there will exist a host of weak "babbling" Nash equilibria in which players may pool on one message or none, or randomize their messages to say whatever they wish, since the result will be the same in any case: universal defection in the PD (under condition (1.4)). This equilibrium is Pareto dominated by the revealing equilibrium described in conditions (1)-(3). But moreover, one can follow Rabin and Farrell to suggest that, given we are assuming a common language, babbling equilibria in this case correspond to willfully ignoring the standard meaning of words. The equilibrium in conditions (1)-(3) is in a way bilaterally self-signaling: if any pair could agree to trust each other's messages, then neither would have an incentive to lie.

1.4 A discrete version of the model.

We have shown above that the equilibrium exists, and in many contexts the ability to identify oneself may well make it focal, so it seems natural to think that people may actually play it. The second purpose of this paper was to test the theory in an experimental treatment. However, establishing a continuous social landscape experimentally is a difficult proposition. Indeed, with a finite number of subjects it is

logically impossible. Therefore, in this section I adapt the model to a similar, yet slightly different, discrete setting[20]. Rather than have a continuous distribution of types around the circle, I let the distribution be "lumpy", although I maintain the overall symmetry. The literature on the minimal group paradigm discussed above suggests that assigning people the same group marker causes them to automatically form some in-group preference. In the model above, the "group" that types identify with on a behavioral level is their neighborhood, or cooperation zone, and is established endogenously as a function of the parameters of social distance preferences. In the minimal group paradigm, by contrast, these groups are exogenously imposed. Since the main phenomenon I wish to test is the ability to credibly transmit group information as cheap talk *given* the behavioral preferences, the transition from endogenous to exogenous groups is without much theoretical cost. On the other hand, it has considerable practical benefit, as it is at the very least a more established method of creating experimental group identities to say "Subjects X, Y and Z are in group t" than to say "Subjects X, Y and Z are in positions $t-1$, t and $t+1$."

It is also plausible that in many social situations, the groups to which people belong are more discrete than the continuous distribution above suggests. Of course, there are many possible criteria on which a person may have greater or lesser social proximity. A given person may be a man, father, athlete, artist, academic, vegetarian and southerner, and identify differently on all those dimensions. As the number of dimensions of discrete differentiation possible increases, in the limit the total difference will become continuous. However, for practical cases, there will still be a discrete number of different combinations of identification possible.

---

[20] This model is a simplified version of Spiegelman (2009). Most of the results remain qualitatively unchanged when assumptions such as equal-sized groups are relaxed.

1.4.1 Model adaptation

Let the density f of the population on the circle be adjusted to have zero measure except at some finite number P of points, each of which represents an *individual* or *person*. Further, let these people be organized into T arbitrarily small, non-overlapping regions called *groups*, located at intervals around the perimeter. The set of groups is G, and a typical person i will belong to group $g_i \in G = \{g_1, \ldots, g_i, \ldots g_T\}$.

Each group is related to some others. In terms of the previous version of the model, this relationship refers to the "close" ones, such that for individuals i and j, $\theta(i, j) < \theta^*(\mathbf{0})$. Since the groups are located on very small intervals, $\theta(i, j) \cong \theta(g_i, g_j)$, for all i and j, and I will assume that for any individuals i, j, k and n such that $g_i = g_k$ and $g_j = g_n$ $\theta(i, j) < \theta^*(\mathbf{0}) \Leftrightarrow \theta(k,n) < \theta^*(\mathbf{0})$. Thus the "close" relationship is defined on groups. Stated differently, this relationship induces a graph on G[21]. Groups are vertices; edges represent the "close" relationship I will refer to as being a *neighbor*, *friend* or *team member*. The collection of groups to whom person i is close (a group including g_i) will be known as i's social *neighborhood*, and will be noted N_i. Vertices that are unconnected are *strangers*. As above, individuals from this population meet for an interaction consisting of (1) a message and (2) an action in a subsequent PD game. To parallel the simplifications above, I will make the following assumptions:

- The measure of each person is $1/P$
- Each group contains an equal number of individuals, denoted p. Thus the measure of each group is p/P.
- N and $T \backslash N$ are both non-empty for all groups.

---

[21] Formally, the relation can be described by a set of indicator functions generating a vector $\mathbf{N}_g = [N_g(1)\ N_g(2)\ \ldots\ N_g(T)]$. The interpretation is that when $N_g(k) = 1$, group k is part of group g's network, or *neighborhood*. However, the formality adds little to the discussion at this point.

- The graph is symmetric, or the groups are equally spaced on the circle, so that all groups have neighborhoods of the same size. The number of groups in one's neighborhood *including one's own* will, with some abuse of notation, also be called N.
- The selection probability distribution is uniform over the whole population.

These assumptions imply that the prior probability of meeting a friend is

$$F \equiv \frac{1}{T}\left(N - \frac{1}{P}\right)$$

As the population P rises, this converges to the fraction of the population comprising the player's neighborhood.

1.4.2 The interaction structure

In the message stage of the interaction, the two selected individuals simultaneously emit statements m from a set of possible statements (a *vocabulary*) with cardinality greater than T. These statements may in equilibrium serve to identify the sender's group membership[22]. Since by assumption all individuals are identical within a group, they will all have the same strategic considerations, and there can never be any strict equilibrium in which members of a group send different signals. A vocabulary of size T is necessary to allow the possibility that each group uniquely identifies itself in equilibrium, which can be considered a very conservative assumption in contexts where messages come from natural languages or other signs invented by the individuals. In both of these cases, the cardinality of the message space is hard even to define. The extra messages allow for the possibility of "silence," taking the place of the empty message in the continuous-space model.

---

[22] Notice that here I skirt what Farrell and Rabin (1996) refer to as the "inessential" multiplicity of equilibria in cheap talk games. With n groups choosing between $n + 1$ messages, the number of possible separating equilibria is $n!$. It matters little which particular group selects which message, from a formal point of view. However, in most real-world scenarios the choice would be tightly restricted by the literal meaning of the messages in question.

Following the message stage, the selected people play a PD game. I will not, in this section, assume that $Q \equiv R - T - S + P = 0$, but will maintain the standard assumption that $T > R > P > S$. The utility payoff is equal to the material payoff, plus a normative utility component similar to that described above. Because I am defining the neighborhood as containing all those groups inside the threshold of $\theta^*(0)$, it follows that the normative incentive is not strong enough to induce cooperation between strangers. Thus I can equivalently assume that there is no (significant) normative incentive of "guilt", except between neighbors. To economize on parameters, I will also drop the "pride" aspect of the model.

As in the previous section, the basic assumption is that, if a player chooses D in the PD game when faced with a neighbor, she pays some utility penalty K, and pays another penalty whenever she receives the "sucker" payment, S. The cheating-cost utility function is

$$U\left(A_i, A_j\right) = x_i\left(A_i, A_j\right) - \phi I^D I^N, \tag{1.10}$$

where I^D is an indicator variable taking a value of 1 if player i chooses D, and 0 otherwise, I^N is an indicator for common group affiliation.

To summarize, the structure of the game is as follows:

- Nature chooses two individuals from G according to an i.i.d. probability distribution, which induces prior beliefs for each individual on the other's unobservable type.
- The chosen individuals simultaneously produce observable messages, and update their beliefs about the opponent's type based on the equilibrium distribution of messages across types.

- They then play a PD game, and receive monetary payoffs. If the payoffs provide more information about the opponent's type, then beliefs are updated again.
- Players experience utility based on their posterior beliefs about the opponent's type, as

$$V_i = \sum_{j \in N_i} \left(\mu_j \left| m_j, A_j \right.\right) U\left(A_i, A_j \left| g_j \in N_i \right.\right) + \sum_{k \notin N_i} \left(\mu_k \left| m_k, A_k \right.\right) U\left(A_k, A_k \left| g_k \notin N_i \right.\right) \tag{1.11}$$

where μ is the posterior belief given observed message m and action A chosen from C and D, and the difference between the two utility functions is determined by the normative incentive. Notice that "along the path" of a separating equilibrium, given a message m_i that is sent in equilibrium by those in group g_i, $\mu_j = 0$ for all $g_j \neq g_i$

Before moving on to the equilibrium with messages, it is worth establishing several benchmark results. In the full model below, the information about neighbor status may be uncertain, as it relies on cheap talk signals. This can be compared with (a) behavior when the information about neighbor status is certain, which is a kind of "best case scenario", and (b) an analogous "worst case scenario" when there is no information about the neighbor status – the babbling condition. In equilibrium, babbling should correspond to a "worst case scenario", since when players have the option of silence (or of complete randomization between messages) no one will send a message that yields an expected utility less than the "no information" scenario.

Taking the worst-case scenario first, and recalling that F was defined as the prior probability (assumed equal for all groups) of meeting a neighbor, the expected utility from cooperating, when there is a subjective probability α that the opponent defects, is

$$EU(C|\alpha) = (1-\alpha)R + \alpha S \tag{1.12}$$

The expected utility from defecting is

$$EU(D|\alpha) = (1-\alpha)T + \alpha P - F \cdot K \tag{1.13}$$

Subtracting (1.13) from (1.12), we find that cooperation is rational as long as

$$K > K^{*}_{mix} \equiv \frac{1}{F}\left[(1-\alpha)(T-R) + \alpha(P-S)\right] \tag{1.14}$$

This expression shows that the "threshold of indifference" value the normative incentive must surpass in order to generate cooperation depends, in general, on the expected play of the opponent. Admitting that all agents are symmetric, suppose, for instance, that everyone would cooperate against a completely unknown (for instance, "babbling," or simply unobservable) other. This implies that $\alpha = 0$, so it requires that $K > (T - R)/F$ – the normative incentive not to cheat, weighted by the probability that it applies, must be greater than the material incentive to do so.

On the other hand, suppose babbling leads to defection. Then $\alpha = 1$, so (1.14) implies that $K < (P - S)/F$. Notice that these conditions need not "match up". If $(T - R) > (P - S)$ (i.e., $Q > 0$, or strategic quasi-complementarity), then for $F{\cdot}K$ in the interval between them, actions become strategic complements, so that there are multiple pure-strategy equilibria without communication. Players will cooperate in cases where they think it is expected, and will defect when they think that is expected.

If the inequality is reversed, i.e., $Q < 0$, then the condition that $(T - R) < F{\cdot}K < (P - S)$ implies that actions are strategic substitutes at the utility level, so that there exists only a mixed-strategy (though potentially degenerate at $\alpha = 1$) equilibrium without messages, with probability of cooperation

$$\alpha = \frac{S - P + F \cdot K}{-Q} > 0 \tag{1.15}$$

In summary, if $Q > 0$, then there will always be a range of normative parameters such that in the utility-denominated game actions are strategic complements; if $Q < 0$, the utility-denominated game will have a region of strategic substitutability. If $Q = 0$, then the material incentives do not depend on the opponent's play, and players will defect if and only if $K < (T - R)/F$.

Incidentally, consider what happens as one becomes more certain that one is faced with a friend, that is, as F rises towards 1. This reduces the threshold for rational cooperation in (1.14), and also the mixed-strategy equilibrium probability in (1.15). The first effect reflects the fact that the normative incentive weighs heavier when it has a higher probability of binding; this translates into the second by tilting the threshold of indifference between actions towards cooperation. Another consequence of an increase in F is a widening of the range of strategic complementarity or substitution, when $Q \neq 0$. When faced with a *sure* friend, one can simply eliminate F in the formulas above. For the results that follow, I will parallel (1.4) by assuming $(P - S) < F \cdot K < (T - R)$.

1.4.3 Equilibrium results

Formally, consider the equilibrium candidate:

- Each group sends a distinct message in the communication stage
- Posterior beliefs, conditional on an observed message, are degenerate on the group that sends that message in equilibrium
- For any message not assigned to any group (off-equilibrium play, or "silence"), posterior beliefs are evenly distributed across all groups.

- Players cooperate in the PD game if and only if each receives a message sent by a neighbor in equilibrium.

Several comments about this potential equilibrium are in order. First, the distinct messages sent may not be unique. Depending on the total set of available messages, there may be arbitrarily many ways for each group to identify itself. In other words, the function mapping messages to types need not be bijective. However it is without loss of generality to redefine messages so that it is.[23] Second, the condition that "silence" is met with a uniform belief across types can be justified by the symmetry of groups. If we relax the assumption that all groups are symmetric, this part of the equilibrium loses some of its plausibility. Finally, while the assumption that the normative incentive binds only with group members is enough to ensure that players will defect along the equilibrium path if they meet someone claiming to be a stranger, the last equilibrium condition also says that players will defect if met with "silence", which requires a slightly stronger parametric constraint.

Notice that since any group message will lead someone to cooperate, which in turn raises the monetary payoff a player receives regardless of her action, "silence" is

---

[23] I also skirt again the "inessential" multiplicity of equilibria, which comes from the fact that there is no "natural language" imposed. The set of messages is quite arbitrary; any message could a priori be chosen to designate any group. This means that any separating equilibrium we describe is actually a family of n! permutations on the signals sent. This may be representative of the arbitrary nature of many group affiliation signals – the colored handkerchiefs used by L.A. gang members in the 1980s and 1990s, for example. However, there are many other cases where certain signal choices are made salient by institutional detail outside the structure of the interaction. For instance, being a Mac or Linux enthusiast is not arbitrary, but rather influenced by the professional and recreational uses the computer is put to. In these cases, the signal – perhaps a penguin sticker on a backpack – is "chosen" as a signal of a certain social affiliation in a relatively non-arbitrary way (although the penguin itself arguably remains arbitrary). Without this kind of guidance it is difficult to see how a number of groups of any significant size could realistically coordinate on full separation. In the formal development I will not concern myself with these problems. A separating equilibrium will assume some institutional mechanism for determining exactly which group signals which message.

never an optimal strategy. Thus any deviation that led to a higher utility than the equilibrium would have to involve an explicitly deceptive strategy, for example the announcement of a group membership other than one's own. The equilibrium therefore holds as long as no type has can expect a higher utility by sending a message used by some other type.

The intuition for the equilibrium is the following. Deviation will be believed in the equilibrium, so each individual has a chance to "trick" her opponent by announcing a different group from her own. This kind of deviation has costs and benefits. To clarify each, consider the figure below. The true type is noted i and the declared type is noted m. The vertices and edges of the graph induced by N have been suppressed, but the neighborhoods of the true and announced types are indicated by ovals.

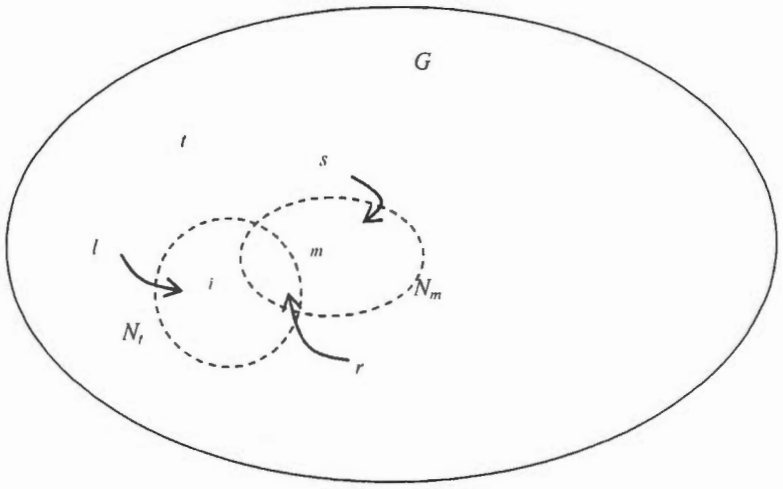


Figure 1.2 Message m different from true group affiliation i.

By declaring himself to be of group m, individual i faces the following results. Other opponents in region s will believe (erroneously) that i is a neighbor and thus they will cooperate with i. Player i, by contrast, believes (correctly) that they are strangers, and

can therefore defect, earning the highest possible payoff and feeling no disutility for doing so. This is the benefit of deviation, and clearly depends on the specific deviation undertaken. On the other hand, individuals from region l will (erroneously) believe that i is a stranger, and will therefore defect. According to the equilibrium, i will return this favor, also defecting, even though he (correctly) believes individuals l to be neighbors. Thus he loses the cooperative monetary payoff, and also feels the psychological burden of defecting against friends. The equilibrium will stand if the benefit of deviation is outweighed by the cost.

Formally, the equilibrium utility

$$V_E = N\pi_{CC} + (1-N)\pi_{DD} = (l+r)\pi_{CC} + (t+s)\pi_{DD} \tag{1.16}$$

must be greater than the deviation utility for all messages m different from that prescribed for one's own group by the equilibrium.

$$V_D = \max_{m \neq l}\left[r(m)\pi_{CC} + l(\pi_{DD} - K) + s\pi_{DC} + t\pi_{DD}\right] \tag{1.17}$$

Subtracting (1.17) from (1.16) to find the net benefit of following the equilibrium, we find

$$V_E - V_D = l(m^*)(\pi_{CC} - \pi_{DD} + K) + s(m^*)(\pi_{DD} - \pi_{DC}) \tag{1.18}$$

for the m^* that maximizes (1.17). It is enough easy to show that, when all groups have the symmetric neighborhoods, the probability of meeting an individual l is the same as the probability of meeting an individual s. Their value will depend on the deviation message m. Intuitively, we can say that the "farther" the message is from the true

group, the larger these values will be[24]. Defining a function d(m) as that distance, and another function ϕ(d) as the value that l – and by extension s – take for a distance d, we can re-write the net benefit of non-deviation as a function of the distance between the true group and the announced group,

$$B(m) = \phi(d)[\pi_{CC} - \pi_{DC} + K] \tag{1.19}$$

The function ϕ is non-negative, since it is a probability. It is also increasing in d. The term in square brackets we recognize as the condition for cooperating with a sure friend. If it is positive, then the material benefit of defection in the PD game is outweighed by the normative cost of cheating a neighbor. Therefore, expression (1.19) tells us that, as long as it is worth cooperating with a friend, it is also worth announcing your true group affiliation in simultaneous announcements. Not only that, but the farther away from the truth is the deviation considered to be, the worse it will seem by comparison. This is summarized below.

> *Proposition 2: Costless signaling can generate positive cooperation rates by enabling coordination with otherwise-unobservable group members.*

An Example:

Consider a set of five equally sized groups A to E, matched to play a PD game with the payoff matrix below.

Table 1.1 Example payoff bi-matrix

	C	D
C	8 8	11 0
D	0 11	5 5

[24] This distance can be more formally defined in graph-theoretic terms as the minimum number of neighbors required to pass from vertex i to vertex m of the graph that the neighborhoods induce on G.

Suppose their neighborhood structure is ring-shaped, so in addition to its own members, each group is neighbors with two others. In Figure 1.3, arrows represent the neighbor relationship, and the dashed shape is *B*'s neighborhood.

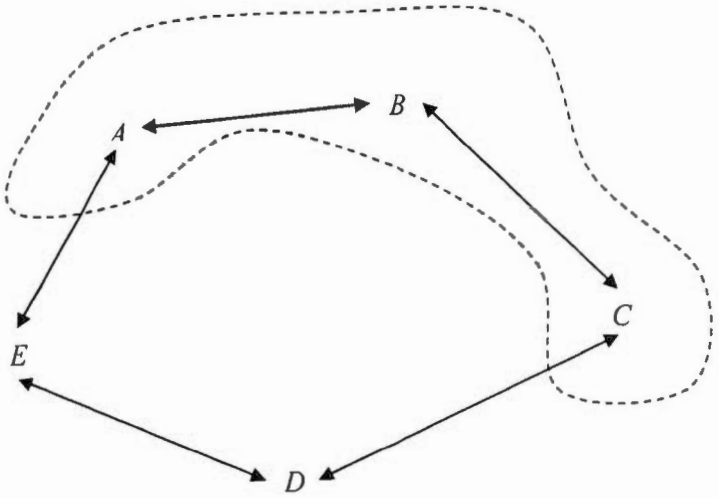


Figure 1.3 Diagram of the discrete case

Under these assumptions, $Q = 2$, and N, the probability of meeting a neighbor with independent draws, is 0.6^{25}. For simplicity, suppose the messages are A to F, in a language where each group is "supposed" to identify itself with its own letter and the letter F represents silence. Consider first what happens if everyone remains silent. In this case, if there is cooperation, then credible messages can only reduce players' utility, and thus clearly cannot be part of any equilibrium, since for the equilibrium to stand, we require non-cooperation without signals. Applying expression (1.14), we find that this implies

---

[25] Again, this holds asymptotically.

$$K < \frac{1}{N}(\pi_{DD} - \pi_{CD}) = \frac{5}{3}(5-0)$$

$$K < \frac{25}{3} \tag{1.20}$$

This means, with dollar-valued payoffs, that the equilibrium will require cooperation with group members to be worth less than \$25/3 ≅ \$8.34, which seems a conservative assumption. Turning to the equilibrium utility, the value of the term of (1.19) in square brackets is [11 – 8 + K] = 3 + K, so the equilibrium will hold for those who have 3 < K < 25/3. Note that $l = s \cong 0.2$ in this example. More specifically, there are three kinds of message possible: players may declare their true group; they may declare to be a neighbor; or they may declare to be a stranger. The symmetry of the group arrangement implies that it does not matter which neighbor or stranger you choose. The expected utility in each case is the following

$$\begin{aligned} U_{true} &= N\pi_{CC} + (1-N)\pi_{DD} = \frac{34}{5} \\ U_{neighb} &= \frac{2}{5}\pi_{CC} + \frac{1}{5}\pi_{DC} + \frac{1}{5}\pi_{DD} + \frac{1}{5}(\pi_{DD} - K) = \frac{37-K}{5} \\ U_{strang} &= \frac{1}{5}\pi_{CC} + \frac{2}{5}\pi_{DC} + \frac{2}{5}(\pi_{DD} - K) = \frac{40-2K}{5} \end{aligned} \tag{1.21}$$

As noted above, as long as K > 3, these utilities are decreasing in the distance between the true and declared group membership. There will be an "extra" message in the equilibrium that will not be reached with positive probability, and to which Bayes' Rule can therefore not be applied. A convenient and intuitive response to this will be to assign equal probability to each group off the equilibrium path, following a message of F. Convenient because, given the result above, it confirms that anyone who chooses F will be met with defection in the PD stage, which minimizes the risk

that anyone will choose that deviation. Intuitive because the basic symmetry of all groups leaves no reason to assume one is more likely than another to keep silent.

1.4.4 Extension to sequential messages

One limitation of the applicability of the model above to social interactions is the simultaneous nature of the communication. This can be interpreted as messages which are chosen before the interaction (a style of dress, for example), and cannot therefore be changed. However, in many cases one message is sent before the other. For instance, the salesman at the door with whom this paper opens observes the signals of the resident's parental status before displaying his own. Intuitive consideration of this example suggests that he may imitate the resident in order to benefit from a social proximity that is, in fact, not all that important to him. Indeed, this "parroting" behavior seems a plausible enough social phenomenon that it is worth briefly expanding the scope of the model – and empirical verification to follow – in order to incorporate it.

Suppose, then, that the players send their messages one after the other, and that the second sender (he) can observe the message sent by the first (she), before choosing what to say in reply. In this case, if the second message sender believes (somewhat naively) that the first will cooperate with those who claim to be neighbors, and defect on others, then his incentives are clear: always send a message corresponding to one of the first sender's neighbors. This message will be truthful whenever the pair are, in fact, neighbors, and will be deceptive in all other cases. This belief is naïve, however, because the first sender, understanding that *everyone* will claim to be one of her neighbors, will treat the message received as utterly uninformative. Since it was required above that players without any information will always defect in the subsequent PD game, such seems to be the inevitable result for the first message sender. Furthermore, invoking the assumption that pride is stronger than guilt, even a second-sender who is a true neighbor will defect, under the certainty that defection awaits from the first. Thus cooperation does not seem to

survive sequential messages. In terms of the messages sent, the fact that they remain uninterpreted means that there is no reason to send one more than another. However, in a context where messages have literal (or at least conventional) meanings, the "naive" strategy of one's own type may be focal would "formative" be better? for the first sender, and imitation of it by the second sender.

1.5 Experimental Procedures

1.5.1 Empirical background:

For nearly 40 years social psychologists and other social scientists have run experiments testing the strength of in-group bias. Many of these follow the *minimal group paradigm* (MPG), first elaborated by Tajifel et al. (1971). In the MPG, in-group bias is demonstrated in completely arbitrary groups, which are themselves artifacts of the experimental procedure. The procedure for inducing the group membership varies with the study. Pinter (2006) provides a survey and a test of different protocols. The method employed here reflects his results. The subsequent literature is vast, largely trying to investigate the determinants and scope of this phenomenon. In the process the phenomenon itself has been found highly robust. Part of the interest in the current project was an additional replication of the effect, under conditions of "diffuse" groups and imperfect information on whether one is in fact playing against an in-group or an out-group member.

1.5.2 Hypotheses and treatments

The theoretical model above presents several empirical predictions concerning how information about group membership will affect behavior. For some range of parameter values, in particular, one should see that players

- Do not cooperate if they do not know the group membership of their co-player, or if they know that their co-player is not a member of their neighborhood

- Cooperate if they know (exogenously) that their co-player is a member of their neighborhood
- When given the opportunity, send "truthful" messages, which correctly identify their group
- Cooperate if their co-player sends them a signal showing neighbor status.

In addition, we saw above that in a *sequential* information setting, in which one player (a leader) sends a message first, and the other (the follower) replies with a message only after having seen that sent by the leader, then messages become meaningless. Given the prediction above about non-cooperation without information, we can therefore now add another prediction: Player should also

- Not cooperate when information is passed sequentially,

And followers may well

- Imitate the leaders with their messages when messages are sequential.

These predictions are tested in the experiment using the following four treatments.

T1: observable groups. Experimenters inform subjects of the group membership of their co-player. This treatment establishes the in-group bias.

T2: no group information. Subjects play the game "blind" as to whether their co-player comes from the same neighborhood or not. This treatment establishes the benchmark of uncooperative action in the PD game.

T3: simultaneous cheap talk. Experimenters allow subjects to send a "message" consisting of one group identification to each other before playing the game. Both players decide which message to send before seeing which was sent by the other. A "truthful" message will be one where the group identification sent in the message is the same as the group to which the player has been allocated. This treatment shows

whether such signaling is credible (i.e., people send truthful messages), and whether it is believed (the in-group bias is maintained when the messages are "cheap talk").

T4: sequential cheap talk. This treatment is essentially a robustness check. One subject (a follower) sees the other's (the initiator's) message before sending her own. Both messages are seen before the game is played. In this case, the model predicts that the initiator's message should be truthful, and the followers' should be truthful only if the groups actually match. The follower's message, in fact, should always be from the same team as the initiator's.

1.5.2.1 Hypotheses:

These predictions translate into several null hypotheses about the data that emerge from the treatments. The variables of interest are: the cooperation rates ρ_T^X and ρ_T^S which will be defined as the proportion of subjects who choose (A) in treatment T, conditional on being matched with (N)eighbors and with (S)trangers, respectively; and ν_T^i the aggregate rate at which players send truthful messages, by role type (initiator and follower). In T3, all types will be considered to be initiators, since there are no followers. The specific hypotheses are as follows:

H1: $\rho_T^N > \rho_T^S$: The overall cooperation rate is greater when matched with neighbors (in-group bias)

H2: $\rho_1^X = \rho_3^X$ for $X = S, N$: The cooperation rate in T1 is the same as in T3 (credible messages), conditional on meeting an in-group member (or someone who claims to be one)

H3: $\rho_2 = \rho_4^N = \rho_4^S$: Initiators should ignore the messages of followers, and play as if no information were transmitted. Followers, knowing this, will also be uncooperative.

H4: $\nu_3^{init} > \nu_4^{foll.}$: players send truthful messages in T3, and followers do not send truthful messages in T4

1.5.3 Specific methodology

Subjects were recruited using the ORSEE system (Greiner, 2004) by the CIRANO Experimental Economics laboratory. The total sample size was 116 subjects, in 6 sessions of 20 subjects each between November 9 and November 20, 2010[26]. Descriptive statistics concerning basic demographics are shown in the table below.

Table 1.2 Demographics of the sample

	mean	N	min	max
Gender	.5	116	0 (F)	1 (M)
Age	26.8087	115	18	58

It has often been found that subjects of different academic backgrounds respond to experimental manipulations differently, particularly in cases of so-called "mixed motive" games, where individual welfare considerations are at odds with social efficiency or fairness concerns. Since the set up in the current study is of this kind, the relative frequency of the various fields of study are presented below. Because this information was collected anonymously by the host laboratory, it cannot be matched to individual subjects. Therefore I cannot control in the results for any effects these majors may have on behavior. However, it can be seen in Table 1.3 that just over 50% of the sample came from science, mathematics, engineering business administration and economics, which have been found to behave more "selfishly" – that is, more in line with the dictates of individual rationality – than other disciplines (Lopez-Perez and Spiegelman, forthcoming).

---

[26] There were actually 7 sessions run at the laboratory. However, due to a programming error, the data in one session had to be discarded. Appendix Y gives more information about the problems with this data.

Table 1.3 Distribution of fields of study

Field of study	Freq.	Percent	Cum.
Non-university/undeclared	16	13.79	13.79
Business and economics	35	30.17	43.97
Science math and engineering	24	20.69	64.66
Other social science	11	9.48	74.14
Languages, literature, humanities	9	7.76	81.90
Health and medicine	9	7.76	89.66
Applied language skills	4	3.45	93.10
Law	4	3.45	96.55
Music	4	3.45	100.00

The structure of the experiment was within-subjects. Each subject participated in each of 4 treatments, differing in the match of the other player and the informational content. The order of the treatments varied somewhat, but was not random. This probably resulted in some order effect in the results, which are discussed below. However, the within-subjects design means that some individual effects can be controlled for – we see some counterpositives that would not otherwise have been possible. The CIRANO laboratory is computerized; each subject was seated at a visually isolated computer terminal. Experimental instructions were provided via a PowerPoint presentation with pre-recorded voice track for consistent presentation[27], and the treatments were programmed and conducted with z-Tree (Fischbacher, 2007).

1.5.3.1 Group assignment

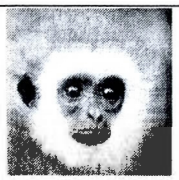
Computers were assigned to groups exogenously to establish the desired distribution. Each session, there were four computers assigned to each of 5 groups[28]. On entering

[27] For this and other tips on the practicalities of running an experiment, I am indebted to Jim Engle-Warnick.

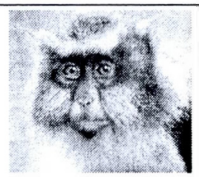
[28] The final session had only 16 subjects; in this case, there were 3 subjects assigned to each of 4 groups, and 4 subjects assigned to the fifth. The last seat was randomly assigned, which should induce symmetric priors on the subjects about the distribution. This is the strategically important consideration.

the laboratory, subjects drew cards informing them of which computer to sit at; this was the source of the random group assignment for each subject. For concreteness, the groups were identified by pictures of the monkeys in Figure 1.4, below. The names were selected so that each monkey would have a color and another, species-related uncommon word in its name[29].


Golden Colombin


Black-masked Gibbon


Grey Langur


Yellow-tufted Mona


Silver Tamarin

Figure 1.4 Group identifiers.

Neighborhoods were imposed on the groups, so that each was in the same neighborhood as 2 other groups; that is, $N_i = 0.6$ for all groups i. Subjects were told that each monkey-group was in "a *team*, or *family*, or *neighborhood*" with two other groups. These terms were used with emphasis, and interchangeably throughout the instructions. During the experimental treatments, the term *team* was retained. Subjects were repeatedly reminded of their own group membership, as well as of the other groups who were on their team.

1.5.3.2 Payment

Subjects were introduced to the particular PD bi-matrix from Table 1.4, which is the game played experimentally. The payoffs are denominated in dollars.

---

[29] The method of image selection was somewhat involved. See Appendix 3 for more details.

Table 1.4 Payoffs from the interaction.

Row \ Column	A	B
A	9 9	11 0
B	0 11	5 5

1.5.3.3 Sequence of events.

On entering the experimental laboratory, subjects signed an informed consent form, and then watched a 12-minute presentation, introducing them to (a) the nature of the groups and (b) the interaction structure. Subjects privately answered questions to make sure they had understood both aspects of the experimental design. Since the number of plays was small (each subject played each treatment only once, for a total of 4 games), the instructional presentation went into some detail concerning the nature of the PD game. The terminology during the experimental sessions and the instruction period was kept as morally neutral as possible: choices were consistently identified simply as A or B. However, it was pointed out that if they chose the same options, players received the same payoff, and that if they chose different options, the player who chose B got $11 – the highest possible payoff – which the player who chose A got $0 – the lowest possible payoff. To make sure that play was not clouded by incomprehension, it was even pointed out that playing B would always earn more money than playing A, but that if *both* play A, each gets a higher payoff than if both play B.

Following the instructional presentation, subjects played four rounds of the game, under informational conditions T1 to T4. As mentioned above, the order of the treatments was varied somewhat across sessions, but was not random. Table 1.5, below, shows the breakdown of treatment orders.

Table 1.5 Order of the treatments

Order	Freq (# sessions in the data)
T1-T2-T3-T4	4
T2-T1-T3-T4	1
T3-T1-T2-T4	1

It will be observed that T4 was always last. Because of the sequential nature of the interaction, this treatment took the longest to complete, and was arguably most complex to perform. In order to keep attention focused and make sure the tasks were clear, subjects therefore started the session with relatively quick, easy treatments. Preliminary analysis of the data indicated order effects (similar to the decay of overcontribution in public goods games, e.g. Burlando and Guala 2002, (Nirel and Gorfine 2003)), which was the reasoning behind the variation in the final two sessions[30].

After the games, one round was selected at random for payment. Finally, a short questionnaire was distributed, among other things, to ask about subjects' motivation for their choices.

1.6 Data and results

In this section I describe the empirical results of the analysis. First I will detail the effects of the treatments on cooperation rates (H1-H3). After that I will investigate the patterns of messages across and within treatments (H4).

1.6.1 Cooperation

Table 1.6 shows the mean cooperation rates, pooled across all treatments, according to whether or not the information that passed suggested that the players came from the same neighborhood.

---

[30] The panel nature of the data allows me to statistically control for timing effects to some extent.

Table 1.6 Pooled results for cooperation as a function of neighbor status.

Choose cooperative strategy number (percent)							
		No		Yes		Total	
Same neighborhood	Yes	134	(60.4)	88	(39.6)	222	(100)
	No	90	(73.2)	33	(26.8)	123	(100)
	No info	83	(72.2)	32	(27.8)	115	(100)
	Total	307	(66.8)	153	(33.3)	460	(100)

Note: Table reports frequency (percentage), rounded to significant digits.

Table 1.6 shows that overall, 39.6 percent of those who saw information suggesting that they come from the same neighborhood cooperated, while 26.8 percent of those who saw some information suggesting that they did *not* come from the same neighborhood as the opponent cooperated. While perhaps not huge in absolute terms, this difference is statistically significant (Pearson chi-square d.f. = 1, $p = 0.017$; Mann-Whitney test $z = -2.38$; $p = 0.017$). This confirms (fails to reject) the hypothesis H1: the data replicates the phenomenon of in-group bias, even though (a) groups were diffuse and (b) in many cases, the information about group status was "cheap." It is also striking how similar the overall results were between those who "knew" they were not on the same team (second row of Table 3.3), and those who simply didn't know (T2, third row). Table 3.3 shows that 73.17% of the former chose the uncooperative strategy, while 72.17 percent of the latter did. Unsurprisingly, a Mann-Whitney test fails to reject the equality hypothesis ($p = 0.863$). Finally, note that the overall average cooperation rate of 33.26 percent is not extraordinarily high or low, considering that there were no opportunities to make explicit promises, and the weighting of the sample towards "math and science" types. (Recall that Sally (1995) found rates ranging from 5% to 96%).

However, pooled results are not the primary interest. The primary interest of the model is to see whether cheap-talk signals of group affiliation are credible. Table 1.7 gives an impression.

Table 1.7 In-group cooperation rates in each treatment.

Treatment	N (in-group)	In-group cooperation rate	Mann-Whitney *p*, H_0: = T1	Mann-Whitney *p*, H_0: = T2
T1: Observable	84	0.464	.	0.006
T2: Blind	116	0.278	0.006	.
T3: Simultaneous	74	0.419	0.568	0.046
T4: Sequential	64	0.281	0.024	0.966
Total	338	0.396	.	.

Note: In all treatments except T2, *N* refers to the number of subjects who had some information that their match was from the same neighborhood.

The data in this table is restricted to subjects whose information was such that they were playing a member of their own neighborhood, when such information was available. For the round where it was not (T2, blind), the entire sample is used. Recall that the observable and blind treatments serve as benchmarks for the maximum and minimum possible information, respectively. It was hypothesized that cooperation in T3 would be the same as that in T1 (H2) and that cooperation in T2 would be the same as in T4 (H3). The Mann-Whitney tests are far from rejecting this hypothesis. The additional hypothesis suggested by a *Homo economicus* model, that cheap talk should be meaningless and therefore that cooperation rates should be the same in all treatments, is rejected. In effect we see two "kinds" of treatment. In treatments T1 and T3, the in-group cooperation rate was relatively high. In treatments T2 and T4, the cooperation rate (in-group for T4; overall for T2) was relatively low. Treatments of the same "kind" are statistically indistinguishable from each other; however, treatments of different kinds are significantly different. This can be taken as evidence

that (a) messages have an effect if they are credible; (b) not all messages are credible; (c) some cheap messages are credible.

So far, the hypotheses presented have all been confirmed by the data. Looking a little deeper into the cooperation rates reveals some complexities, however. Table 1.8 expands on Table 1.7 to show both in-group and out-group cooperation rates in each treatment. In T2 (blind), again, overall rates are presented, since there was no information about the group match. The final column presents the difference in the cooperation rates, and a p-value for a Mann-Whitney test of equality.

Table 1.8 Cooperation rates by in-group/out-group status in each treatment.

Treatment	*Type*	*N*	*mean*	*Difference (p)*
a. Observable	Ingroup	84	0.464	
	Outgroup	32	0.406	0.056
	Total	*116*	*0.448*	(0.5795)
b. Blind	Ingroup	-	.	.
	Outgroup	-	.	.
	Total	*116*	*0.276*	
c. Simultaneous	Ingroup	74	0.419	
	Outgroup	42	0.238	0.181
	Total	*116*	*0.357*	(0.0512)*
d. Sequential	Ingroup	65	0.292	
	Outgroup	51	0.196	0.096
	Total	*116*	*0.25*	(0.2369)
e. Total	Ingroup	223	0.399	
	Outgroup	125	0.264	0.136
	Total	*348*	*0.332*	(0.0114)**

Note: Differences in cooperation rates are accompanied by Mann-Whitney p-values. In the blind treatment there was no information on in- versus out-group matching; the "total" information in panel e. thus refers to the other three treatments only. * significant at 10% level; ** significant at 5% level.

Inspecting Table 1.8, we see that in each treatment, the direction of the difference in cooperation rates is as predicted by H1 (i.e., more cooperation when faced with a group member). The conundrum of Table 1.8 resides in Panel a., the observable treatment (T1). According to the predictions from the model the

information should be most reliable in this case, and so a greater effect of in-group status expected. On the contrary, this treatment had the closest results between in- and out-group matches. Likely not coincidentally, it also had the highest cooperation rates overall, particularly among out-group members. Indeed, the anomalous entry in the table appears to be out-group cooperation rates with observable groups. This result appears to be in conflict with in-group bias. In the case where the group knowledge was sure, there was no difference between in-group and out-group matches!

There are at least three departures of the current experimental design from the canonical minimal group paradigm that might, alone or in combination, contribute to this result. First, it could be that the MGP simply does not generalize to overlapping groups. Second, it could be that endogenous group identification weakens the credibility too much for the phenomenon to hold. These are key features of the underlying model, and if they were the source of the anomaly, that would make the paradigm unsuitable to testing the theory. However, it seems safe to minimize the risk from these factors. Mainly, this is due to the fact that the anomaly occurred in one treatment, while both of the factors above were present throughout the study. Indeed, the anomalous treatment was the only one with *exogenous* group identification! A third, and more incidental, departure from the canonical design involves the level of "anthromorphism" in the group identifiers. Often groups are identified by preference for paintings, or by random draw of a color of card. Here, the identifiers were monkeys with faces that do not escape the adjective "cute". It may also be important that to recall from Table 2 that in 4 out of 6 sessions, the observable treatment (T1) came first. One might well conjecture that, particularly in early rounds, before the (relatively complex) group structure had been assimilated, simply observing that one's "opponent" was a cute monkey would be enough to reduce social distance between any individuals. This effect, if present, might drown out the in-group bias, at least until sufficient repetition had reinforced the group structure. The result would be a sort of time-trend, where the in-group bias would grow stronger over the course of

play. A panel regression using the round (1 to 4, regardless of the treatment) as the time variable and an individual random effects variable by subject should filter out this time-trend. A random-effects estimation is preferable to a fixed-effects estimation for at least two reasons. First, if there is little correlation between the explanatory variables, the random-effects measure is more efficient, and second, fixed-effects models are inconsistent in a maximum-likelihood estimation framework, which means they cannot be applied to the standard estimation techniques for discrete choice data (Greene, 2003). By design, an experimental framework seeks to minimize the correlation among explanatory variables. Finally, a Hausman test comparing the estimates from linear (OLS) fixed- and random-effects regressions utterly fails to reject the hypothesis of equality ($p = 0.9671$), which in the linear case is taken to be evidence that random-effects is legitimate. Thus in Table 1.9, below, I report the results of three random-effects probit estimations, regressing the probability of cooperation on an indicator for being matched with someone from the same team, as well as dummies for treatments T1, T3 and T4. Since there was no team information in T2, this treatment was dropped from the regression. In regression 1, T4 was dropped for the comparison point; in regression 2 and 3, T1 was dropped.

Table 1.9 Probit regression results

	Regression 1: T4 omitted			Regression 2: T1 omitted			Regression 3: Interactions		
	Coef.	(SE)	*p*	Coef.	(SE)	*p*	Coef.	(SE)	*p*
Matched with team member	0.455	0.199	0.022	0.455	0.199	0.022	0.067	0.347	0.847
T1 (observable)	0.749	0.214	0.000		-omitted-		-omitted-		
T3 (simultaneous)	0.436	0.209	0.037	-0.313	0.197	0.112	-1.023	0.414	0.014
T4 (sequential)		-omitted-		-0.749	0.214	0.000	-0.942	0.388	0.015
T3xteam match							1.024	0.511	0.044
T4xteam match							0.268	0.486	0.582

Note: random-effects probit regressions of cooperation on dummies for team-match and two of the three periods where such information was available. For each regression, $N = 345$ in a balanced panel of 115 individuals over 3 periods.

The first two regressions yield several important points. First, matching with a team-member is, overall, a significant positive force in promoting cooperation (p = 0.022). Further, while compared to T4 (panel (a)), both T1 and T3 had significant, positive effects (p = 0.000 and 0.003, respectively), compared to T1, T3 is insignificant (p = 0.112), but T4 has a significant negative effect (p = 0.025). This can be interpreted to mean that there is no significant difference between the cooperativeness in observable and simultaneous treatments, but that both are significantly more effective than the sequential treatment in engendering cooperation. Regression 3 adds more detail to this picture. There the insignificant base effect of team matching (p = 0.847) illustrates again that in the comparison treatment (T1), there was no effect. The significant negative direct effects of T3 (p = 0.014) and T4 (p = 0.015) show that those who did not match cooperated much less in those treatments. However, the significant positive interaction for team matching in T3 (p = 0.044) shows that in that treatment, those who did match cooperated. Indeed, magnitude of the coefficients of the direct and interacted effects in T3 sum to essentially zero, showing that those who met neighbors in T3 cooperated essentially at the same level as the comparison treatment (T1). Meanwhile, the interaction on T4 is not significant (p = 0.582), showing that meeting a neighbor had no effect there. Thus, except for the lingering anomaly of indiscriminate cooperation in T1, these data support the model predictions. Moreover, the regression lends weight to the conjecture that the cuteness of the monkeys swamped the group bias in the first round.

1.6.2 Message choice

A further aspect predicted by the model concerns the truthfulness of messages. It was predicted that in T3 (simultaneous messages), the monkey sent would correspond to the actual group membership of the message sender. T4, by contrast, makes no such prediction. A plausible conjecture, based on a naïve interpretation of the interaction that may become focal in the "babbling" sequential design predicts that first movers

may announce their true group, perhaps due to some kind of (unmodeled) lie aversion, and subsequently ignore second-mover messages. The naïve prediction of the effects is that all second-movers should announce a group in the same team as the first-mover's announcement; those who are actually in the same group will tell the truth, and those in a different group will lie.

In the data, lies are common. There seems to be a "baseline" lie rate of about 50%. This may be linked to another phenomenon, which is an apparent strong idiosyncratic tendency for individuals to lie or tell the truth in all treatments. For instance, Table 8 shows the bivariate breakdown of individuals' lie decisions in T3 and T4. We see that more than half of those who told the truth in T3 also did so in T4, while nearly 78% of those who lied in T3 also lied in T4. A Chi-square test (d.f. = 1) yields a p-value of 0.0000, indicating that indeed individuals T3 and T4 lie decisions were not independently made.

Table 1.10 Lie rates across the message rounds

	T4Lie		
T3Lie	no	yes	Total
no	31. 54.4	26 45.6	57 100
yes	13. 22.0	46 78.0	59 100
Total	44. 37.9	72 62.0	116 100

Note: Frequency and row percents shown. Rows: message in T3 corresponded to true monkey; Columns: message in T4 corresponded to true monkey

However, there is also an unequivocal increase in the lie rate from T3 to T4. Table 7 shows that the overall rates are 50.86% (T3) versus 62.07% (T4), which represents a nearly 24% increase over the baseline level, and is a significant difference (paired t-test of equality $t = -2.11$, $p = 0.037$). And despite the idiosyncratic

tendencies mentioned above, it can be seen from Table 8 that twice as many (26 subjects) lied in T4 but not T3 as lied in T3 but not T4 (13 subjects). This difference permits a sign test of the hypothesis that the average change in behavior was zero, yielding a marginally significant p-value of 0.053, lending modest initial support to H4. While the significance of this result is not stellar, recall that in T4 there is actually no prediction that players should lie; the strict theory predicts pure babble, which might as well take the form of truth as anything. Consider the naïve interpretation conjectured above, however. This predicts that second-movers should lie more than first movers, since they have a clear (though still not exactly equilibrium) incentive to imitate the first-mover. Repeating the two-way table exercise from Table 8 by first- or second-mover status reveals the highly suggestive fact that, while first-movers tended to reproduce their T3 messages (Fisher's exact test p-value against independence = 0.001), second-movers T4 messages were statistically unrelated to their T3 messages (p = 0.260). The second-movers were apparently motivated by some other factor beyond the "inertia" that T3 messages show. What was causing the second-movers to change the way they had played from T3? Had they simply been babbling randomly, one would expect that they would end up in the first player's team about 60% of the time, since neighborhoods occupy 60 of the social space. In fact, nearly 76% of pairs (88 out of 116) in T4 matched teams, a significantly higher value (two-sided binomial sign test p = 0.000). I take this as evidence that people were in fact playing the naïve strategy conjectured, which means that they were using their messages in something other than a "babbling" manner.

Additional evidence consistent with this conjecture concerns the prediction that not all second-players had an incentive to lie, even by the naïve strategy. Specifically, second-movers who found that the first mover announced a team member could imitate that player without lying. Table 1.11 therefore shows the average lie rate for first and second message senders, depending on whether the message from the other came from the cooperation zone. Notice that the first sender had not yet seen the

message when this choice was made, so there is no surprise that these numbers are similar, and statistically indistinguishable from the baseline lie rate. Similarly, second senders who see a first sender's message from their own team lie at the baseline rate. However, those who receive a stranger's message lie well over the baseline rate, at 81.8%. Despite the small samples that remain from this splitting of the group, a Mann-Whitney test of the difference in second-mover messages shows a significant difference ($p = 0.0071$).

Table 1.11 Lie rates by move order and first-movers team match

	First movers lie rate (N)	Second movers lie rate (N)
Other claimed to be off team	0.67 (18)	0.81 (33)
Other claimed to be on team	0.52 (40)	0.48 (25)
Total	0.57 (58)	0.67 (58)

1.7 Conclusion

This paper has presented a theoretical model of social distance that generates in-group bias based on an action-oriented normative incentive. This work bridges the gap between two existing literatures. On the one hand, it relates to studies on the effects of social distance, which generally highlight the cooperative impulse any two people might share when they become "personalized". On the other, it uses the social distance concept to generate effects of in-group bias, which study how people are discriminately, or selectively cooperative towards those with whom they feel they share some common characteristics. This model is presented in two forms. The first is in continuous space, and finds a "cooperation zone" such that any individual cooperates in a one-shot prisoners' dilemma game if he is convinced the other player comes from a social location within the zone. This in turn supports a Bayesian equilibrium in which individuals costlessly, yet credibly, disclose their true social location in a simultaneous-messages game. The second form, easier to apply to

common situations, starts with the cooperation zones as overlapping groups, or teams, and derives similar results. I then briefly demonstrate how the cooperative cheap talk equilibrium does not survive the relaxation of simultaneity in messages. When one player sends a message first, and the other replies, then it is always in the second sender's interest to deviate from the "honest" equilibrium and claim to be from within the first sender's team. Knowing this, the first sender will not put any credibility in the message. The model therefore makes no predictions about the messages sent, however I conjecture a particular deviation as a plausible, though somewhat "naïve" behavioral pattern.

Following this, I describe an experiment carried out to test certain predictions the model makes. The experiment generalizes the minimal group paradigm (MGP) to a case with endogenous group identifiers and overlapping groups. The results are, broadly consistent with the theory. In particular, I find that evidence that a match with a social neighbor raises cooperation rates when messages are simultaneous, but not when they are sequential. In addition, the overall cooperation rates with sequential messages are the same as those with exogenous identification. Cooperation rates with sequential messages are significantly lower, but are not significantly different from rates when there is no information at all. Also in line with the theory, I find that the rate of deceptive messages is significantly higher in the sequential treatment than in the simultaneous, an effect largely due to second-movers imitating the team offered by the first-mover. All this seems to suggest – as the theory predicts – that simultaneous messages are credible, while sequential messages are a kind of naïve "babble".

The greatest departure from the theoretical predictions is that when group identifiers are exogenously credible, the case arguably most comparable to standard MGP studies, the in-group bias does not manifest. Given the robustness of in-group bias as an empirical phenomenon, it seems sensible to look for causes of this effect in the peculiarities of the current design. Three main departures from the canonic MGP

design were discussed. Two of these (endogenous credibility of the group signal and overlapping groups) are key to the theory that undergirds the experiment. One hopes, therefore, that the fault does not lie there. Fortunately, these apply to all rounds, including those where the in-group bias was not found. Indeed, the endogeneity of the group identification does not apply to the treatment where the effect was the weakest! I therefore conclude that these are not at fault. An additional, and more incidental, difference concerned the degree of personification of the group identifier. Standard MGP procedures have subjects randomly allotted to groups based on relatively impersonal markers (Pinter and Greenwald 2010). In the current study, the identifiers shown in Figure 1.4 were chosen with an aim to compensate for the more abstract groups with more concrete images. However, it may have occurred that in so doing, the images initially reduced the perceived social distance between all groups. ("We're all just monkeys, after all!") It may be conjectured that some time and experience with the imposed group structure was required to overcome this initial induced cooperativeness. In addition, despite the alterations in the order of treatments, they remain significantly correlated; T1 was in the first round in 80 of the 116 observations. Spearman's rho is 0.586 ($p = 0.000$). Thus the bulk of the time of diminished group identification would have occurred in T1, which would generate just the results obtained.

This scenario is, of course, surmise. However, as Sherlock Holmes would say, it fits the facts available. When more facts become available, we shall have ample time to revise the theory. Follow-up experiments to the one described above could be conceived. Three particular changes I would envisage would be as follows. First, replace monkeys with simple solid colors. These colors could be arranged in a wheel such that primaries (red, yellow, blue) are in the same team as the mixed colors they comprise. For example, red would be in the same color as orange and purple. Thus mixed colors would be in the same team as the primaries of which they are composed; orange would be in the same team as red and yellow. Each neighborhood

would therefore be equal to half the total social space. The second change would be to include an explicit "empty signal". Since these are dominated strategies in the theory, they were excluded from the experiment. However, it may be worth including them. Finally, I would change the payoff structure to make $Q = 0$, for better approximation of the continuous version of the model.

Bibliography

(2010). "What is the human library?"

Ali M, A. (2007). "Group identity, social distance and intergroup bias." *Journal of Economic Psychology* 28(3): 324-337.

Andreoni, J. (1989). "Giving with impure altruism: applications to charity and Ricardian equivalence." *Journal of Political Economy* 97(6): 1447-1458.

Battigalli, P. and M. Dufwenberg (2007). "Guilt in games." *American Economic Review* 97(2): 170-176.

Battigalli, P. and M. Dufwenberg (2009). "Dynamic psychological games." *Journal of Economic Theory* 144(1): 1-35.

Charness, G. and U. Gneezy (2008). "What's in a name? Anonymity and social distance in dictator and ultimatum games." *Journal of Economic Behavior & Organization* 68(1): 29-35.

Charness, G., E. Haruvy, et al. (2007). "Social distance and reciprocity: An Internet experiment." *Journal of Economic Behavior & Organization* 63(1): 88-103.

Chen, Y. and S. X. Li (2009). "Group identity and social preferences." *American Economic Review* 99(1): 431-457.

Crawford, V. (1998). "A Survey of Experiments on Communication via Cheap Talk." *Journal of Economic Theory* 78(2): 286-298.

Crawford, V. and J. Sobel (1982). "Strategic information transmission." *Econometrica* 50(6): 1431-1451.

Deffains, B. and C. Fluet (2009). Legal liability when individuals have moral concerns. *CIRPEE Working Paper 09-51*: 42 pgs.

Eckel, C. (2007). "People playing games: the human face of experimental economics." *Southern Economic Journal* 73(4): 840-857.

Eckel, C., A. Kacelnik, et al. (2001). "The value of a smile: Game theory with a human face." *Journal of Economic Psychology* 22(5): 617-640.

Eckel, C. C. and P. J. Grossman (1996). "Altruism in Anonymous Dictator Games." *Games and Economic Behavior* 16(2): 181-191.

Eckel, C. C. and R. Petrie (2008). Face Value. *Experimental Economics Working Paper Series*, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University.

Farrell, J. and M. Rabin (1996). "Cheap Talk." *Journal of Economic Perspectives* 10(3): 103-118.

Fernandez-Dols, J.-M., P. Aguilar, et al. (2010). "Hypocrites or maligned cooperative participants? Experimenter induced normative conflict in zero-sum situations." *Journal of Experimental Social Psychology* 46(2010): 525-530.

Fiedler, M., E. Haruvy, et al. (2011). "Social distance in a virtual world experiment." *Games and Economic Behavior* 72(2): 400-426.

Green, J. R. and N. L. Stokey (2007). "A two-person game of information transmission." *Journal of Economic Theory* 135(1): 90-104.

Greene, J. D., S. Morelli, et al. (2008). "Cognitive load selectively interferes with utilitarian moral judgment." *Cognition* 107(3): 1144-1154.

Guth, W., M. Ploner, et al. (2008). Determinants of in-group bias: group affiliation or guilt-aversion? *Jena Economic Discussion Papers 1008-046*. Jena, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

Hargreaves Heap, S. P. and D. J. Zizzo (2009). "The Value of Groups." *American Economic Review* 99(1): 295-323.

Hoffman, E., K. McCabe, et al. (2008). Chapter 49 Social Distance and Reciprocity in Dictator Games. *Handbook of Experimental Economics Results*, Elsevier. Volume 1: 429-435.

Kartik, N. (2009). "Strategic Communication with Lying Costs." *Review of Economic Studies* 76(4): 1359-1395.

Ledyard, J. (1995). Public Goods: A survey of experimental research. *Handbook of Experimental Economics*. J. H. Kagel and A. E. Roth. Princeton, Princeton University Press: 111-194.

López-Pérez, R. (2008). "Followers and leaders: Reciprocity, social norms and group behavior." *Journal of Socio-Economics*.

Lundquist, T., T. Ellingsen, et al. (2009). "The aversion to lying." *Journal of Economic Behavior & Organization* 70(1-2): 81-92.

Nirel, R. and M. Gorfine (2003). "Nonparametric analysis of longitudinal binary data: An application to the intergroup prisoner's dilemma game." *Experimental Economics* 6: 327-341.

Pinter, B. and A. G. Greenwald (2010). "A comparison of minimal group induction procedures." *Group Processes & Intergroup Relations*.

Sally, D. (1995). "Conversation and cooperation in social dilemmas." *Rationality and Society* 7(1): 58-92.

Sally, D. (2005). "Can I say "bobobo" and mean "There's no such thing as cheap talk?"" *Journal of Economic Behavior & Organization* 57(2005): 245-266.

Spiegelman, E. (2009). The real cost of cheap talk. *Mimeo*. Montreal, UQAM, CIRPEE.

Spiegelman, E. (2011). The (Human) Nature of Economic Choice. *Actes du Premier Colloque des Cycles Supérieurs de l'ADÉPUM*. Montreal, ADÉPUM.

Wichardt, P. (2008). "Identity and why we cooperate with those we do." *Journal of Economic Psychology* 29: 127-139.

Wu, Y., M. C. Leliveld, et al. (2011). "Social distance modulates recipient's fairness consideration in the dictator game: An ERP study." *Biological Psychology* 88(2-3): 253-262.

Yamagishi, T. and T. Kiyonari (2000). "The group as the container of generalized reciprocity." *Social Psychology Quarterly* 63(2): 116-132.

Yamagishi, T. and N. Mifune (2008). "Does shared group membership promote altruism? Fear, greed and reputation." *Rationality and Society* 20(1): 5-30.

APPENDICE A: IDIOSYNCRATIC K

Suppose people are heterogeneous regard to their sensitivity to the guilt normative incentive, K. Thus there may be non-zero probabilities that an opponent would, for instance, defect with a sure friend, or cooperate against a full mix. This then induces a probability that a given, randomly selected neighbor will cooperate, function of the distribution of the normative sensitivity and the strength of the material incentives that have to be overcome. The probability corresponds to the proportion of the population whose normative sensitivity exceeds a given threshold.

Thus we see that the threshold cost of cheating required to ensure cooperation falls as the information about the neighbor status of the opponent improves. Expression (1.14) implies that a player facing a known neighbor must, to ensure rational cooperation, have a cost satisfying

$$K > K^{*}_{friend} \equiv \alpha(\pi_{DC} - \pi_{CC}) + (1-\alpha)(\pi_{DD} - \pi_{CD}) \qquad \text{(A1)}$$

I will assume that K distributes randomly in the population, with support on some non-negative interval of the real line and continuously increasing CDF $\theta(K)$.

Denoting the perceived probability that the opponent will cooperate as α, equation (A1) implies that players should cooperate with neighbors whenever

$$\alpha(\pi_{CC} - \pi_{DC}) + (1-\alpha)(\pi_{CD} - \pi_{DD}) + K > 0 .$$

Thus, those individuals will not cheat whose cost satisfies

$$K > K^* = (\pi_{DD} - \pi_{CD}) - \alpha Q \,. \qquad \text{(A2)}$$

where $Q = \pi_{CC} - \pi_{DC} - \pi_{CD} + \pi_{DD}$ is a kind of strategic quasi-complementarity[31]. Notice that this also implicitly defines α as $1 - \theta(K^*)$. For instance, if K is uniform between $\underline{K}$ and $\bar{K}$, then an interior α will be equal to

$$\alpha = \frac{(\pi_{DD} - \pi_{CD}) - \bar{K}}{Q - (\bar{K} - \underline{K})}, \qquad \text{(A3)}$$

assuming the denominator is non-zero.

---

[31] Notice that as Q rises, the threshold level of cost that ensures cooperation falls. More generally, if $Q = 0$, then $\pi_{CC} - \pi_{DC}$, the net cost of cooperation if my opponent cooperates, is the same as $\pi_{CD} - \pi_{DD}$, the net cost of cooperation if my opponent defects. Defection in this case increases my score by the same amount, regardless of whether or not my opponent also defects. The actions of other players affect only the level of the payoff profile, not the "marginal" (more precisely, incremental) cost and benefit of my own actions. If this difference is positive, by contrast, that means that my opponent's defection hurts me more if I cooperate than if I defect. So if I expect the opponent to defect, this pushes me towards defection myself. If I expect the opponent to cooperate, then the same argument pushes me towards cooperation. Regardless, it should be well noted that defection is a dominant strategy. It is for this reason that I call Q "quasi" complementarity.

APPENDICE B : MONKEY IMAGE SELECTION

The group images used were selected from internet sites by the experimenter on the basis of trying to find faces that were both roughly equal in attractiveness, and also distinctive enough to make recognizable groups. However, in a pilot run with just over 100 college students, it appeared that there were significant disparities in people's preferences for the monkeys. Students were asked to "vote for a monkey", via a Web-based interface, without further explanation. Out of 118 students, 84 replied with a vote. A full 40% (34 votes) of the sample voted for one monkey, and another garnered a meager 5% (4 votes). Unsurprisingly, a Chi-square goodness of fit test (DF = 4) returns a statistic of 34.2, with a p-value of 0.000.

It seemed that assigning subjects to groups that had such a heterogeneous appeal could only lead to confounds in the resulting behavioral patterns. To solve the problem, a larger group of monkeys was selected and redistributed to the same sample. Individuals were directed to eliminate monkeys, and return a subset they "liked." A single vote was then distributed equally among the chosen images. If an individual returned 4 images, for example, each received a weighted vote of 0.25. Images were then compared by the total "weight" of the votes they received. This procedure is somewhat *ad hoc*; however, for the five monkeys included it yielded statistically indistinguishable results in terms of numbers of votes (Chi-Sq = 2.770, DF = 4, P-Value = 0.597), and as can be seen below, the distribution of weighted second-round votes among the monkeys in both rounds is visually very similar to the number of votes they received in the first round, thus it seems to yield similar results to a direct vote.

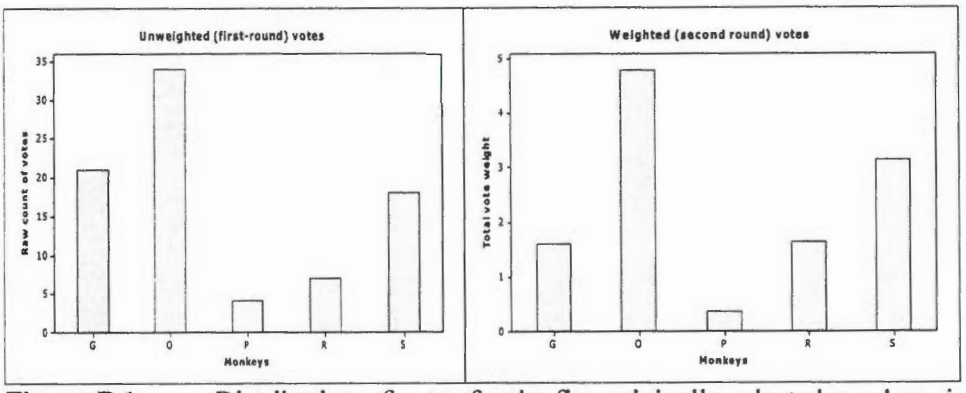


Figure B.1 Distribution of votes for the five originally selected monkeys in two voting procedures.

Again, there were significant disparities in the measured appreciation for the different monkey images. The monkey images chosen in the end represent the "middle of the road"; they were neither in the most favored nor the least-favored[32]. The following graph shows the monkeys ordered by vote weighting. The selected monkeys were numbers 4, 8, 10, 12 and 13.

[32] Incidentally, the same image that won 40% of the first-round votes also had the highest weight in the second round.

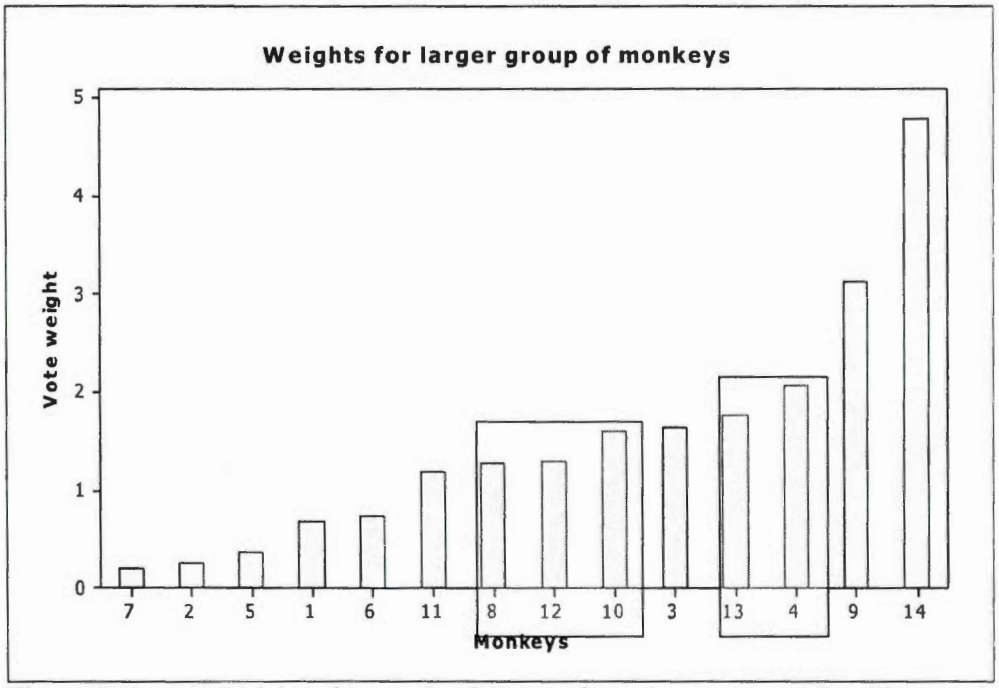


Figure B.2 Weight of votes for full set of monkeys; chosen five illustrated in boxes.

Monkey 3 was not selected because it is quite different-looking from the others, in that it has more background, you can see one hand, and it is eating something. Monkey 4 was selected over monkey 11 because 11 again had quite different lighting, and a different kind of expression. The full set of monkeys is available from the author on request. It may be noted that the chi-square goodness of fit test for equal proportions on the number of votes each of these monkeys yields a statistic of 0.286 (d.f. = 4, p-value = 0.991).

CHAPITRE II

WHY DO PEOPLE TELL THE TRUTH? EXPERIMENTAL EVIDENCE FOR PURE LIE AVERSION[1]

Abstract: Recent experimental literature shows that truth-telling is not always motivated by pecuniary motives, and several alternative motivations have been proposed. However, the relative importance of motivations behind lie aversion in any given context is still not totally clear. This paper investigates the relevance of pure lie-aversion, that is, a dislike for lies independently of their consequences. We propose a very simple design where pure lie aversion predicts positive truth-telling. Other motives considered in the literature, by contrast, predict zero truth-telling. Thus we interpret the finding that more than a third of the subjects tell the truth as evidence for pure lie-aversion. Our design also eliminates confounds with another motivation (a desire to act as others expect us to act) not frequently considered but consistent with much existing evidence. We also observe that subjects who tell the truth are more likely to believe that others will tell the truth as well.

Résumé: Une littérature expérimentale récente montre que les gens s'abstiennent souvent de mentir quand les incitations monétaires les poussent dans l'autre sens. Parmi plusieurs motivations possibles pour cet excédent d'honnêteté, nous isolons de manière expérimentale une aversion pure au mensonge, c'est-à-dire une désutilité de l'acte de mentir, indépendamment des conséquences. Nous trouvons que plus d'un tiers des 258 sujets résiste à la tentation de mentir. Qui plus est, les sujets qui disent la vérité sont plus portés à croire que les autres la diront aussi.

---

[1] Chapter co-authored with Raul Lopez-Perez. Thus the first-person plural pronoun will be used throughout.

2.1 Introduction

In many important economic settings, people may increase their expected material gain by providing information that they believe to be false – in short, by lying. Immediate examples include accounting, auditing, insurance, job interviews, labor negotiations, regulatory hearings, and tax compliance. Based on the standard *Homo economicus* assumption that all agents are self-interested money maximizers, economic theory predicts that people will always respond to these situations mendaciously. Much of the mechanism design and Principal-Agent literature is indeed aimed precisely at avoiding this result. However, a recent experimental literature shows that people often tell the truth in several cases where the theory would not predict it. The broad research question this paper addresses is: Why do they do that? One potential motivation for honesty that has attracted attention in the literature is *pure lie-aversion*, that is, the idea that people suffer a utility cost when they tell a lie (e.g., Ellingsen and Johannesson, 2004; Kartik, 2009). Intuitively, people care about social norms or ethical principles that forbid lying, such as those based on religions like Christianity and Islam, and feel bad if they utter a lie.[2] In this paper, we investigate experimentally the extent to which people are motivated by pure lie-aversion.

Experimental methods are ideal to explore the relevance of lie-aversion. The existence of truthfulness seems to falsify the *Homo economicus* assumption in some situations, but more controlled decision contexts are required to further discriminate between potential motivations for honesty. For instance, an altruistic person might tell the truth not because she values honesty per se, but because she believes that a deceived co-player could make a harmful choice. More subtly, communication might reduce social distance, and that could in turn reinforce players' altruistic feelings

[2] Ellingsen and Johannesson (2004) and Gneezy (2005) review some psychological literature on lie-aversion. Gneezy (2005) also considers the views of some classical philosophers on the morality of deception.

(Bohnet and Frey, 1999), thus fostering truth-telling. Similarly, open-form communication might create some sort of social identity (Orbell et al., 1990, Buchan et al., 2006) and thus again foster altruism.

Apart of altruism, another potentially important factor that might affect honesty is guilt-aversion. We note in this respect that guilt is multifarious in the psychological literature (Baumeister et al., 1995; Gore and Harvey, 1995; Tangney & Dearing, 2002; Tilghman-Osborne et al., 2010): It may be caused by impersonal transgressions, harming another person, trust/oath violation, or more generally by the acknowledgement of a wrongful commission or omission of acts. In our paper, (*payoff-based*) guilt-aversion refers to the idea, first introduced in Dufwenberg and Gneezy (2000), and more formally elaborated in Battigalli and Dufwenberg (2007, 2009), that people want their co-player to get the payoff that (they think) she expects, and suffer a utility cost when she does not. This phenomenon can clearly be influenced by communication whenever that affects beliefs about the co-player's expectations – the so-called "second-order expectations." For instance, models of guilt-aversion offer an explanation for costly truth-telling in settings where the lie increases the divergence between the co-player's (expected) expectations and her actual payoffs. Such a hypothesis has been object of experiments including Charness and Dufwenberg (2006), and has received some verification.

Alternatively, one can think of other specifications for guilt-aversion, such as what we call here *act-based guilt-aversion*, that is, the theory that people like to act as others expect them to. In particular, a message sender likes to tell the truth when she believes that the message receiver(s) expect it.[3] While this theory has not received

---

[3] We stress that the crucial difference between pure lie-aversion and guilt-aversion is that the latter posits a relation between beliefs and the activation of the bad feelings, whereas pure lie-aversion assumes that the bad feelings are activated simply by uttering a lie. The name given to such feelings, in contrast, is largely immaterial for the distinction. Pure lie-aversion does not rule out that the bad feelings are what psychologists call guilt.

much attention in the literature (see Peeters et al., 2007 for an exception), it is potentially important because it is consistent with much experimental evidence and with behavioral patterns studied in psychology. For instance, Rosenthal (2003, p.151) asserts the existence of "hundreds of studies [demonstrating] that one person's expectations for the behavior of another person can actually affect that other person's behavior", even when those expectations are not made explicit.

Our experimental design allows us to discriminate between pure lie-aversion and the previously mentioned motivations, including both versions of guilt-aversion, an issue that previous studies have not addressed. We consider two treatments of a very simple, one-shot game where one player must send a (truthful or false) message to another one. Act-based guilt-aversion predicts a positive but different rate of truth-telling across treatments, whereas pure lie-aversion predicts the same positive rate in both treatments. Altruism and payoff-based guilt-aversion predict zero truth-telling. In both treatments, furthermore, we measure first and second-order beliefs about truth-telling, which should be correlated with behavior according to act-based guilt.

Overall, 38.76 percent of the subjects choose to tell the truth in our study, but there is not a significant difference in the rate of truth-telling across treatments. Our results therefore suggest that pure lie-aversion is a significant force behind honesty, whereas act-based guilt-aversion is not, at least in this context. Yet we do find that first- and second-order beliefs co-vary significantly with behavior. For instance, subjects who tell the truth in our study are significantly more likely to believe that others will tell the truth as well. This suggests a need to enrich the assumptions surrounding the theory of lie-aversion (as suggested in Bicchieri, 2005; Erat and Gneezy, 2009; or López-Pérez, 2010).

Our study contributes to an already large experimental literature examining deception and honesty, reviewed extensively in the next section.[4] It is most closely related to Peeters et al. (2007), Fischbacher and Heusi (2008), Erat and Gneezy (2009), and Sánchez-Pagés and Vorsatz (2009), which provide evidence consistent with the existence of lie-averse agents, controlling for several potential confounds. Our paper provides additional evidence in line with lie-aversion and complements the previous literature with a design that can discriminate between lie-aversion and act-based guilt,[5] controlling moreover for the effect of altruism, social distance, social identity, and payoff-based guilt-aversion on truth-telling.

Following the literature review, we formally describe act-based guilt-aversion in section 3. Section 4 describes our experimental design and procedures. Section 5 describes the results from our experiment and section 6 concludes.

2.2 Related literature

Our main research goal is to investigate the relevance of pure lie-aversion, preventing any confounds with other potential motivators of honesty. Although the focus of previous experimental studies on deception and honesty differs slightly from ours, a detailed review of such literature can help to clarify what we already know on this issue and moreover illustrate the numerous motivations that can subtly affect honesty, in particular act-based guilt-aversion.[6]

---

[4] It is also worth noting that a substantial literature in social dilemmas and coordination games over the past 30 years has found that costless, non-binding communication is a robustly effective force affecting strategic behavior (see the surveys in Ledyard, 1995; Sally, 1995; and Crawford, 1998).

[5] Peeters et al. (2007) run an experiment with a sender-receiver game played over 100 rounds with re-matching, and analyze the performance of a model of "consequentialistic preferences" with characteristics similar to what we denominate here act-based guilt-aversion, and another of "deontological preferences" similar to pure lie-aversion. Although some results tend to lend weight to the idea of act-based guilt-aversion, a test of such model in the repeated sender-receiver game is hindered by the fact that it predicts multiple equilibria. In the conclusion, we discuss how some of our results could help to understand dynamic play in repeated games like theirs.

[6] Our main focus will be on studies that specifically investigate lying and honesty, but it is worth noting at the outset that a substantial literature in social dilemmas and coordination games over the

Gneezy (2005) considers three sender-receiver games conceptually similar to those in Crawford and Sobel (1982). One player (the *receiver*) must choose between two allocations of money (A and B) for herself and another player (the *sender*). Allocation A gives more money to the receiver than does allocation B, whereas the opposite happens for the sender. However, only the sender knows the true payoff constellation: the receiver's only guidance is a message from the sender, which is restricted to "Option A will earn you more money than option B." (which is true) or the opposite statement (which is a lie). Note that the receiver can never ascertain whether the other player lied or even had an incentive to do so. Gneezy elicits the expectations of some senders and finds that 82 percent expect the receiver to follow their message. If any such sender wants to maximize her monetary payoff, therefore, he should deceive the receiver. Yet the rate of deception varied between 17 and 52 percent in the three games.[7]

Several motivations could arguably explain why some senders tell the truth even at a cost. First, pure lie-aversion is obviously consistent with this fact. Second, altruism could also play a partial role. Since most senders in Gneezy (2005) expected messages to be trusted, lies were expected to reduce the co-player's payoff. It follows that altruistic senders could tell the truth in order not to harm the receiver.[8] Third, communication might reduce social distance among the subjects and interact with the sender's altruistic concerns. The study by Frohlich et al. (2001) suggests one possible reason for this (see also Bohnet and Frey, 1999). They argue that subjects may have

---

past 30 years has found that costless, non-binding communication is a robustly effective force affecting strategic behavior (see the surveys in Ledyard, 1995; Sally, 1995; and Crawford, 1998).

[7] The games considered by Gneezy vary the benefit of deception (to the sender, if the message is followed) as well as its harm (to the receiver); he generally finds that the rate of deception rises in the first and falls in the second. See Hurkens and Kartik (2009) for some caveats to this finding

[8] On the other hand, Gneezy compares behavior in his sender-receiver games with behavior in dictator games with the same payoff constellation. Significant changes in behavior in this context indicate that truth-telling in the sender-receiver games is not only due to the sender's preference for a certain payoff distribution.

doubts about the veracity of an experimental design if it presents a high level of anonymity and social distance, and those doubts can affect their behavior. For instance, their data suggests that altruism in dictator games is negatively affected by doubts about whether dictators are paired with real people. After communicating in sender-receiver games, conversely, senders might gain in confidence that there is another person on the "other end" of their decisions, and that could make them more altruistic. Fourth, for act-based guilt-aversion we recall again that most senders expected the receiver to believe the message sent. Senders believed that the receiver expected truth-telling, which by our assumptions should make them more likely to tell the truth. *Fifth*, selfishness could also explain part of the truth-telling, since 18 percent of the senders expected the receiver not to act as recommended and hence they could tell the truth even if they were selfish.[9] *Finally*, payoff-based guilt-aversion might also explain honesty, provided that the sender has appropriate second-order payoff expectations. In effect, if the sender believes that the receiver expects a payoff larger than the minimum possible (recall in this respect that the receiver is uninformed about the payoff constellation), the sender is more likely to tell the truth in order not to disappoint her.

Charness and Dufwenberg (2006) use a simplified trust game with a random shock and three treatments: (1) No pre-play communication; (2) the first mover (*investor*) can send an open message to the second mover (*trustee*) before the first move; (3) same as (2), but now the trustee is the message sender. Further, the authors elicit the investor's beliefs regarding the trustee's behavior, and the trustee's second-order beliefs in this respect. Their key findings are: (a) Compared with no communication, cooperation and overall efficiency rise significantly when the trustee sends a message (treatment 3) but not when the investor sends the message (treatment 2); (b) investors anticipate these patterns, and trustees anticipate this anticipation by

[9] Sutter (2009) demonstrates that people actually follow this strategy to some extent.

the investors; (c) there is a correlation between cooperation by the trustee and her second-order beliefs (i.e., those who think that the investor expects cooperation are more likely to actually cooperate); (d) messages coded as "promises" increase cooperation and efficiency more than other messages.

What theory could explain these results? Obviously, the standard model of selfish players fails to predict cooperation in any treatment. As explained in detail in López-Pérez (2010), pure lie-aversion predicts findings (a), (b), and (d), although it is ambiguous regarding point (c). As Charness and Dufwenberg (2006) note, however, lie-aversion could explain point (c) if the assumption of a (false) consensus effect is added, so that cooperative players tend to believe that other players are cooperative as well. In turn, payoff-based guilt-aversion can explain results (a) to (d) provided that in the absence of a promise to the contrary, investors expect trustees to act selfishly. Note that since the game is set up so that lies, actions and payoffs are correlated, act-based guilt has the same predictions as payoff-based guilt.

In Vanberg (2008), subjects play a binary dictator game with a choice set similar to the trustee's in Charness and Dufwenberg (2006). Before knowing their role (dictator/recipient), however, subjects are matched in pairs and can send at most two open-form messages to the co-player. Afterwards, half of the pairs are re-matched and the subjects' roles are determined. Further, dictators are informed whether they were re-matched (in that case, they can see the messages sent and received by their new co-players), whereas recipients *are not*. Then dictators choose. Subjects play 8 rounds according to this protocol (they are re-matched each round) with no feedback, except own payoff in each round. Recipient's first-order beliefs about the dictator's choice and dictator's second-order beliefs are elicited in an incentive-compatible manner.

Consider a player who announces during the communication stage that, provided that he remains matched with the other player and becomes the dictator, he

will make a generous choice. If that player is sufficiently lie-averse, such a promise should foster generous behavior if the proviso holds, but not if he is re-matched with another player, to whom no such promise was made.[10] If that player is (payoff) guilt-averse, in contrast, his generosity should be equally affected by others' promises as by his own. In effect, generosity should depend only on second-order beliefs, which are unaffected by the re-matching because recipients do not know whether they have been re-matched. In contrast to this prediction, re-matched dictators facing recipients who had received a promise from someone else had higher second-order expectations, but were not more likely to act generously than those facing recipients who had not received such a promise. In line with lie-aversion too, dictators who promised to act generously were significantly more generous if they were not switched than if they were re-matched with a recipient who had received a promise from another subject, even when second-order beliefs were not significantly different in both cases.

As in the previous studies, pure lie-aversion could play a role in explaining these results, but there are many other potential forces. For instance, the communication protocol uses open-form messages, which communicate more than intentions: players can transmit information about their personal characteristics, their economic needs, or simply make jokes. This may create some sort of social identity (Orbell et al., 1990; Buchan et al. 2006) and highlight or increase a feeling of sympathy or altruism.[11] If a dictator likes a person with whom she has communicated,

---

[10] Should a lie-averse player ever make such a promise in equilibrium? Yes, but only in a conditional manner, that is, only if the co-player makes the same promise as well –a proof of this is available from the authors. Note that an unconditional promise imposes a cost and provides no benefit, as the game is not played repeatedly with the same co-player.

[11] In this respect, Charness and Dufwenberg (2009) use the same game as Charness and Dufwenberg (2006), but with closed rather than open messages. More precisely, trustees may either send a promise to cooperate, or stay silent. Maybe unsurprisingly, most subjects send a promise. Yet investors' behavior was not significantly different from the no-communication treatment in Charness and Dufwenberg (2006), while the trustees who sent promises were only marginally more cooperative than those who did not. Hence, this comparison *across studies* seems to point against both lie-aversion and

therefore, she would be probably more likely to make a promise, thus partially explaining the results. In addition, some form of reciprocity could play a role, as dictators who are not re-matched probably face a recipient who promised to give money before. It may also be noted that this experiment cannot distinguish between pure lie aversion and act-based guilt aversion. In effect, a re-matched dictator who had promised to be generous with *someone else* does not lie if he is not generous with his final co-player, and thus can feel no guilt for doing so: act-based guilt aversion must also predict less generosity on re-matching. Finally, the results are at odds with the assumption that *all* players display payoff-based guilt-aversion, but could still be replicated if a fraction of the agents were guilt-averse and another, say, lie-averse.

In the baseline treatment of Fischbacher and Heusi (2008), subjects roll a six-sided die, and are paid according to a self-report of the number which comes up.[12] While the researchers cannot discern whether any specific individual lied, the aggregate rate of deception can be estimated based on differences between the observed and expected number of observations for each report, assuming a fair die. The authors find that not all players declare the profit-maximizing answer. Indeed, some even declare the lowest-payoff outcome. Furthermore, among those who lie, some are "incomplete liars," reporting a value higher than that which they rolled, but which does not maximize payoff. These results are broadly robust to tripling the stakes, adding externalities, repeated play, and double-blind anonymity.

This study is comparable to ours because it eliminates any effect of altruism, social distance, social identity, and payoff-based guilt-aversion on truth-telling. Yet several other factors apart of pure lie-aversion could play a role in the decision

---

guilt-aversion, while pointing to the role played by open-form communication in enhancing social cooperation.

[12] The payoff from a die roll of (1, 2, 3, 4, 5, 6) is CHF(1, 2, 3, 4, 5, 0), respectively. Thus a report of "5" earns subjects the most money. Notice that no number should come up significantly more or less than one sixth of the time if everyone were truthful.

whether to be honest, so that the relevance of this motive is somehow unclear. To start, participants in this study send a message to the experimenter when they report the number, and act-based guilt-aversion predicts that the expectations of the experimenter should be important to them. Second, subjects were asked to enter the number that they had thrown, so that a lie involved (a) making an untrue statement, but also (b) cheating, that is, surreptitious violation of a formal rule, and also (c) contravening a direct instruction from an authority figure. Any of these considerations could foster honesty in this context –e.g., Milgram (1963) famously showed people's reluctance to do (c). Finally, subjects in Fischbacher and Heusi were told that the die-roll was designed to determine how much they would be paid for filling out a survey. It might be that some subjects feel that it is somehow "unfair" for the same task to be paid differently to different people. It is not totally clear how this rule could affect the results. It could push up the rate of lying, especially if people find 4 or 5 CHF a fair payment for the questionnaire.

Erat and Gneezy (2009), used a game whose the basic structure of the game is similar to that in Gneezy (2005). However, the design includes nine different games to cover several different varieties of lie: altruistic white lies, which hurt the sender and help the receiver; Pareto white lies, which help both parties; and selfish black lies, which help the sender and hurt the receiver. Most relevant for our study, the authors find some subjects refrain from lying even when it results in an increase in both parties' payoffs (Pareto white lies). This lends strong support to pure lie-aversion. Since subjects' expectations were not measured, however, it is also possible that act-based guilt-aversion was driving the results.

Peeters et al. (2007) study a sender-receiver game played over 100 rounds with re-matching. In some rounds, the receiver has an option to sanction the sender, reducing both players' payoffs to zero rather than accepting the resultant payoff. Although the standard equilibrium with selfish players involves randomization by

both players and no sanctions, the authors find that some players tell the truth more than predicted. These players also tended to trust messages, and sanction those who lie more often. The authors study the performance of a model of "consequentialistic preferences" with characteristics similar to what we term here act-based guilt-aversion, and another model of "deontological preferences" similar to pure lie-aversion. Some of their results tend to lend weight to the idea of act-based guilt-aversion, although the model of Peerters et al. predicts multiple equilibria in the sender-receiver game (including the standard one).[13]

Finally, Sánchez-Pagés and Vorsatz (2009) consider a sender-receiver game where the sender can tell the truth, lie or remain silent (in that case, she pays a small cost). In this manner, the authors discriminate between a preference for truth-telling (i.e., getting a utility payoff if one tells the truth) and pure lie-aversion (which predicts silence under certain equilibrium conditions).[14] The game is played fifty times (with re-matching and change of roles); the authors report that senders tell the truth more often than predicted by a standard model of selfish players, and that the rate of choice of silence is significantly larger than zero. While this latter result points again to the importance of lie-aversion, the relevance of this motive is unclear, as other motivations like act-based guilt-aversion could also play a role.

In summary, while the existing literature has significantly improved our understanding of the incentives behind truth-telling, some uncertainties remain. The relevance of pure lie-aversion is still unclear because in previous studies other factors could have affected honesty as well, like the interaction between communication and

---

[13] In the conclusion, we discuss how some of our results could help to understand dynamic play in repeated games like this.

[14] Our experimental design cannot distinguish between these two closely related motives, and we leave this for further research. Yet, as we suggest in the conclusion, a small variation of our design could allow us to do so.

altruism, payoff-based, and act-based guilt-aversion. In what follows, we propose a design to investigate pure lie-aversion that eliminates potential confounds.

2.3 Act-Based Guilt-Aversion

Consider an extensive-form two-player game[15] with players A and B, and let $x_A(z)$ denote A's monetary payoff at terminal node z. At some point in the game, suppose that A has an opportunity to communicate something to B, and may tell the truth (i.e., send a message consistent with her beliefs) or lie (send a message inconsistent with her beliefs). Let I(z) denote an indicator taking value 1 at terminal node z if A lied in the history of z, and value 0 otherwise. Suppose further that at any terminal node z, player B has beliefs about the probability that A has told a lie in the history of z, and let $\mu(z)$ denote A's second-order beliefs about B's beliefs. That is, A thinks that B thinks that A has lied with probability $\mu(z)$. Finally, let $G_A(z) = \max [0, I(z) - \mu(z)]$ denote the intensity of A's guilt . This set up is partially inspired in Charness and Dufwenberg (2006). To formalize the idea that people dislike feeling guilty, player A's preferences $\succ$ are defined over the set of vectors $[x_A(z), G_A(z)]$, and satisfy rationality, plus two axioms. The first is a monotonicity axiom: other things equal, people prefer more money to less:

Axiom G1 (monotonicity): Given $G_A(z) = G_A(z')$, $[x_A(z), G_A(z)] \succ [x_A(z'), G_A(z')]$ if $x_A(z) > x_A(z')$.

The second axiom is a continuity hypothesis. Intuitively, this states that an increase in guilt can be compensated by a sum Δ of money, which is a function of two factors:

---

[15] We focus on the two-player case for simplicity; the ideas would generalize easily.

Axiom G2 (continuity): For any x_A and any $G_A < G^*_A$ there exists some positive amount of money Δ which depends on x_A and $(G^*_A - G_A)$, such that $[x_A, G_A] \approx [x_A + \Delta(\cdot), G^*_A]$.

We make two remarks. First, $\Delta(\cdot)$ can be different for each player A. In fact we will assume in Section 3, when discussing our experimental game, that there exists some measure of A-players for any possible value of Δ (we do not need to be precise regarding the distribution of players). Second, if we additionally posit that the function $\Delta(\cdot)$ strictly increases with the difference $(G^*_A - G_A)$, axioms 1 and 2 imply the principle of act-based guilt-aversion. That is, a greater deviation from what one thinks was expected evokes stronger feelings of guilt:

Principle GA (guilt-aversion): Given x_A, $(x_A, G_A) \succ (x_A, G'_A)$ if $G_A < G'_A$.

It is worthwhile to underline some differences between (a) act-based and (b) payoff-based guilt-aversion (as in Charness and Dufwenberg, 2006). To start, the expectations in theory (b) are about *B's payoffs*, while in theory (a) they are about *A's actions*. Note also that in payoff-based guilt-aversion the lie itself is not normative. A lie can increase, decrease, or leave guilt unchanged, depending on how it affects the payoff expectations; in act-based guilt-aversion, in contrast, the lie itself is normative: a player feels guilty for lying regardless of the payoff effect, albeit less guilty as she thinks that the expectations that she will lie will grow. These ideas can be thought of as incorporating the principle of so-called psychological game theory, that second order beliefs enter directly into the utility function, into other, more traditional preferences. Theory (b) is like a "psychologised" altruism, in which A cares about B's payoff to the extent that (A thinks) B expects to get a high payoff. Theory (a) is like a "psychologised" lie-aversion, in which A intrinsically cares about lying to B, to the extent that (A thinks) B expects A not to. To finish, we can also compare act-

based guilt and pure lie-aversion. With our previous notation, lie-aversion assumes that preferences depend on vectors $[x_A(z), I(z)]$, and posit a dislike for lies (that is, $I(z)=1$), other things equal. Therefore, the main difference between act-based guilt and lie-averse preferences is that the latter do not depend on the agent's beliefs, but only on whether he/she lies.

2.4 Experimental Design and Procedures

Our design uses a very simple, one-shot game with two players (A and B). Player A privately observes a random signal on the computer screen (more precisely, a green or a blue circle) and must then choose a message for player B. Two messages are always possible: "The green circle has appeared" or "The blue circle has appeared". Monetary payoffs are as follows: A gets 15 Euros if he announces the green circle and 14 if he announces the blue one, whereas B always gets 10 Euros. Note three things: (i) A's payoff depends on the message sent, not on the realization of the signal, (ii) A faces a dilemma between honesty and material interest if the signal happens to be blue, as telling the truth is then costly, and (iii) B does not observe the realization of the random signal and hence cannot verify whether the message received is false or true, but does know the payoff set.[16] The experiment has two treatments (High and Low); they differ only in the probability of the blue circle, which is 0.25 in Low, and 0.75 in High (this probability is always common knowledge in the corresponding treatment).

We conducted 20 computerized sessions (10 High and 10 Low) at the Universidad Autónoma de Madrid, with a total of 258 participants. The sessions were

---

[16] This distinguishes our study from "deception games" (e.g. Gneezy, 2005), in which the receiver does not know the payoff set. In those studies, the receivers' ignorance eliminates the concern that the sender's decision is influenced by his knowledge that the receiver knows whether the action was harmful, leaving the harm intact. In our study, this concern does not arise because the receiver is not harmed by the sender's choice. Furthermore, the receiver's knowledge of the sender's incentives plays an integral part in the experimental treatments, as these incentives induce the expectations we attempt to manipulate with our (High and Low) treatments, described below.

conducted in two waves, the first (106 subjects) in November 2010, the second (152 subjects) between September and October 2011. The software used for our sessions was z-Tree (Fischbacher, 2007). Participants were students from different disciplines, and the distribution of disciplines was similar in both treatments (Chi-square(7) = 7.583; p = 0.371).[17] Participants were not students of the experimenters. After being seated at a visually isolated computer terminal, each participant received written instructions that described the game (see Appendix I). Subjects could read the instructions at their own pace and we answered their questions in private. We used neutral language and avoided terms such as "lie". Understanding of the rules was checked with a control questionnaire that all subjects had to answer correctly before they could start making choices.

The instructions attempted to diminish potential demand effects or other confounds. For instance, we stated that this was an experiment on decision-making and that "there are no tricky questions, you must simply choose as you prefer". A potential motivation by any subject to behave so as to 'please' the experimenter, therefore, arguably put no constraints on her choice. Additionally, the instructions did not contain any indication to be truthful. Finally, we speculated that subjects might tell the truth not because they dislike lies, but to increase the aggregate rate of truth-telling in our study. Subjects might think that a low rate of truth-telling, if published, could have detrimental effects on the credibility of messages in our society (a public good), and hence the efficacy of communication. To reduce this potential effect, we informed subjects that their session was part of a large study with more than 40

[17] This is important because, as we show in a short note accompanying this paper (López-Pérez and Spiegelman; 2011), there exists a correlation between honest behavior and the subject's discipline. Since we have a similar distribution of disciplines in both treatments, we can be sure that any potential difference in behavior across treatments is not due to differences in the subjects' studies. We further note that there were not significant differences across treatments in the average values for political position (p = 0.683; Mann-Whitney test), gender (p = 0.452), or religiosity (p = 0.165).

participants. In this manner, they could ascertain that any individual choice was going to have a small effect on the aggregate.

Participants were anonymously matched in pairs. Before their roles (A/B) were randomly determined, all chose as if they had role A. Since the B-players are totally passive, this cannot affect their choices afterwards.[18] We used the strategy method to elicit the decision; that is, before knowing the actual realization of the random signal, subjects indicated what message they would send for each contingency (blue/green).[19] This method maximizes the amount of data gathered, provides information which facilitates the test of the theories, and permits the elicitation of subjects' beliefs in a manner that facilitates comparisons across treatments.

We elicited two beliefs from all subjects immediately after they had indicated the messages they would send. First, we asked each subject to estimate the percentage of all subjects who chose to send the message "green" when the signal was blue – in other words, their expectations about deception when the signal was blue. We will refer to this number in what follows as a subject's first-order belief. Second, we asked each subject to estimate the average percentage estimated by all subjects in the previous question. We call this estimation a subject's second-order belief – according to act-based guilt, this belief should be correlated with A's decision.[20] Both first and

---

[18] One could think of an alternative design in which the A-players send messages to the experimenter, and hence there is no need for the B-players. In this case, however, act-based guilt predicts that the subjects' second order beliefs about the experimenter's expectations should affect their decision. Controlling for such beliefs could be difficult. In addition, the degree (and relevance) of lie-aversion could depend on the status of the recipient.

[19] In principle, the strategy method might induce different behavior than the specific-response method, where participants know the realization of the signal. We have run a control treatment to check for possible effects of the strategy method, and as we discuss later, we observe no significant effect. We also note that Brandts and Charness (2009) review the experimental studies that use both methods and find no treatment differences in most of them.

[20] More precisely, act-based guilt predicts a correlation between (i) A's choice and (ii) A's belief about B's belief about A's choice: $E_A[E_B[\text{signal} = \text{blue} \mid \text{message} = \text{"green"}]] = \Pr[\text{blue} \mid \text{"green"}]$ in

second-order beliefs were elicited in an incentive-compatible manner, as we paid 3 Euros when the absolute error was less than or equal to 5 percentage points.[21] Since beliefs were mentioned after subjects had made their choices, the belief elicitation could not affect them. Only after beliefs were elicited, one subject in each pair was randomly selected as the real A, the other as B. The color of the circle was generated based on the relevant probabilities (High or Low), the actual A-player informed of the color, and the message previously selected by A sent to B. At the end of the experiment, subjects answered a brief questionnaire which included some socio-demographic information and a question about their reasons for their message choice when the circle was blue. Subjects were paid in private by an assistant who was not informed about the details of the experiment. Each session lasted approximately 40 minutes, and subjects earned on average 12.70 Euros.

2.4.1 Discussion

The goal of our design is to investigate the relevance of pure lie-aversion, eliminating confounds from other motives, and in particular controlling for act-based guilt-aversion. To understand this, it is convenient to mention the predictions by several relevant utility theories. First, it is clear that a selfish player A would always announce 'green' in any treatment. Second, the same prediction is shared by any theory assuming that people are altruistic or that communication affects altruism, although for a different reason. Altruism can reduce lies if they are expected to harm the receiver if trusted, but in our design B's payoff is not affected by A's choice, so that there is no altruistic reason ever to announce 'blue'. Third, note also that truth-telling cannot be motivated by inequity aversion as in Fehr and Schmidt (1999) or

---

equilibrium . Since subjects do not know with whom they will be paired, the average percentage asked in this question provides a measure of Pr["green" | blue], which will correlate with (ii), and hence should also correlate with (i), according to the theory.

[21] First-order (second-order) beliefs were paid only if the subject was later selected into role B (A). We did this in order to avoid payoff asymmetries. Our belief-elicitation protocol is simple and rather easy to describe in instructions, and is not marred by any hedging problem.

Bolton and Ockenfels (2000), as these models assume that aversion to advantageous inequity is never so strong that A-players would 'hurt' themselves to reduce it, thus choosing (A, B) payoff allocation (14, 10) instead of (15, 10). Fourth, payoff-based guilt-aversion as in Charness and Dufwenberg (2006) cannot explain any "blue" messages either. This is because it is common knowledge that B's payoff is constant across messages, as are, therefore, A's second-order expectations.

Consider now lie-aversion. Clearly, a lie-averse player will tell the truth in any treatment i.e., announce 'blue' ('green') if the signal is blue (green) if the utility cost of lying is large enough. Given the payoff constellation in our study, it seems safe to assume that this will be the case for some types.[22] Further, the probability of appearance of the blue signal is irrelevant for a lie-averse player, so that this theory makes the following prediction:

Prediction LA (lie-aversion): The rate of truth-telling (and hence also of lying) will be the same across treatments.

With respect to act-based guilt, we can apply axioms 1 and 2 introduced in Section 2. Let p_B denote the probability of a blue signal and consider an A-player who observes the green signal. From axiom 1 and the definition of $G_A(z)$, it is clear that such player will never lie, i.e., announce 'blue', as he can get a higher payoff and suffer no guilt by telling the truth. It follows that message 'blue' will always be trusted, so that the second-order belief after sending message 'blue' is $\mu^B = 0$. For an A-player who observes the blue signal, in turn, let μ^G denote the second-order belief that A has lied if she announces 'green'. Taking into account $\mu^B = 0$ and axiom 2, and assuming that there exists some measure of A-players for any possible value of

[22] If the cost of telling the truth was higher, some lie-averse types could decide not to tell the truth. We would be unable, therefore, to provide an accurate estimation of the percentage of subjects who dislike lies –i.e., of the relevance of pure lie-aversion.

Δ, it follows that for any μ^G there is a strictly positive fraction f of A-players that will say 'green' when the signal is blue (i.e., lie). By Bayes' theorem, therefore, the chance that message 'green' is a lie is

$$\Pr\left[blue \middle| \text{"green"}\right] = \frac{f \cdot p_B}{1 - p_B + f \cdot p_B}. \tag{1.4}$$

This has to coincide with μ^G if beliefs are consistent. Since this conditional probability depends positively on p_B, it follows that people should expect more lies as p_B rises. This will reduce the guilt associated with lying, and therefore increase the lie rate.[23] A simple example of a perfect Bayesian equilibrium that implements this phenomenon can be found in Appendix II. The general prediction is the following:

Prediction ABGA (act-based guilt-aversion): The rate of truth-telling will be positive in both treatments, but lower in treatment High; in other words, the High treatment will display the highest rate of lying.

There are two intuitions behind this result: (i) lying by A is more likely if she believes that B expects a lie with high probability, and (ii) B expects a lie with higher probability in High because the blue signal is more likely in that treatment, and therefore there are more occasions to get a larger payoff by lying after the blue signal (recall that lies are only predicted when the signal is blue; otherwise they are costly). Since senders in this theory tend to do what receivers expect them to do, a decrease in truth-telling follows. In summary, a constant rate across treatments allows us to reject act-based guilt as an incentive in this context, and moreover provides by elimination of other factors an estimation of the relevance of lie-aversion.

---

[23] Notice that f will therefore change as well. This will have a complementary effect, as (1) also rises in f.

2.5 Results

Player A has four possible pure strategies in the game. Denoting them as the message sent upon seeing a green (G) and blue (B) circle, respectively, they are: "payoff maximizing" (G, G); "honest" (G, B); "mythomaniac" (B, G); and "payoff minimizing" (B, B). Table 1 indicates the percentage of subjects who played each strategy in each treatment (High/Low), and in aggregate.[24] Note that we could obtain this information because we used the strategy method in the experiment. As we can see, the most frequent choices in both treatments correspond to strategies (G, G) and (G, B), while other strategies are much less frequently chosen.[25]

Table 2.1 Percentage of choice of each strategy in each treatment

	Strategies				
Treatment	(G, G)	(G, B)	(B, G)	(B, B)	Total
High	47.0 %	39.4 %	2.27 %	11.4 %	100%
Low	54.8 %	38.1 %	2.38 %	4.76 %	100%
Aggregate	50.8 %	38.8 %	2.33 %	8.14 %	100%

Note: N = 132 and 126 in treatment High and Low, respectively.

Our design includes two controls to discriminate between lie-aversion and act-based guilt. For the first control, let $f(S)_T$ denote the frequency of choice of strategy S in treatment T (T = H, L). According to act-based guilt only strategies (G, G) or (G, B) should be chosen, and moreover (Prediction ABGA), the null hypothesis $f(G, G)_H \leq f(G, G)_L$ should be rejected in favor of the alternative $f(G, G)_H > f(G, G)_L$. As Table 1 indicates, this will not be possible. Pure lie-aversion, in contrast, predicts no

[24] We pool the data from the two waves of subjects (November 2010 and September-October 2011), as they are statistically identical in terms of their strategy choices. A Chi-square analysis of the joint distribution fails to reject independence (d.f. = 3; stat = 2.637; p-value = 0.451)

[25] None of the theories so far considered in this paper can explain why some small fractions of the subjects chose the payoff minimizing strategy (B, B) or the mythomaniac one (B, G) in both treatments. We discuss this issue later.

difference in the rate of choice of the strategies (G, G) or (G, B) across treatments. As observed, no difference is found and a Mann-Whitney test fails to reject hypotheses $f(G, G)_H = f(G, G)_L$ ($p > 0.2$), and $f(G, B)_H = f(G, B)_L$ ($p > 0.8$). Similar results follow from a Chi-square analysis if we restrict attention to the strategies (G, G) and (G, B), the only two predicted by our theories;[26] the Pearson Chi-Square statistic is 0.495 (DF = 1, p-value = 0.482), while Fisher's Exact test yields a p-value of 0.509.

As a second control, we can use the subjects' first and second-order beliefs about deception. To analyze this issue from a theoretical point of view, we start by assuming that beliefs correctly anticipate behavior, as usual in equilibrium analysis. Since act-based guilt predicts a higher rate of choice of (G, G) in the High treatment, it correspondingly predicts higher expectations of deception in that treatment. Lie-aversion, in contrast, predicts no difference in behavior and therefore, in beliefs across treatments. Table 2.2 reports data about subjects' average beliefs in each treatment and in aggregate; these are remarkably constant at around 70%. Unsurprisingly, a Mann-Whitney test indicates that neither first-order ($p = 0.81$) nor second-order ($p = 0.97$) beliefs are significantly different across treatments.[27]. Hence, the evidence seems inconsistent with act-based guilt-aversion and more in line with lie-aversion.

Table 2.2 Average beliefs about deception in each treatment

Treatment		Total (N=258)
High (N=132)	Low (N=126)	

[26] A Chi-square analysis of Table 1 is invalid because the expected counts of the payoff minimizers are too low.

[27] Similar tests also reveal that the two waves of subjects are identical on first-order expectations ($p = 0.838$) and second-order expectations ($p = 0.990$).

	Mean (S.E.)	Mean (S.E.)	Mean (S.E.)
Average first-order beliefs	70 (2.50)	69 (2.61)	70 (1.80)
Average second-order beliefs	70 (2.28)	71 (2.30)	70 (1.62)

Note: S.E. = Standard error of the mean. Means have been rounded to two significant digits. First-order beliefs reflect the answer to the question: *"What percentage of subjects will send the green message on seeing the blue light?"* Second-order beliefs reflect the answer to the question *"What will be the average answer to the question above?"*

What if we drop the standard assumption that priors (i.e., beliefs) are common and correct, and assume instead that players have heterogeneous priors and play optimally given their own priors? In this case, act-based guilt predicts that player A's decision to lie after the blue signal should depend positively on his belief that the co-player expects him to lie. According to this theory, therefore, a participant in any treatment who reports a high second-order belief about deception is more likely to send message 'green' after the blue signal. Our data in Table 2.3 are emphatically in line with this: Pooled across treatments, subjects who planned to send the message 'green' when the circle was blue had on average a second-order belief of deception of 81.5 percent; the value for subjects who sent message 'blue' was 58.1 percent.[28] This difference is significant (Mann-Whitney $p < 0.0001$); interestingly, second-order beliefs are also significantly more dispersed (variance ratio test $p < 0.0001$).

Table 2.3 Summary statistics on second-order beliefs, by strategy.

Behavior	**N**	**Mean**	**SE**
Honest (G,B)	100	59	2.69
Minimizer (B,B)	21	53	6.32

[28] Note well that the first percentage refers to the subjects choosing either (G, G) or (B, G), whereas the second one refers to the subjects playing either (G, B) or (B, B).

Blue message overall	*121*	*58*	*2.48*
Maximizer (G,G)	131	82	1.58
Mythomaniac (B,G)	6	59	10.78
Green message overall	*137*	*81*	*1.62*
Total	258	70	1.62

Figure 2.1 provides further illustration of the correlation between beliefs and behavior. Each circle (squared) point in this figure represents a subject who sent message 'blue' ('green') after the blue signal in our treatments, placed according to her second- and first-order beliefs. We can see that the subjects choosing message 'green' are highly concentrated in the upper end of the scale, that is, second-order beliefs significantly predict the decision to lie after the blue signal.

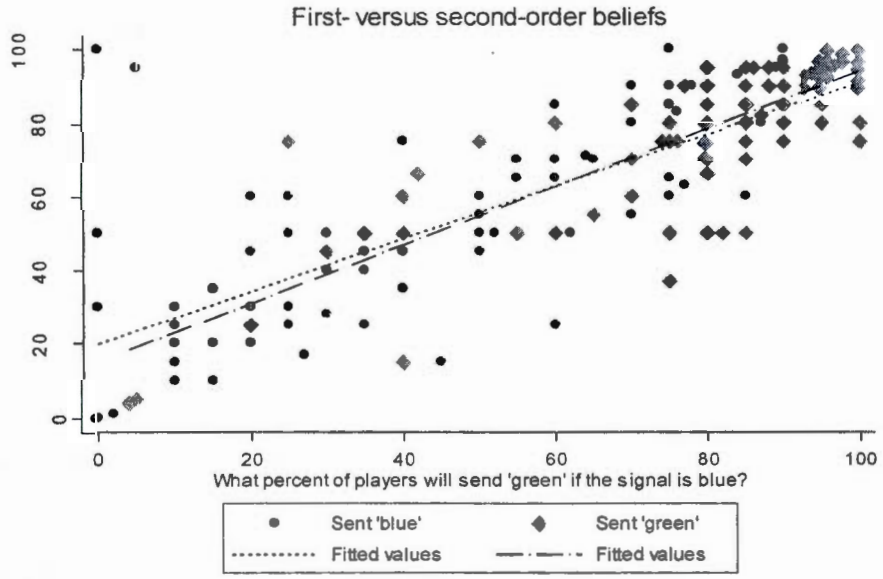


Figure 2.1 Beliefs on choosing each message after the blue signal

In addition, Figure 2.1 also shows by means of two regressions lines that first and second-order beliefs are correlated for each group of agents. Since we saw before that second-order beliefs are correlated with the decision to lie, it follows that first-

order beliefs are also correlated with the decision to lie. That is, people who expect many subjects to lie are also more likely to lie. For instance, pooled across treatments, subjects who chose the strategies (G, B) and (G, G) respectively expected that 54 and 84 percent of the participants would lie. This is again a significant difference; Mann-Whitney test, $p < 0.0001$. In summary, therefore, our data indicates a clear correlation between beliefs and behavior (see also Lundquist et al., 2009 on this), but first-order beliefs already capture this relation, and second-order beliefs do not appear to provide much more insight.

It is worthwhile to note that, while a relation between beliefs and behavior is predicted by act-based guilt-aversion, it is not incompatible with pure lie-aversion, simply because the latter theory makes no definite prediction in this respect once one allows for heterogeneous beliefs between players. However, our results suggest that a theory of lie-aversion might be complemented with some assumptions in this respect. For instance, in a context of heterogeneous priors, the data are consistent with the idea that people are averse to break norms of honesty, particularly if they do not expect others to break those norms, i.e., to lie (Bicchieri, 2005; López-Pérez, 2010; Cialdini, Reno, and Kallgren, 1990).[29]

In making the suggestion that lie-aversion and honest behavior interact with first-order beliefs, we are well aware of two caveats. First, even if the conjecture is correct, the measured association may not be valid. Participants in our experiment stated their beliefs after choosing their actions, and if those who lied themselves would "prefer to believe" that others were lying, too, then they might come to believe such thing in order to (unconsciously) avoid cognitive dissonance. Note however that this effect should be at least attenuated by the payment for accurate beliefs. Second, the association between honesty and stated beliefs is vulnerable to the now relatively

[29] If this were true, our design could underestimate the relevance of pure lie-aversion. In effect, some people could be lie-averse but lie in our experiment because they expect most others to lie as well.

well-known argument about the (false) consensus effect (e.g. Ross et al., 1977). According to this hypothesis, people project their own behavior, attitudes and beliefs onto others, tending to overestimate the extent to which others act and think the same way they do. Thus subjects who do not lie tend to think others don't lie either, and by extension that others expect them not to lie. Hence the beliefs do not drive the action; rather, some personal characteristic is driving both the beliefs and the action, generating a spurious relationship. In this respect, the strong correlation between first and second-order beliefs (Spearman's $\rho = 0.827$, p-value < 0.0001) shown in Figure 2.1 is consistent with the consensus effect.[30] Further research should clarify whether the relation between honest behavior and beliefs is just a result of this effect, or due to an interaction between lie-aversion and beliefs.

We continue by arguing that the results mentioned in this section are not an artifact of the use of the strategy method. Recall in this respect that subjects in the High and Low treatments had to indicate the message to be sent in any possible contingency; i.e., they made hypothetical decisions for both circle colors before discovering the true color of the circle. Given the hypothetical nature of these decisions, one might argue that emotions like guilt are less vivid in this case, and that could have an effect on behavior. To check for this, we ran a control treatment without the strategy method, which coincided with our High treatment in everything except that subjects made their choice after seeing the randomly selected circle color in their screens, and that beliefs about deception were conditional on having seen the blue signal. We focused on the High treatment simply because lies are most likely when the signal is blue; recall also that we found previously no significant behavioral differences across treatments.

---

[30] Yet we note that second-order expectations showed a "regression towards the mean." Thus, subjects with very high first-order expectations had second-order expectations systematically lower than the first, while subjects with exceptionally low first-order expectations "recognized" that others would guess higher than they. This is reflected in a regression slope significantly different from 1 ($p < 0.005$)

A total of 40 subjects participated in this control treatment, and the distribution of gender and major was similar to our two other treatments. In effect, a Mann-Whitney test of the hypothesis of equal gender distributions yields a p-value of 0.700, while Fisher's exact test of the equality of the major distributions yields a p-value of 0.796. Our main results in this section are replicated. First, among the 30 subjects who saw the blue circle in their screens, 40 percent sent the truthful message 'blue'. This is not significantly different than the analogous rate in the High treatment (Mann-Whitney $p = 0.909$). We also observe a correlation between honest behavior and beliefs. Subjects who sent a false 'green' message reported both first and second-order beliefs of deception that were significantly higher than those reported by subjects who chose to send a truthful 'blue' message (first order: $p < 0.001$; second order: $p < 0.005$). Table 4 shows subjects' average beliefs about deception depending on history of play (i.e. the color of the circle observed, and the message sent afterwards). First and second-order expectations were again strongly correlated ($r = 0.786$, $p < 0.0001$).

Table 2.4 Average beliefs, conditional on history of play

Circle color	Average beliefs	Message sent		Mann-Whitney p
		Green	Blue	
Blue	first-order	90	58	0.0008
	second-order	90	63	0.0010
	N	18	12	
Green	first-order	81	52	0.087
	second-order	81	68	0.290
	N	8	2	

Note: Mann-Whitney test of equality of beliefs across messages sent.

We finish this section with a brief discussion of the behavior of the 8.14% of subjects who chose the 'minimizing' strategy (B, B). One potential explanation of this behavior is that those subjects were trying to avoid the receiver's suspicion that they might be lying, perhaps because they expect that such a suspicion would bring

disapproval (on disapproval-aversion, see López-Pérez and Vorsatz, 2010). Some evidence points in this line. First, this theory predicts that choice (B, B) should be more frequent in the High treatment. Since the blue (green) message is likely to be trusted in the High (Low) treatment, disapproval-averse agents should choose (B, B) in High and (G, G) in Low. Indeed, the only significant effect that the different treatments seem to have induced is a (marginal) difference in the minimization behavior (Mann-Whitney test; p-value = 0.053). Second, we asked the subjects at the end of the experiment to write an open-form reason for their message choice if the circle was blue. Frequently, the justifications of those subjects choosing strategy (B, B) made reference either to the probability of the blue signal or to the B-party. For instance, one subject choosing (B, B) in treatment High justified sending the blue message after the blue signal in the following manner: "The color was actually blue and moreover the blue circle had a likelihood of appearance of 75% so that participant B would consider me sincere with 75% of probability". Other examples included "the blue circle was the most likely to appear" [from treatment High], and "since the green circle is most likely to appear, B would very likely think that the green circle would appear" [from someone choosing strategy (G, G) in treatment Low].

2.6 Conclusion

This paper reports the results from an experiment investigating whether pure lie-aversion affects truth-telling. Our design allows us to discriminate between this and other potential motivations for truth-telling that have been considered in the literature. Participants in our design know that uttering a lie will increase their own money payoff and at the same time will inflict no harm on anybody, or affect anyone's payoff expectations. Further, they are isolated from each other: They do not know anything about their co-player and their decisions are anonymous. As a result, nobody should tell the truth in our setting for altruism or to shape the receiver's payoff expectations. In contrast, people could tell the truth if they dislike lies. We consider

two variants of this idea: (i) Pure lie-aversion and (ii) act-based guilt-aversion. The first predicts truth-telling if the cost is low, irrespective of other variables. The second one implies that people will tell the truth if it is not very costly and moreover they believe that others expect truth-telling from them. Our two treatments permit us to discriminate between these motivations.

Our main results are the following: (1) Overall, nearly 40% of the subjects choose the strategy consistent with pure lie-aversion; (2) we find no significant evidence for act-based guilt aversion; (3) there is a correlation between beliefs and honest behavior, so that people telling the truth expect a higher fraction of others to tell the truth as well. These results suggest that pure lie-aversion is a widespread motive, possibly influenced by beliefs (as suggested by Bicchieri, 2005), and have implications for understanding behavior. For instance, surveys are often used to explore societal trends, and questionnaires are also employed in experiments (such as the current one, for example). Responders often get no reward for their answers, so that one could expect them to answer in a random manner and hence consider their answers as simply 'hot air'. Yet our study suggests that some responders might tell the truth even if they suffer a small cost, so that truth-telling should arguably be more pronounced if it involves no cost (as in most surveys, if not all). Of course, other factors may affect responses. A desire for privacy or an aversion to disapproval from the people running the survey may lead to biased responses to sensitive questions. In addition, respondents may not perfectly recall the information the questions require. For instance, the reasons that subjects give to justify past decisions may be psychologically distinct from their motivations at the moment of action. While these factors should not be ignored, our results suggest at least that a complete disregard for surveys or questionnaires is not warranted.

Our results might help to understand previous experimental results. For instance, Peeters et al. (2007) note in a repeated game that some subjects tell the truth

in most rounds but not always, which seems inconsistent with pure lie-aversion. However, if the behavior of the lie-averse types depends on their first-order beliefs and these beliefs change with repetition, we could observe some lie-averse players who change their behavior accordingly. Finally, our results might also provide a benchmark for new experiments. In this respect, we propose three questions with which we hope to suggest future experiments. First, how strong is lie-aversion? Truth-telling in our study was cheap (just 1 Euro), would it decrease radically if its cost increases to, say, 5 Euros? Second, do people dislike telling lies or do they enjoy telling the truth (Sánchez-Pagés and Vorsatz, 2009)? One can distinguish between these accounts in a slight variation of our basic game where player A also has the option to remain silent, in which case she gets 15 Euros (B always gets 10). When the circle is blue, a lie-averse player would maximize her utility choosing silence, whereas a player who (sufficiently) enjoys telling the truth would choose the 'blue' message. Third, neuro-economic research (e.g. Zak, 2011; Sommer et al, 2010) suggests that moral cognition contains emotional (quick, instinctive and crude), as well as reasoned (slow, deliberative and sophisticated) components. It is conceivable that these components manifest themselves as different kinds of "other-regarding" preference. For instance, the aversion to lying seems instinctive and emotional, which could explain its prevalence. On the other hand, theories such as guilt aversion, which require some interpretation of the co-player's expectations, may be part of the reasoned moral arsenal. One could hence think that such 'reasoned' factors have an effect only if conveniently primed by the context.

Bibliography

Battigalli, Pierpaolo, and Martin Dufwenberg, (2007). "Guilt in Games", American Economic Review, 97(2), 170-176.

_____ (2009). "Dynamic psychological games", Journal of Economic Theory, 144(1), 1-35.

Baumeister, R., A. Stillwell, and T. Heatherton (1995). "Personal Narratives About Guilt: Role in Action Control and Interpersonal Relationships", Basic and Applied Social Psychology, 17, 173–198.

Bicchieri, C. (2005). *The Grammar of Society*. Oxford University Press.

Bohnet, Iris, and Bruno Frey (1999). "Social Distance and Other Regarding Behavior in Dictator Games: Comment", American Economic Review, 89, 335–339.

Brandts, Jordi, and Gary Charness (2009). "The Strategy versus the Direct-response Method: A Survey of Experimental Comparisons", mimeo.

Bolton, G. E., and A. Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition", American Economic Review, 90(1), 166-93.

Buchan, N. R., E. J. Johnson, and R. Croson (2006). "Let's Get Personal: An International Examination of the Influence of Communication, Culture and Social Distance on Other Regarding Preferences", Journal of Economic Behavior and Organization, 60, 373-398.

Carter, John R., and Michael D. Irons (1991). "Are economists different, and if so, why?" Journal of Economic Perspectives. 5(2), 171-177.

Charness, G., and M. Dufwenberg (2006). "Promises and Partnerships", Econometrica, 74(6), 1579-1601.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places." Journal of Personality and Social Psychology, 58(6), 1015-1026.

Crawford, V., and J. Sobel (1982). "Strategic Information Transmission", Econometrica, 50, 1431-1452.

Crawford, V. (1998). "A Survey of Experiments on Communication via Cheap Talk", Journal of Economic Theory, 78, 286-298.

Dufwenberg, M., and U. Gneezy (2000). "Measuring Beliefs in an Experimental Lost Wallet Game", Games and Economic Behavior, 30, 163-182.

Ellingsen, T., and M. Johannesson (2004). "Promises, Threats, and Fairness", The Economic Journal, 114, 397-420.

Erat, Sanjiv and Uri Gneezy (2009). "White Lies", mimeo.

Fehr, E. and K. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation", Quarterly Journal of Economics, 114(3), 817-68.

Fischbacher, Urs (2007). "z-Tree: Zurich toolbox for ready-made economic experiments", Experimental Economics. 10(2), 171-178.

Fischbacher, U., and F. Heusi (2008). "Lies in Disguise: An Experimental Study on Cheating", mimeo.

Gneezy, U. (2005). "Deception: The Role of Consequences", American Economic Review, 95(1), 384-394.

Gore, Edmond J. and O. J. Harvey (1995). "A factor analysis of a scale of shame and guilt: dimensions of conscience questionnaire", Personality and Individual Differences, 19(5) , 769-771

Hurkens, S., and N. Kartik (2009). "Would I Lie to You? On Social Preferences and Lying Aversion", Experimental Economics, 12, 180-192.

Kartik, N (2009). "Strategic Communication with Lying Costs", Review of Economic Studies, 76, 1359-1395.

Ledyard, J. (1995). "Public Goods: A Survey of Experimental Research", in J. Kagel and A. E. Roth (Eds.), *Handbook of Experimental Economics*, Princeton Univ. Press.

López-Pérez, R. (2010). "The Power of Words: A Model of Honesty and Fairness", mimeo.

López-Pérez, R. and M. Vorsatz (2010). "On Approval and Disapproval: Theory and Experiments", Journal of Economic Psychology 31, 527–541.

López-Pérez, R. and E. Spiegelman (2011). "Do economists lie more?", mimeo.

Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson (2009). "The Aversion to Lying", Journal of Economic Behavior and Organization, 70, 81-92.

Orbell, J., R. Dawes, and A. van de Kragt (1990). "The Limits of Multilateral Promising", Ethics, 100, 616-627.

Peeters, R., M. Vorsatz, and M. Walzl (2007). "Truth, Trust, and Sanctions: On Institutional Selection in Sender-Receiver Games", mimeo.

Rosenthal, R. (2003). "Covert Communication in Laboratories, Classrooms, and the Truly Real World", Current Directions in Psychological Science, 12(5), 151-154.

Ross, L., D. Greene, and P. House. (1977). "The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Processes", Journal of Experimental Social Psychology, 13, 279-301.

Sally, D. (1995). "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992", Rationality and Society, 7(1), 58-92.

Sánchez-Pagés, Santiago and Marc Vorsatz, (2009). "Enjoy the Silence: An Experiment on Truth-Telling", Experimental Economics, 12(2), 220-241.

Sommer, Monika, Christoph Rothmayr, Katrin Döhnel, Jörg Meinhardt, Johannes Schwerdtner, Beate Sodian, Göran Hajak, (2010) "How should I decide? The neural correlates of everyday moral reasoning." Neuropsychologia 48, 2018-2026.

Sutter, Matthias (2009). "Deception through telling the truth!? Experimental evidence from individuals and teams", The Economic Journal, 119, 47–60.

Tangney, J. P., and R. L. Dearing, (2002). *Shame and Guilt*. New York: Guilford Press.

Tilghman-Osborne, Carlos, David A. Cole and Julia W. Felton (2010). "Definition and measurement of guilt: Implications for clinical research and practice", Clinical Psychology Review, 30, 536–546.

Vanberg, C. (2008). "Why do people keep their promises? An experimental test of two explanations", Econometrica, 76 (6), 1467-80.

Zak, Paul J. (2011). "Moral Markets", Journal of Economic Behavior & Organization, 77(2), 212-233.

APPENDICE C

EXPERIMENTAL INSTRUCTIONS

Thank you very much for participating in this experiment, which is financed by a research fund. Our aim is to study how people make decisions. In total, more than 40 people will participate in this study, in several sessions. There are no tricky questions, you must simply choose as you prefer. At the end of the experiment, you will be paid some money; the precise amount will depend on chance and your decisions during the experiment. It is very important that you do not talk to any other participant. If you do not follow this rule we will have to exclude you from the experiment and you will not earn any money. If you have questions, please raise your hand and we will assist you.

Description of the Experiment

In this experiment there are two types of participants (A and B). The basic task of each A is choosing a message for B. More precisely, towards the end of the experiment, A's computer will randomly reveal either a blue circle or a green one – the probability of a blue circle is [75% in the High treatment, 25% in the Low

treatment]. A will observe the circle in the screen and then send to B one of the following two messages: (i) 'the blue circle has appeared', or (ii) 'the green circle has appeared'. Payoffs are as follows: A will always get 14 Euros if he/she announces that the blue circle appeared and 15 Euros if he/she announces green. The payoff of any B is 10 Euros in any case. We remark that B will not observe the colour selected by the computer, but only receive A's message.

Since we want to know the message that you would send in any possible contingency, we will proceed according to the following protocol. To start, each of you will choose as if you were an A-participant. In addition, *before* knowing the color (blue/green) selected by the computer, you must indicate the message that you would send to B in two possible cases: (a) if the blue circle were selected and (b) if the green circle were selected. Afterwards, each of you will complete a short and anonymous questionnaire. Only then will your actual type be randomly determined (A or B with probability 50% each) and revealed to you. Moreover, each A-participant will be randomly matched with a different B-participant. If you happen to be A, you will see the color of the circle in the screen, and your corresponding message will be sent to B. If you are chosen to be B, you will receive the message chosen by A, and your previous responses to (a) and (b) will have no effect. Note well that you will never know the type of any other participant, nor will any other participant get to know your type. The decisions in this experiment are anonymous, that is, no participant will ever know which participant made which choice. For this reason, no participant will know the identity of the person with whom he/she is paired.

The experiment will end with another short and anonymous questionnaire. Your payment will be made in private in an adjoining office by an assistant unrelated to this study. This assistant will only know your final payoff in this experiment, but not what you actually chose in the experiment.

Before we start the experiment, please answer the following questions. Raise your hand when you are done so that we can verify the answers.

In a hypothetical example, assume that the computer selects color _____ (choose either 'green' or 'blue' to construct your own example), and that A had decided in that case to send message _____ (choose either 'green appeared' or 'blue appeared' to construct your own example).

For this hypothetical example, answer the following questions,

- What would A's payoff be? _____
- What would B's payoff be? _____
- What would A's payoff be, if she/he had chosen the other message? _____
- What would B's payoff be, if A had chosen the other message? _____

In addition:

- Will B ever know the color that actually appeared? Yes No
- After your type has been determined and provided that you are of type A, can you change your prior choices? Yes No
- Can A ever affect B's balance? Yes No
- Can B ever affect A's balance? Yes No
- Do A's choices affect A's balance? Yes No
- What is the probability that the circle is blue-colored? _____

APPENDICE D

EXAMPLE EQUILIBRIUM

Suppose that utility can be described by the function $U = x - \gamma \max[0, I - \mu^G]$, where x is the monetary payoff, I is an indicator for lying, μ^G is as defined in Section 2, and γ is an idiosyncratic sensitivity parameter randomly distributed over some non-negative interval according to a cumulative density function F. A simple extension of the PBE concept requires player A to maximize utility given beliefs, and that A use Bayes' rule, where possible, to define player B's (second-order) expected beliefs. Predictions are as follows. First, never lie when the circle is green. Second, since the monetary gain from lying is 1 in our game, it follows that A will lie on seeing the blue circle if $\gamma < 1/(1 - \mu^G) \equiv \gamma^*$. Thus, as long as the support of γ includes this value, Bayes' rule will always be defined. The fraction of A-players who will lie can be written as $f = F(\gamma^*)$. Combining this with (1.4) in Section 3 implicitly defines f as an increasing function of p_B. For instance, if γ distributes uniformly on [0, Γ], for Γ sufficiently large, then an interior solution will satisfy $f = \frac{1 - p_B}{\Gamma(1 - p_B) - p_B}$, which rises with p_B. It may be interesting to note that under the (restrictive) assumptions above, the previous expression permits one to estimate the upper bound of the distribution, Γ. The value of f can be estimated from the data, and p_B is imposed in the treatment. The estimate will be

$$\hat{\Gamma} = \frac{1}{f(G,G)} + \frac{p_B}{1 - p_B} \qquad \text{(D1)}$$

where f(G,G) is the proportion of the sample that chose that strategy. Based on our (High, Low) lie rates of (0.54386, 0.58974), we can calculate estimates of the upper bound of 4.84 and 2.03, respectively. Since the assignment to these groups was random, such a large difference in preferences seems unlikely. We therefore take this as evidence rejecting the model. Of course, we cannot determine whether it is act-based guilt, the distribution of γ, or the simple formulation of our utility function that is to blame for the rejection.

CHAPITRE III

THE BLIND AND THE BLINKERED: WHEN SELF-DECEPTION FRAMES MORAL CHOICE

Abstract: This paper models the organizational effects of one kind of rule-following behavior. The model features a behavioral (moral) rule prescribing costly effort, but to an extent that is a matter of some uncertainty. An Incumbent (she) is informed of the salience of the rule, and an uninformed Entrant (he), must try to learn from observing the Incumbent's choices. The Incumbent has no incentive to manipulate the information that the Entrant receives. However, she may try to *self-deceive* in a manner modeled as a signaling game she plays against herself. For the Incumbent, the model generates distorted signals, with the level of distortion increasing with moral sense. Analysis of the Entrant compares the case in which he must rely on the Incumbent's "words" – that is, the message she sends to herself – with that in which he must rely on her "actions" – the effort level furnished, and that where he has access to both indicators. Words are found to give more precise information than actions, but less than both together. Further, Entrant effort will be positively correlated with the Incumbent's moral sense whenever effort is observable, but negatively correlated when only messages are observable.

Résumé: Ce papier modélise les effets organisationnels d'un comportement caractérisé par une adhésion aux règles. Le modèle contient une règle comportementale (dite « morale »), qui prescrit de l'effort coûteux, mais jusqu'à un point qui n'est pas connu avec certitude, *a priori*. Un « ancien » (*Incumbent*) est informé de l'importance de la règle tandis qu'un « nouveau » (*Entrant*), non informé, doit essayer d'apprendre en observant les choix de l'ancien. Celui-ci ne veut pas manipuler l'information que le nouveau reçoit. Or, il peut essayer de *s'aveugler* d'une manière modélisée par un jeu de signal qu'il joue contre lui-même. Le modèle génère des distorsions de signaux qui augmentent avec la sensibilité morale. L'analyse du nouveau nous porte à comparer les cas où (a) il observe les « mots » de l'ancien – son message du jeu de signal -, (b) il voit ses « actes » - le niveau d'effort fourni - et (c) il voit les deux. Les mots donnent de l'information plus précise que les actes. De plus, l'effort du nouveau sera en corrélation positive avec la sensibilité morale de l'ancien dans tous les cas où l'effort est observable, mais en corrélation négative quand le nouveau ne voit que des messages.

3.1 Introduction

In many social situations, including the workplace, clubs, buses, elevators and the in-laws' house, people learn about the norms of appropriate behaviour not just from explicit, formal direction, but also from simply watching what other people say and do. We may have general ideas about the kind of behaviour that is accepted since in any culture there are principles which transcend any given situation and can be applied with some confidence to new contexts. However, there is also substantial variation in the level of acceptable behaviour across individual circumstances, and it is natural to try to pick up clues as to when one may use one's fingers when eating, or what kind of attitude to take with regards to a person making a toast, by means of surreptitious gleanings from other individuals. In a potentially more serious case, employees in an organization may be faced with various venal temptations, ranging from private use of company vehicles to accepting bribes. Particularly when they are in some doubt about whether or not these practices are "acceptable", they may attempt to interpret others' behaviour to form their opinions, without explicitly asking for direction. However common, the process of using clues to define acceptable behaviour is fraught with peril. In this paper, I focus on two specific kinds of danger the potential learner faces. First, there is no guarantee that the person from whom one takes his cues actually cares much for the rules at hand. If we admit that there are rules which are sometimes important to obey for their own sake, for instance, stock analysts' diligent reporting of the status of client companies, then it is possible that one may learn the wrong thing from watching an unscrupulous analyst. Furthermore, when following the rules for the task at hand is also costly in some way, then even those (analysts) who intrinsically care for the rules may try to self-deceive,

understating their importance to avoid the necessity of full compliance. This self-deception may contaminate other individuals who are attempting to form opinions based on watching the self-deceiver.

For the purposes of this paper, I examine these phenomena in a relatively stark form. A follower, called the Entrant (or he) in the model, learns about how rigorously a rule must be followed from observing the leader, called the Incumbent (she), without that Incumbent actually aiming to have such influence. There is no strategic interrelationship in the payoffs between players in this paper. They do not try to influence one another's behaviour at all[63]. Nevertheless, in equilibrium (some of) the participants do manage to influence (some) others. The mechanism for this channel of influence is close to the "herding" behaviour of Bikhchandani, Hirshleifer et al. (1992) (BHW). But while the basic Bayesian learning mechanism may be the same, the model in this paper is quite different from the BHW setup[64]. For instance, BHW focuses on interactions between identical individuals, while I allow heterogeneity of type. Moreover, I allow the possibility that the Incumbent may attempt to deceive herself about the rigor required. To illustrate the impact of this self deception, consider the comparison with effect identified as "cascading" in BHW. In BHW, agents sequentially adopt an action if it seems valuable, and reject it otherwise, having observed the previous decisions. Individuals late in a sequence may well ignore their own information if it is subject to error, and contradicts the bulk of what is already accumulated. In the situation in the model presented here, the first mover may distort her perceptions to yield an interpretation different from the truth. For

---

[63] While this runs counter to a strong intuition about social interactions that is without doubt important ground for future research, we will find that the model is complicated enough for one paper even in this highly pared-down form.

[64] For instance, I drop the two arguably most central features of their model: (a) a long sequence of players, (b) each of whom observes an i.i.d. signal. These are not relevant for my work, so I assume there are only 2 players, and only one observes the "signal".

example, she may try to "convince herself" that her chosen action is valuable where the evidence suggests it is not.

Such self-deception raises two important questions: why, and how? Each of these questions is certainly involved, and a full treatment is left to future work. As a beginning to understanding the framework for these question, I explore the idea that there is a fundamental ambiguity about the context of the in which these question arise. Prior works suggests that the ambiguity can arise from several sources. Bénabou (e.g., 2009) considers savoring utility, reflecting that the actual benefit of an action is uncertain because it is in the future. Experiments such as those of Dana, Weber et al. (2007) and Haisley and Weber (2008) introduce uncertainty in the payoff constellations. They conclude that generally, the more vivid or obvious the evidence for the true state of the world is, the harder it will be to self-deceive about it. The degree of precision in the information is almost certainly important in determining its vividness. Perfect information will be more vivid than imperfect information, and more precise imperfect information will be more vivid than less precise imperfect information. It may also be that there are some other qualities that affect the vividness of information. In the model below, I simplify these considerations, and consider two classes of information. "Vague" information is sufficiently imprecise, that individuals are able to self-deceive concerning it. By contrast, I define "vivid" information as that which precludes self-deception.

In addition to this uncertainty, self-deception requires that the self deceiver have preferences over the possible states of the world. If we allow any benefit at all from believing true propositions, then a person will only try to convince herself that state p is really state not-p if she strictly prefers state not-p. In the model below, this state-preference comes from a state-contingent moral obligation that I call the "Moral

Compatibility Constraint" (MCC)[65]. The MCC obliges people to exert a level of effort (a cost) that varies depending on the particular state. For instance, stealing is always wrong, but taking $100 left in an automated teller machine by some forgetful person might make you feel less bad than taking the money directly out of somebody's wallet. The aim of self-deception is essentially a relaxation of the MCC. The self deception can be considered a re-interpretation of the situation so that it requires a lower effort (cost). In the model below, the deception is operationalised in a signalling game that the self-deceiver plays against herself. The interpretation of this "message" is the "motivated reasoning process" (Mele 1997) by which she attempts to manipulate her beliefs. Her actions may also involve some observable behavior (rehearsing the desired beliefs, for instance) or the actions may be hidden. I thus model self-deception as an intentional action. Such a strategy should be considered an application of the standard "as if" assumption. Although many actual processes of motivated reasoning are likely unconscious in psychological experience, and hence seem outside the bounds of rationality, for the purposes of this paper I assume that these processes conform to what would happen if she intentionally manipulated her beliefs.

A longstanding question in the study of self-deception concerns how a rational individual may both (as self-deceiver) know proposition p and (as self-deceived) be fooled into thinking something different. In this model, I address this issue as a Perfect Bayesian Equilibrium (PBE) in which the Incumbent acts both as the sender and receiver of a message. She sends herself a message determined by an equilibrium function, and then interprets it based on the kind of actual situation that would lead her to send that message. Thus self-deception succeeds only in those cases where the

---

[65] The terminology is an allusion to the "Incentive Compatibility Constraint" (ICC) well known from Principal-Agent problems. Just as the ICC is a condition that "compels" the Agent to carry out the actions that the Principal desires, so this is a condition that "compels" the players to obey the rule. I do not, however, explicitly model a Principal who establishes this constraint.

equilibrium beliefs following reception of a message m are different from the state of the world which generates m, which implies that self-deception requires some pooling of the equilibrium messages. For self-deception to occur in the model, it must be because the message elicited could have come from more than one observed salience. Otherwise, even distorted messages will be perfectly decodable. In Section 2 I establish the extent to and conditions under which the Incumbent succeeds in such self-deception. In Section 3 I extend the model to the Entrant's rational interpretation of the Incumbent's observable actions. Section 4 offers some additional discussion and Section 5 concludes.

3.2 A model framework

3.2.1 Preliminaries

In this model, two players, an Incumbent (she) and an Entrant (he) face a moral rule of the type "take due care" or "tell no lies" or "read the papers you cite." The actual amount of care they take (Deffains and Fluet forthcoming), threshold "shade of gray" of lie they tell (Erat and Gneezy 2009), or importance of the paper they merely accept from a "reliable" reference list (Simkin and Roychowdhury 2003), is considered to be a continuous variable. It is denoted e for "effort exerted", and takes non-negative real values. As the examples above illustrate, this effort may have some external social value. However, it is costly to exert. For simplicity[66], I assume that the net marginal cost of effort is linear, normalized to 1. The cost to an individual of exerting effort level e is e, and therefore the individually rational level of effort to exert is 0.

The actual level of effort exerted is determined by the product of two parameters. The first, γ, denotes the individuals "moral sense" or inherent sensitivity to the rule, and the second, σ, denotes the general relevance of the rule to the current

[66] This assumption will necessarily shape the form of the equilibrium results. Generalizations are to other "well behaved" functions should yield qualitatively similar results, but are left to further work.

decision context. The parameters are viewed as independent random variables distributing on known functions $G(.)$ and $H(.)$, respectively. I will generally assume that both are uniform distributions. The support of G will be $[0,\gamma^H]$; for H it will be $[0,1]$.

Contrary to the view developed in, for instance, Bénabou and Tirole (2010), I assume that γ is vivid[67] – in other words, people are clear on their own moral position. No self-deception is possible with regards to γ. In the model presented here, the quality of the information about σ is one of the main differences between the Incumbent and the Entrant. The Incumbent observes σ as "vague" information; the Entrant does not observe σ at all. The interpretation of this assumption is that the Incumbent has more experience in the kind of dilemma at hand, and thus is able to better discriminate the salience of the rule. For the Entrant, all dilemmas look the same[68]. Thus the type of an Incumbent is two-dimensional, taking values in $\Gamma\times[0,1]$. The Entrant's type, by contrast, is scalar, taking values in Γ.

I stated above that the product of the parameters γ and σ determines effort. However, it will be noticed that γ is the only one that is known, in general, with certainty. The information that players have with respect to σ depends upon their type. Incumbents have perfect information initially, but this may be marred by their later self-deception. Entrants have no information initially beyond the prior distribution H, and must form an opinion based on observing the Incumbent. I assume

---

[67] I am very receptive to the contrary view, in which "self-image management" is a strategy to provide evidence of one's own good nature. My assumption, by contrast, means that people know what force the rules will have on their behavior, when the time to act arrives. This is not necessarily the same as their "true" nature.

[68] Compare this with Shavell (2002), in which the implication of moral rules to particular contexts is not always clear.

that Entrants do not self-deceive[69]. Rather, the entrant uses the information that can be extracted from the Incumbent's choices to form a Bayesian expectation of the value of σ, and this expectation then multiplies with γ to determine effort.[70] Denoting the final expected value of σ as $\mu(.)$, where "." refers to "whatever information is available to the player," the effort level chosen will therefore take the form

$$e(\mu(.)) = E[\sigma|.]\gamma \equiv \mu(.)\gamma \tag{3.1}$$

Expression (3.1) corresponds to "moral dumbfounding" (Haidt 2001) since it is a behavioural tendency that goes against the grain of natural desires, and is what I call the "moral compatibility constraint" (MCC). The MCC is the assumption that gives players an incentive to try to manipulate their interpretation of the salience of a clear moral rule. Notice also that as γ increases the incentive increases. The dependence of incentive on γ will generate a sorting condition comparable to the single-crossing condition. Although irrelevant for the Incumbent, who observes γ, this sorting will have importance for the Entrant, who does not.

Psychological studies have corroborated the intuition that manipulation of beliefs is costly. One cannot simply believe whatever one wants. For simplicity, I posit a (point) "target belief"[71] called *m*. Consistent with much literature, I adopt a

---

[69] This is done out of concern with tractability, not realism. The Entrant's process of inference will be relatively involved; I leave to future research the additional complexity of his self-deception.

[70] Throughout the paper I skirt issues of the "true" normative importance of the moral rule, and therefore the interpretation of this differential effort given the same moral rule remains a matter of "persistent uncertainty". Kant (2005) – see also Hausman and McPherson (1993), White, (2004). Van Staveren, (2007). This might suggest that the value of γ reflects the *appropriate* level of effort for that individual. On the other hand, it could also be that individuals of low moral sense simply don't care, and so exert less than the "appropriate" level. The difference is largely immaterial to the model; its relevance to interpretations of the results will be commented, where appropriate.

[71] This terminology is seen in the literature on moral reasoning to represent the belief that a person would like to hold. See for instance DePaul (1993). Reflection may upset this target belief, and the person may rationally end up believing something else, but the target is the belief that, for one reason or a balance of reasons, seems preferable. In the present context, it represents the result of an arbitrage between the cost of distorting the evidence and the cost of leaving the evidence intact.

simple cost function, assuming the costs of m to be quadratic in the distortion it represents, and proportional to a constant k, which I assume to be the same for all players. The value of k likely encompasses such considerations as the difficulty in "cooking the mental books" to direct one's attention away from undesirable evidence, or some other motivated reasoning mechanism.

These assumptions lead to a utility function that has the following form:

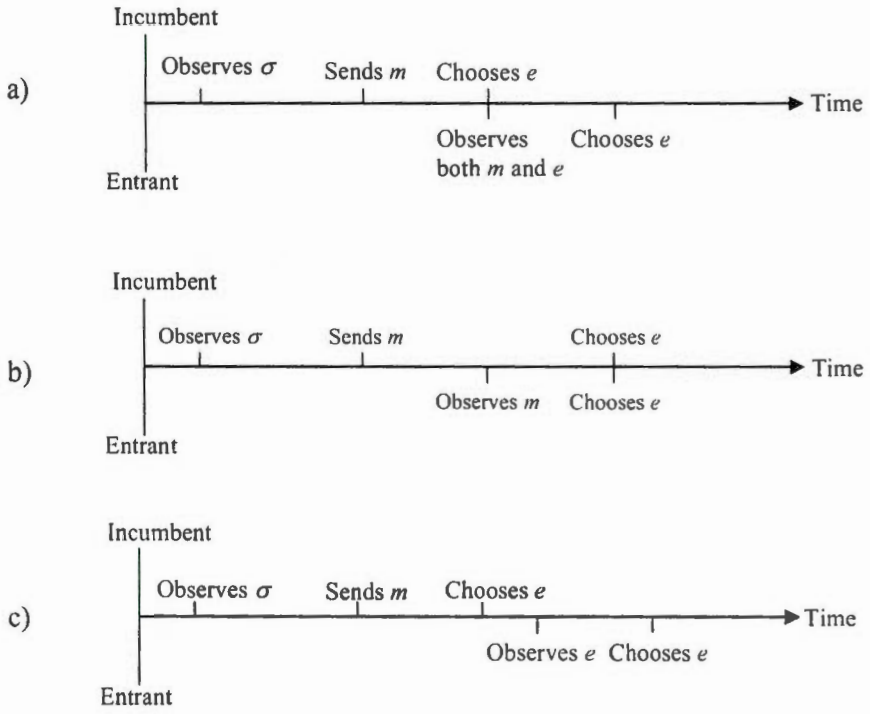
$$V = -e - \frac{k}{2}(\sigma - m)^2 \tag{3.2}$$

where e will be determined by the MCC and the constraint of Bayesian interpretation shown in (3.1). Notice that the single-crossing property applies, as those who observe a lower value of σ have a lower marginal cost of distortion for given $m < \sigma$. As mentioned, the information upon which e is based depends on whether the player is an Incumbent or an Entrant. It is further determined by the equilibrium strategies played by Incumbents.

The structure of the game is as follows. First, the Incumbent observes a "signal," σ, and her own γ. She then chooses a "message" m for herself, which may or may not be observable to the Entrant. She interprets this message using a function that constitutes the beliefs of a PBE. The equilibrium strategies therefore determine the expected value of σ given m, which together with γ then determines the Incumbent's effort according to (3.1). Again, the effort exerted may be public or private information.

The Entrant, having observed at least one of the Incumbent's choices (message or action), makes use of the strategies specified by the PBE to calculate an expected value of σ. Together with his γ_E, the expected value of σ determines the Entrant's effort level. As mentioned, I assume that the does not Entrant self-deceives. However, the information the Entrant can observe varies over three cases. Either the

Entrant observes the message, m, or the effort, e, or both. The Entrant never observes either σ or the Incumbent's level of γ, which I further assume to be completely independent from his own. The figure below illustrates the structure.



Note: In panel (a), the Entrant observes both m and the Incumbent's e before choosing. In panel (b), he observes only m and in panel (c) he observes only e. These correspond to the cases developed in Section 4, below.

Figure 3.1 Timing of the game.

Notice that the strategic interaction is entirely one-way in the game, as it is, for example, in BHW or Battaglini, Bénabou et al. (2005). This means that the decisions for the players can be calculated forwards in time. The Incumbent's decisions will determine the Entrant's reaction but the Entrant's reaction will not

affect the Incumbent actions. To briefly preview the results, I report the following principal findings. The Incumbent's shows downward distortion in her messages ($m < \sigma$) with a partially-separating equilibrium structure that, *for each* γ, is essentially identical to that observed by Kartik (2009), with a lower pool and separation above a threshold σ_γ. Higher values of γ send more distorted messages, leading to larger pools, and hence a smaller probability of separation. An important result for the results described below concerns the divergence between messages and signals. A higher-γ Incumbent will send lower messages than a lower-γ Incumbent for a given σ, but will produce more effort.

As for the Entrant, there are two essential kinds of questions. First is the question of contamination. How does the moral sense (and thus the self-deception) of the Incumbent affect the Entrant's behavior? In equilibrium there is always a correlation between the Incumbent's γ and the Entrant's effort in equilibrium. Interestingly, however, the direction of the correlation changes with the case considered. When e is observable (Cases (a) and (c) in Fig.1), the Entrant's effort is positively correlated with the Incumbent's γ, and the correlation is stronger in case (a) than in case (c). This is intuitive, since more information (e and m, rather than just e) is available. This correlation can be interpreted as the transmission of a kind of organizational culture. When only m is observable, by contrast, we see that the Entrant's effort will actually fall as γ rises. This is due to the fact that, for a given σ, at higher γ Incumbents send lower messages. Even though the Incumbent's lies do not always, and on average cannot, fool herself, they propagate throughout the less-experienced members of the group (here represented by the Entrant), reducing the effort exerted.

A second kind of question concerns the average, or overall differences in the informational cases (a) to (c). In terms of the effort provided, Incumbents are "Bayesian self-deceivers," who use all the information available in their manipulated

target beliefs, and thus on average have the same beliefs they would have without self-deception. Entrants, in turn, have no vague information, and hence cannot self-deceive at all. On the other hand, while the average behavior is the same in all cases, the beliefs that generate it are not. Perhaps unsurprisingly, Entrant's posterior distribution of σ is more dispersed in the partial information cases than in the full information case. More interesting, the posterior distribution of σ is also more dispersed in the "actions", effort-only case than in the "words" case. "Actions speak louder than words" is the saying, but words in this model speak more precisely.

3.2.2 Analysis of the Incumbent

The Incumbent's manipulation of beliefs corresponds to her interpretation of the rule. She would "prefer to believe" that σ is low, as this requires a lower level of effort e given γ according to the MCC. The question is what doubts she will be able to cast into her own mind as to the importance of the rule. As outlined above, I model this manipulation as an intra-personal persuasion game. The Incumbent's decision about the rule has two stages. In the first stage, the she observes σ and chooses a message $m(\sigma)$ or simply m from [0,1]. In supposing that the messages come from the same interval as the true values of σ, the game assumes that messages have an established literal meaning. This is the same tactic used in Kartik (2009) and Emons and Fluet (2009). In this last paper, the qualitative form of the equilibrium strategy is different than the one employed here because the message space there is unbounded. A subsequent version of the model with natural upper and lower bounds to the state-space (and hence also message-space), yields results very similar to those found here.

In the second stage of the model, the Incumbent no longer has access to σ, but only to her message. However, she recognizes that she has the tendency to send $m \neq \sigma$, and is Bayesian enough to attempt to decipher the "real" meaning behind her message. She therefore calculates a conditional expectation, which makes use of the endogenous, equilibrium interpretation function based on Bayesian reasoning,

denoted $\mu(m,\gamma)$. Stated differently, the informed player chooses a message for herself in an attempt to induce beliefs of a lower σ, but based on the knowledge that she will later interpret this message according to Bayes' law (where that is possible) and be forced by the MCC to play in accordance.

If in equilibrium observations of any σ in the (non-empty) set Σ_m lead to the message m, then

$$\mu(m,\gamma) = \frac{\int_{\Sigma_m} z h(z)\, dz}{\int_{\Sigma_m} h(z)\, dz} \tag{3.3}$$

If Σ_m is a singleton (that is, $m(\sigma)$ is invertible), then it will be identical to $\mu(m,\gamma)$. In this case I will refer to the belief as $\hat{\sigma}(m,\gamma)$. Further, because I assume that the Incumbent retains the knowledge of γ, and therefore that argument will be constant across all decisions any Incumbent might make, I will simplify the notation when possible, writing $\hat{\sigma}(m)$. If Σ_m is empty for any m, in other words if there are messages that are not sent in equilibrium, then Bayes' Rule is undefined for those messages, and beliefs are ambiguous, and will be chosen in a manner respecting the equilibrium while striving to be "sensible".

The equilibrium I will focus on has two parts, defined as follows by a strictly monotonic function[72] $m(\sigma)$ and a threshold σ_γ, both of which depend on γ,

$$m^*(\sigma) = \begin{cases} m(\sigma) & \text{if } \sigma > \sigma_\gamma \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

---

[72] This is the equivalent of the "separating function" from Kartik (2009), and I will occasionally refer to it by the same name.

$$\mu(m) = \begin{cases} \hat{\sigma}(m) & \text{if } m > 0 \\ E\left[\sigma \middle| \sigma < \sigma_\gamma\right] & \text{otherwise} \end{cases} \tag{3.5}$$

Let us take the upper, interior case first. The strategy will be defined so that

$$m(\sigma) = \arg\max_m -\gamma\hat{\sigma}(m) - \frac{k}{2}(\sigma - m)^2 \tag{3.6}$$

Given the continuity of (3.2), when the strategy function is monotonic it can readily be shown[73] that it must also be continuous. As a result, the belief function , $\hat{\sigma}(m)$ must also be continuous. Therefore we can take the first-order condition that

$$\hat{\sigma}'(m) = \frac{k}{\gamma}(\sigma - m) \text{ for all } \sigma \text{ where } m > 0 \tag{3.7}$$

Notice that the quadratic distortion cost structure implies that the second-order conditions are satisfied. The monotonicity of the strategy further implies the existence of an inverse function, which means the inference will be exact: $\hat{\sigma}(m(\sigma)) \equiv \sigma$ and expression (3.7) can be re-written

$$\hat{\sigma}'(m) = \frac{k}{\gamma}(\hat{\sigma}(m) - m) \tag{3.8}$$

The idea that in any separating equilibrium the least-preferred type will be identified, and thus have no incentive to distort – the Riley condition – in this case implies a boundary condition of $\sigma(1) = 1$. The single-crossing property (of the signal

---

[73] Any strictly monotonic strategy results in an invertible strategy function. This means that $\mu(m)$ is degenerate on some singleton σ for all m. Suppose a monotonic strategy is not continuous. Then there exists a σ' such that the limit of the strategy from the left is not the same as the limit from the right. This means that a deviation from $m(\sigma' - \varepsilon)$ to $m(\sigma' + \varepsilon)$ for some arbitrarily small positive ε will cause an arbitrarily small change in $\mu(m)$, but a discretely large change in (3.2), thereby generating an "unraveling" deviation chain towards the lower-cost deviation.

distortion) then implies that an observer detecting a lower σ will respond with lower messages. Such a property indicates the strategies will indeed be monotonic. Further, because (3.8) must hold for every value of m, we can solve it as a differential equation for a globally optimal strategy function for the expected value of σ, given m (Mailath 1987). The form of (3.8) implies an exponential-distortion solution,

$$\hat{\sigma}(m) = m + \frac{\gamma}{k} + e^{\left(\frac{km}{\gamma}\right)} C. \tag{3.9}$$

where C is the constant of integration. Using the boundary condition identified above then yields a closed-form interpretation function,

$$\hat{\sigma}(m) = m + \frac{\gamma}{k}\left(1 - e^{\frac{k}{\gamma}(m-1)}\right) \tag{3.10}$$

Expression (3.10) describes the interpretation of a message. The message sent will be $m(\sigma)$, its inverse. Several further useful features of the relationship described in (3.10) may be distinguished[74]. First, we may verify the monotonicty in m. Second, all messages are understatements since for all m in [0,1], $m(\sigma) \leq \sigma$, with equality only at $\sigma = 1$, for any positive values of γ.[75] Third, any observed σ less than a threshold $\sigma_0 \equiv \gamma/k(1 - e^{-k/\gamma})$ will associate with a negative m. This threshold is greater than zero for any value of $\gamma > 0$, and is increasing in that parameter, rising towards 1

---

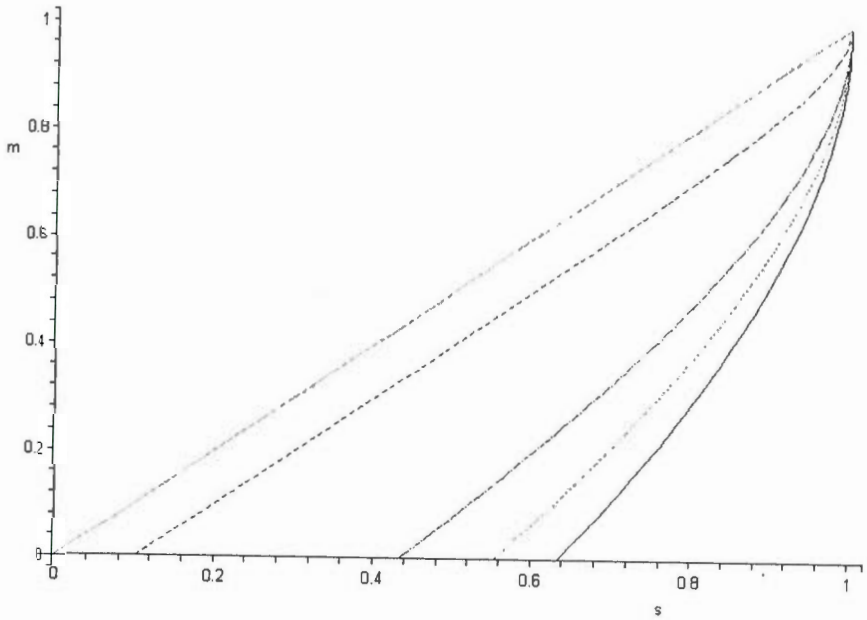
[74] One point upon which I will not dwell is that (3.10) implies that V decreases in γ and σ. This illustrates what Sen (1976) calls the "counter-preferential" nature of the behavioral restriction. However, I cannot make any actual welfare statements without further information about the normative value of the rule itself. And this may require a wider philosophical scope than I wish to set in this paper.

[75] Strictly speaking, these properties hold in the range of $\sigma \in [0,1)$. It may be noticed that technically, (3.10) also admits messages greater than 1, but is invertible only on [0,1). Allowing "neologisms" – messages from outside [0,1], would result in a "hyperbolic" equilibrium in the sense both of hyperbole and hyperbola, which I investigate no further here, despite the etymological interest the two terms suggest.

as γ goes to infinity[76]. This threshold limitation implies any individual with $\gamma > 0$ will hit the bound of feasible messages for some $\sigma > 0$, and those with higher γ will hit the bound at a larger value of σ than those with lower. Fourth, if we let γ vary, the interpretation given to a particular message rises with γ. This relationship implies that, given an observed value of σ, the equilibrium message falls in γ: reintroducing the full notation and denoting derivatives with subscripts, $m_2(\sigma,\gamma) < 0$, or in words, higher-γ Incumbents will respond to a given σ with lower messages than will lower-γ Incumbents. This is the other sorting condition, and can be interpreted to mean that γ is or the marginal incentive to lie. The condition is intuitive, since higher values of γ are constrained to exert higher levels of effort. Figure 3.2, below, shows how the separating function changes as γ rises in comparison to k. It can be seen as "level curves" in (σ,m,γ)-space, rising in σ and falling in m. We therefore see that the relationship is quasi-convex: the lower contour set of a given value of γ - the set of (σ,m) that imply sensitivity lower than γ - is a convex set.

---

[76] In the limit as γ goes to zero, this function falls to $m(\sigma) = \sigma$, which hits the horizontal boundary only at 0.



Note: (Dashed, dot-dash, dot, solid) curves show strategies as γ/k rises through (0.1, 0.3, 0.5, 0.75) as level-curves in (σ,m,γ)-space. Figure plotted using Maple software.

Figure 3.2 Separating function as γ changes.

Summarizing the points so far, we have the following observation:

Observation 1: The separating function $m(\sigma,\gamma)$ represents a downward distortion in the message that Incumbents report to themselves, and satisfies the following properties

a) *m is invertible on $[0,1)$; also $m(1) = 1$*

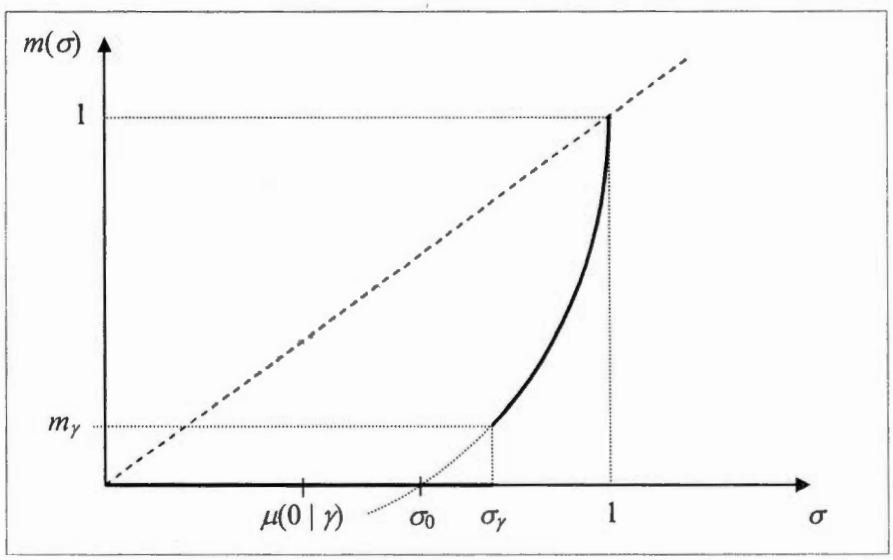
b) *$m_1 > 0$; $m_{11} > 0$; $m_2 < 0$; $m_{12} > 0$*

c) *For any $\gamma > 0$, there exists a $1 > \sigma_0 > 0$ such $m(\sigma_0,\gamma) = 0$.*

The equilibrium strategy for the Incumbent will consist of the separating function implied by (3.10) for all σ greater than a threshold I have called σ_γ, and $m = 0$, otherwise. Let us pass now to the second part of the strategy to consider the threshold σ_γ at which Incumbents switch to $m = 0$. The interesting point[77] here is that, for all $\gamma > 0$, $\sigma_\gamma > \sigma_0$. To see that this must be so, consider the interpretation of $m = 0$. Using the notation from (3.3), Σ_m is not a singleton for $m = 0$. In particular, it is the interval $\Sigma_0 = [0,\sigma_\gamma]$. This means that the interpretation of a zero message will be less than σ_γ. If the strategy switched at σ_0, then the limit of the interpreted value of σ given m as m goes to zero, which is σ_0, and is strictly greater than the interpreted value of $m = 0$. Naturally, this inequality cannot stand in equilibrium. It will lead marginal individuals who saw a σ just to the right of σ_0, and who are therefore prescribed an $m > 0$, which will be perfectly decoded, to deviate and also choose $m = 0$. As they do so they increase σ_γ, and hence the interpreted value of the zero-message. Also, as marginal σ-types deviate to $m = 0$, they must deviate farther from their optimal message. Thus the "benefit" of the deviation falls and the "cost" increases as σ_γ rises. Eventually, either all types will send the message $m = 0$, in which case $\mu(0)$ is just the unconditional expectation of σ, or there will be some type who finds that the utility from deviating is just equal to the utility of the positive message. For this latter type, the benefit of the lower interpretation is just balanced by the cost of the higher distortion. Figure 3.3, below, illustrates this latter case.

---

[77] Just as the separating function is like that in Kartik, (2009), so this "bunching" phenomenon is the same as in that paper. A similar phenomenon can also be seen in, for instance, Bernheim, (1994) and indeed is the same principle used to generate the thresholds in Crawford. and Sobel (1982).



Note: The highest signal σ at which this Incumbent will separate (σ_γ), and the corresponding lowest positive message she will send (m_γ), are also noted.

Figure 3.3 Illustration of "bunching."

The threshold value of σ_γ will be where the point of indifference between sending the separating message $m(\sigma)$ and the pooling message $m = 0$. The "benefit" side of the pooling decision is the lower interpretation of the salience, $\mu(0 \mid \gamma)$, the "cost" side is the larger distortion required. Re-introducing the explicit dependence of m on γ, at the threshold itself,

$$V\left(m\left(\sigma_\gamma,\gamma\right),\sigma_\gamma\right) \equiv -\gamma\sigma_\gamma - \frac{k}{2}\left(\sigma_\gamma - m\left(\sigma_\gamma,\gamma\right)\right)^2 = -\gamma\mu\left(0\right) - \frac{k}{2}\sigma_\gamma^2 \equiv V\left(0.\sigma_\gamma\right) \quad (3.11)$$

Equation 2.11 implicitly defines a relationship between m, γ and σ such that (a) for a given γ, only signals greater than σ_γ will elicit a positive message, with messages following the separating function; (b) for a given σ_γ, only Incumbents of type less than γ will send a positive message, with the message sent falling in γ; (c) for a given positive message, the signal must have been greater than σ_γ, and the Incumbent's type less than γ, again according to the separating function.

Notice that since $\sigma_\gamma > \sigma_0 > 0$, all positive-$\gamma$ Incumbents will necessarily have ranges of σ that lead to the message $m = 0$. *The "corner solution" branch of the equilibrium always exists.* The remaining question is therefore: what values of γ and σ that *will* lead Incumbents to play the separating function? The answer, detailed in Appendix A (Proposition 1d), is *the "interior solution" branch of the equilibrium exists for Incumbents with* $\gamma < k$. This fact motivates the assumption later on (see note 78, below) that the upper bound of the distribution of γ, which I have called γ^H, coincides with k.

A final point is required to fully characterize the Incumbent's equilibrium strategy. Beliefs are well defined by Bayes' Rule when $m = 0$, and for any $m > 0$ sent in equilibrium. However, the fact that there is a minimum positive acceptable message for any $\gamma > 0$ implies that, unless $\gamma = 0$, (which happens with measure 0, and in which case e always equals zero, so (3.2) is optimized by choosing $m = \sigma$), there will be some messages that are never chosen in equilibrium, and for which, therefore, Bayes' Rule does not define beliefs. I assign these messages the beliefs that would correspond to the separating function. In other words, (3.10) will supply the beliefs for any positive message in the equilibrium, whether that message be on or off the equilibrium path. In fact, this was implicit in condition (3.11), which compared deviations from the separating function to zero messages, and that condition guarantees that the equilibrium is sequentially optimal.

In summary, then, we can state the following Proposition:

> *Proposition 1: A single-player equilibrium of this model exists, defined by (3.4) and (3.5), where $\sigma(m)$ is defined as (3.10) and $m(\sigma)$ is its inverse, and σ_γ is as determined by the threshold (3.11). In this equilibrium, the Incumbent will (a) always send a message less than the observed value of σ, except in the limiting cases when $\sigma = 1$ or 0; (b) nevertheless be able to*

perfectly decode her message, and be forced to play the according e, whenever the message is greater than zero; (c) send a message of $m = 0$ if and only if $\sigma < \sigma_\gamma$ – at this point she will be pooling with other σ-types who (would have) sent the same message, and so will play $e = \sigma_\gamma/2$; (d) exhibit a threshold, and hence a range of separation, that grows in γ. However only those with $\gamma = 0$ will never pool on $m = 0$, for any value of σ, and only those with $\gamma < k$ will ever send $m > 0$.

Corollary 1: The effort level $e(\sigma,\gamma)$ is strictly increasing in the second argument, and non-decreasing in the first. Further, it is strictly increasing in both arguments when $\sigma > \sigma_\gamma$. The message sent $m(\sigma,\gamma)$ is non-decreasing in the first argument and non-increasing in the second. Further, it is strictly increasing (decreasing) in the first (second) argument when it is greater than zero.

Corollary 1 will be very important in the analysis of the Entrant's problem. It implies that there is an important qualitative difference between the two kinds of information that an Entrant can receive. In particular, a higher m can either mean a higher σ or a lower γ, while a higher e means a higher σ or a higher γ.

Before moving on to the Entrant's problem, I will add a note on the extent to which self-deception is possible in this model. Proposition 1 states that, for high-salience situations, the Incumbent will be forced, despite her "protests" in the form of downward-distorted messages, to recognize the importance of the rule for what it is, and play accordingly. Thus she can only self-deceive in relatively low-salience problems. In addition, however, it states that all low-salience situations are treated identically, as having the salience expected conditional on $\sigma < \sigma_\gamma$. This means that the effort provided in very low-salience situations, specifically, those less than $\sigma_\gamma/2$ in

the uniform-distribution case, will be treated as more important than they are. We therefore have a model in which the Incumbent can (a) "make mountains out of molehills", inflating the importance of inconsequential problems; (b) "make molehills out of mountains," deflating the importance of bigger problems; and (c) accept truly big problems for what they are. These results are summarized below:

> *Proposition 2: On average, Incumbents playing the equilibrium from Proposition 1 will exert "too much" effort for low-salience contexts ($\sigma < E[\sigma \mid \sigma < \sigma_\gamma]$), and "not enough" in higher-salience contexts ($E[\sigma \mid \sigma < \sigma_\gamma] < \sigma < \sigma_\gamma$). Also, both effects will be stronger for Incumbents with higher values of γ_I.*

3.3 The Entrant

I now consider the second player[78]. For each of the three cases illustrated in Figure 3.1, I will address three questions. First, what is the nature of the inference that the Entrant can make, given the information at his disposal? Second, what is the effect of a change in the Incumbent's (unobserved) moral sense, γ_I on the Entrant's behavior? The answer to this question addresses the issue of contamination, or hierarchical influence within the organization, and will provide the content of the most of the propositions in the section. The question will turn on the correlation between the Incumbent's moral sense and the Entrant's behavior. For example: can an organization with unusually principled Incumbents expect its Entrants also to show

[78] In this section I maintain another simplifying assumption, namely that $\gamma^H = k$. As noted above, this implies that (a) all Incumbents except the highest have play the interior branch of the equilibrium for σ high enough, and (b) for any $\sigma < 1$, there is a measure of Incumbents who will respond to it with $m = 0$. This limits the generality of the discussion, but the benefit in clearing away a "clutter of cases" is worth the cost.

higher levels of effort? The third question is what the nature of the Entrant's perception of the salience, after observing the messages and signals from the Incumbent.

To begin, recall that the Entrant's payoffs are entirely separate from those of the Incumbent. The only channel of influence between them is the information that the Incumbent passes on to the Entrant, and the Incumbent has no direct incentive to try to influence the Entrant's behavior with this information. This is why I can calculate the PBE for the Incumbent independently of the Entrant's decision, and need not readjust it to take account of this reaction. Further, I will assume that the Entrant does not self-deceive, and forms an expectation of σ based on observable choices by the Incumbent. Thus the Entrant's problem is similar to that of "followers" in BHW. The fact that he does not observe an independent signal simplifies the Entrant's decision, and yet he faces a more complex task of interpretation than the agents in BHW. Each distinct Incumbent will respond to a given σ differently. Thus there are an infinite number of Incumbents types who will produce any given message or effort level, for a corresponding interval of different saliences. The Entrant's problem, upon observing a message or an effort level, is to determine what the salience "probably" is, based on (i) what type of Incumbent might have sent that message or exerted that effort, and (ii) what kind of σ would induce her to do so. Thus the Entrant will generate the beliefs in his PBE about the expected value of σ based on Bayesian updating of the Incumbent's equilibrium play, given the observed behavior. The sequentially rational behavior will be defined as that constrained by the MCC, given the equilibrium beliefs. In contrast to the Incumbent, the Entrant's

beliefs are always defined by Bayes' Rule[79], so no off-path beliefs need to be specified.

3.3.1 Case A: *m* and *e* observable

The first result shows that observing *e* and *m* together is sufficient to inform the Entrant as fully as is possible, once the original salience has been overridden by the Incumbent's message.

> *Lemma 1: if the Entrant has access to both m and e, he can recover all the information available to the Incumbent at the time of action.*
>
> Proof: See Appendix A.

Stated otherwise, if the Entrant sees a positive message and the effort that follows it, then he will know exactly the level of σ that the Incumbent saw originally (and tried unsuccessfully to dissimulate). If the Entrant sees a zero-message along with effort by the Incumbent, he will be able to determine the exact threshold σ below which the true value must fall. In both cases, he has the exact same information as the Incumbent. The intuition for this result hangs on the dual-sorting behavior of the Incumbent. Of all the (σ,γ) Incumbents who might send a given (positive) message, each will follow it with a different level of effort, and the higher the effort, the higher the γ. Therefore, observing both the message and the effort allows the Entrant to identify precisely the Incumbent's bivariate type. Similarly, all the Incumbents with different γ will follow a message of $m = 0$ with different levels of

---

[79] Although each γ-type of Incumbent has a cut-off point, and thus puts some positive probability mass on a message of zero, and has some range of messages which are not sent in equilibrium, from the Entrant's point of view, all messages may be seen with positive probability. The lowest message an Incumbent will send was defined by $m(\sigma_\gamma)$, and is positive for any positive γ. However, the limit of σ_γ as γ falls to zero is 0. Therefore for any positive value z of σ, there is a value y of γ such that $\sigma_y < z$; thus there is some measure of Incumbents (specifically $G(y)$), who will give separating messages when $\sigma = z$.

effort. However it is still true that the higher the effort, the higher the γ that must have induced it. Thus, observing the effort level is equivalent to observing γ, and allows the Entrant to infer that $\sigma < \sigma_\gamma$. This implies that the Incumbent's value of γ will have an influence on the Entrant when $m = 0$, if the Entrant can observe both m and e. In this case, for any given σ a higher Incumbent's value of γ will raise the Entrant's effort level too, since the Entrant will make an inference from a higher threshold σ_γ. Since σ and γ are independent, this relationship will hold on average across different σ_γ. These considerations provide the content of the next Proposition

> *Proposition 3: When m and e are both observable, Entrant effort will rise with γ, given σ, and Incumbents with high values of γ will be more likely to have an effect on Entrant behavior than those with low values of γ.*

Proposition 3 addresses the hierarchical influence result for the first case in Fig. 2.1. Even when all the information is available, there is still some endogenous hierarchical influence. An increase in the Incumbent's γ will exert an upward force on the level of effort exerted by less-informed incumbents for any σ they observe. This implies that across all the possible encounters (letting both γ and σ vary), there will be some correlation of players' effort levels. The Incumbent and Entrant's effort choices will not be conditionally independent.

3.3.2 Case B: *m* only observable

When e is not observable, the Entrant can no longer differentiate between the types of Incumbent who might send a given message. Furthermore, the set of different Incumbents who send a message will each do so having observed a different salience. Thus before we can address the questions of organizational or hierarchical influence, we must describe how the Entrant's interpretation of the observed message proceeds. In principle, it is relatively straightforward. Any given message m will be interpreted

as an average of the saliences that might induce an Incumbent to send m, each weighted by the likelihood of meeting just the γ_1 who would.

To begin, consider the Entrant's interpretation of $m = 0$. We saw above that all Incumbents have a threshold σ, below which they will send a message of $m = 0$. Put from the Entrant's perspective, this means a message of 0 could have been sent by any Incumbent. The assumption that $\gamma^H = k$ further implies that it could conceivably been sent following any value of σ.[80] The Incumbent's expected value of σ given $m = 0$ *and* γ_1 was defined above as $\mu(0 \mid \gamma_1)$. Thus, the Entrant's expected value of σ given only $m = 0$ will be an average over all the values of $\mu(0 \mid \gamma_1)$, weighted proportionally to the probability that the γ_1 in question sends a message of 0. It can be shown that this average is less than the unconditional expectation of σ, but – at least in the uniform distribution case, greater than half that value.

When $m > 0$, the compatible values of σ become much more restricted. First, the fact that all messages are downward distortions implies that the salience must be at least as high as the message. Second, there is also a highest σ compatible with any m, above which all incumbents either send a message higher than m, or send $m = 0$[81]. This is the next result.

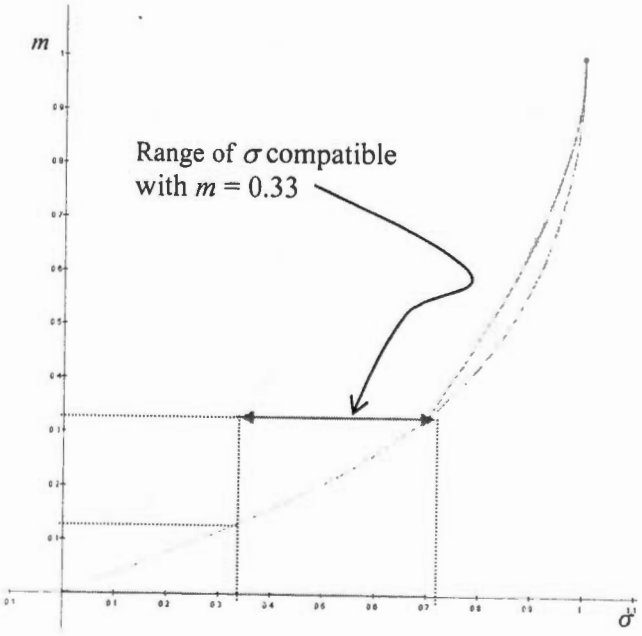
> *Lemma 2: For any message m such that $0 < m < 1$, there exists a $\sigma_{max}(m) < 1$ such that m must have been sent following a signal from the interval $[m, \sigma_{max}(m)]$.*
>
> Proof: See Appendix A

---

[80] More precisely, this requires that $\gamma^H \geq k$.

[81] Again, relaxing the assumption that $\gamma^H = k$ would make this true in only two of three cases, specifically where $\gamma^H \geq k$.

Notice that this result implies that the posterior distribution of σ given only a message m is a strict subset of [0,1] for any $m > 0$, otherwise it has full support on [0,1]. Figure 3.4 shows the locus of $\sigma_{max}(m)$ as γ rises from 0 to k. A positive message must come from a value of σ between the 45-degree line and the curve. For instance, the double-headed arrow illustrates that a message of 0.33 is compatible with σ between 0.33 and about 0.7. An illustrative strategy (upper curve) shows how the separating functions cut the locus from above.[82]



Note: The range of possible σ compatible with $m = 0.33$ is shown in the double-headed arrow. The full strategy – separating and pooling portions – for the *highest* compatible γ, in this case $\gamma_{0.33} = 0.5$, is also shown. Higher values of γ would result in curves that cut the locus at higher values of σ, and hence are not compatible with the message $m = 0.33$. Graphic adapted from an implicit plot of the original equations made with Pacific Tech Graphing Calculator software.

Figure 3.4 The locus of σ_{max}.

[82] In cases where $\gamma^H < k$, the upper limit of σ would follow the black curve up to the maximal value of γ, and then follow a separating function like the red curve beyond that.

The expectation of σ over this restricted set of possible values can be found by taking an average over the γ's compatible with it. Notice that given an m, the Incumbent for whom $\sigma_\gamma = \sigma_{\max}(m)$ is the highest-γ type who will send that message. This level of γ, denoted γ_m, can be found by eliminating σ from (3.2) and (3.10)[83]. Thus, the expected salience will be

$$E\left[\sigma \middle| m\right] \equiv \hat{\sigma}_m(m) = \frac{\int_0^{\gamma_m} \mu\left(m \middle| \gamma\right) g(\gamma)\, d\gamma}{G(\gamma_m)} \tag{3.12}$$

where

$$\gamma_m \equiv \frac{km^2}{2km - \gamma_m} = m + \frac{\gamma_m}{k}\left(1 - e^{\frac{k}{\gamma_m}(m-1)}\right) \tag{3.13}$$

and $\mu(m|\gamma)$ is as in (3.10). It can be shown that: (a) γ_m increases in m; (b) $\gamma_0 = 0$; (c) $\gamma_1 = k$.[84] It is also clear that $\hat{\sigma}_m$ will rise with m, both because the upper bound of the integral rises, and because all types who do not send a zero message send higher messages as σ rises – that is, $\mu(m)$ is an increasing function. This allows us to answer the hierarchical influence question.

> *Proposition 4: When the Entrant observes only $m > 0$, Entrant effort levels will fall with Incumbent γ for any value of σ, and Incumbents with high values of γ will be less likely to have an effect on Entrant behavior than those with low values of γ.*

---

[83] This Incumbent is also the highest who will send *any* positive message given σ. In cases below where this is the fact of interest, she will be denoted γ_σ. Notice also that her *message*, $m_{\min}(\sigma)$, the lowest positive message that will be sent following σ.

[84] If $\gamma^H < k$, then there will a case of some σ for which there is no σ_γ. In the reverse case, there would be a positive probability mass of Incumbents for whom $\sigma_\gamma = 1$.

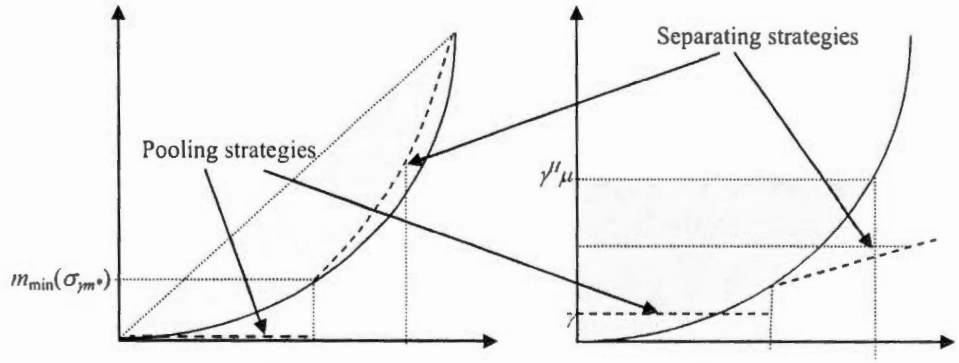
Proof: The first part is established by the facts that (a) higher values of γ reduce the message that the Incumbent sends, but increase their effort for a given σ, and (b) lower messages reduce the Entrant's interpretation which reduces their effort level. For the second part, note that the higher the value of γ, the smaller the chance *a priori* of sending a positive message. Q.E.D.

Proposition 4 is somewhat counterintuitive. In the model, Incumbents with high levels of γ "talk a lot," sending themselves relatively low positive messages for any given value of σ. However, they fail to "walk the talk" because they perfectly decode positive messages and recover their initial beliefs. The Entrants, on the other hand, have a belief that monotonically increases with the (positive) message. Thus for any given value of σ, lower messages will correspond to higher Incumbent beliefs and lower Entrant beliefs. Independence of the variables again implies that effort levels will be negatively correlated overall. One interpretation is that "naïve" Entrants are "taken in" by the Incumbents' empty talk. For instance, suppose the Incumbent tries to reduce her exposure to the rule, say, by making some comment in the lunch room about how "Byzantine bureaucratic rules are made to be bent, if not broken." Back in her office, it could be that when faced with the actual rule in question, the MCC will not let her break it. However, the Entrant who overheard her speaking goes back to his office feeling that he has actual license to do so.

3.3.3 Case C: e only observable

I turn now to the third case, in which the messages that the Incumbent sends (to herself) are private, or perhaps even subconscious. Once again, the first question to answer is the nature of the inference the Entrant can make. In this case, the Entrant must glean what information he can from observing the effort level that the Incumbent exerts, although still with full knowledge of the equilibrium she plays, and hence of the possibility that the observed effort may well be based on self-deception.

To begin, recall the general form that this effort will take. Just as with the messages, the effort of each γ_1 will be discontinuous at σ_γ. For any value of $\sigma < \sigma_\gamma$, the effort level will be constant in σ, at a level of $e = \gamma E[\sigma \mid \sigma < \sigma_\gamma]$, which is $\gamma\sigma_\gamma/2$ in the uniform-distribution case. Otherwise, the effort will rise linearly along the ray from the origin of slope γ_1. This will result in an upward-jump discontinuity in the Incumbent's effort at σ_γ, illustrated in Figure 3.5, below. The Entrant, however does not see this. From the Entrant's perspective, a given level of effort e might be due to a (particular) low-σ, high-γ Incumbent who sent herself a message of $m = 0$, or it might be any one of a continuum of high-σ, low-γ Incumbents who sent themselves the equilibrium positive messages.



(a) Message strategies (b) Effort strategies

Note: Panel (a) shows that all positive messages signify a value of σ between m and some maximum, denoted $\sigma_{\gamma m}$. By contrast, Panel (b) effort levels up to $\gamma^H\mu(0|\gamma^H)$ are compatible with *any* value of σ.

Figure 3.5 Strategies in m versus e

Figure 3.5 illustrates the phenomenon that while zero-messages are distinct from positive messages (panel (a)), the effort they engender will be the same as the effort engendered by a continuum of other (γ,σ) combinations (panel (b)). Thus any

effort level that could conceivably be send after $m = 0$ is compatible with the entire support of σ. This is the content of Lemma 3.

Lemma 3: Any effort up to $\gamma^H/2$ is compatible with any salience in [0,1].

Proof: See Appendix A.

This fact makes finding the expected value of σ given e, $\hat{\sigma}_e$, surprisingly involved. It will be equal to a weighted average of (i) the expected value of σ given that it is greater than the minimum σ that would lead any Incumbent to exert e after a positive message (i.e., the lighter part of figure 5(b), below the curve), and (ii) the expected value of σ given that it is less than the maximum value that would lead any Incumbent to exert e after a zero-message (the darker part, above the curve). Both of these values clearly increase as e rises. However, so does the weight put upon the latter, lower component. Thus the sign of the derivative $\hat{\sigma}'_e(e)$ is in general ambiguous. However, the following lemma can be proved for the current case

Lemma 4: The interpretation of σ given e increases in e.

Proof: see Appendix A.

Lemma 4 is wholly intuitive – one would expect higher observed effort levels to suggest higher levels of σ. The fact that it is not as obvious as it appears occurs because the Entrant cannot tell high-σ, low-γ effort levels apart from the inverse, in general. Moreover, a high effort level is relatively more likely to come from a high-γ Incumbent, who would exert high effort no matter σ. While one could probably construct "pathological" utility variants that would reverse the result, they would have to differ markedly from that proposed in equation (3.2), above. With this established, we can immediately adduce the following:

Proposition 5: When the Entrant observes only the Incumbent's effort level, Entrant effort will be positively correlated with the Incumbent's γ_I

Proof: When γ_I rises, so does the effort level. This raises the Entrant's interpretation of σ, and hence also his effort. Q.E.D.

This Proposition may correspond to the idea of a "culture of corruption" (for low values of γ), or at the very least an "organizational culture". In cases where an Entrant happens to see a high-γ individual, he will also tend to make higher than average interpretations of the true salience.

3.3.4 Informativeness of the cases

We can now turn briefly to an overall comparison of the information that the Entrant receives in the three cases above (diagrammed in Figure 3.1) We can immediately take it that case A is the most informative, since it provides the Entrant with informative data lacking in each of the other cases. Consider therefore the posterior distribution of beliefs in cases B and C, for a given (γ,σ) Incumbent. The criterion I will use for this comparison is the celebrated "garbling" condition of Blackwell (1953). Intuitively, this partial ordering of so-called "experiments", which are abstract procedures that generate uncertain information about the state of the world, holds experiment x to be more informative than experiment y if the distribution of outcomes from y can be modeled as equal to that of x, plus some "noise". The paper established an equivalence between this condition and the sufficiency of the statistic arising from x for the statistic arising from y.

I take the comparison of the message-only case (B) and the effort-only case (C) in two parts, distinguishing (γ,σ) types that generate positive messages from those that generate zero-messages. If $m = 0$, beliefs in case B consist of an average of all the pooling intervals, each weighted by the probability that the Incumbent to which it corresponds sees a σ low enough to generate the zero-message. As long as $\gamma^H \geq k$,

the resulting distribution places some positive weight on all positive intervals of σ in [0,1]. In Case C the beliefs combine (i) a uniformly-likely region corresponding to the unique pooling zone that would induce some specific γ (and incidentally the correct one) to exert the observed e after a zero-message, with (ii) all the values of σ that would induce some Incumbent to exert e after sending a positive message. These come from an interval that includes $\sigma = 1$, and which overlaps with the pooling interval. Therefore in Case C, too, (γ,σ) combinations leading to a message of zero result in beliefs with support on the full interval of [0,1]. From this, we can see that the two cases cannot be compared. Notice that component (i) of the beliefs in case C is one of those which are combined to form the beliefs in Case B. As far as this goes, therefore, Case B "garbles" the information in Case C, and is therefore a clear example of better information. However, the region (ii) adds noise to the Case C beliefs, relative to what is found in Case B. This is because the (γ,σ) combinations that these beliefs entertain result in positive messages, and are given zero weight in the beliefs of Case B. Thus each represents a garbling of parts of the other, and overall they remain unordered by Blackwell.

When $m > 0$, more can be said. The beliefs in Case C are the same as they were above, since m remains unobserved. However, the beliefs in Case B take the interval illustrated in Figure 3.4. Thus the support of the posterior beliefs in case B is a subset of the support in Case C. For a given distribution, this relationship would imply that the information in Case B was better. Furthermore, while the beliefs in the intersection are uniform in Case C, because no information is available about what the value of σ might be, in Case B the beliefs make use of additional information that comes from the equilibrium strategies. Therefore even in this overlap, the beliefs in Case C can be seen as a garbling of those in Case B. Overall, therefore, when $m > 0$ the beliefs in Case C correspond to a garbling of the beliefs from Case B. Because in

all cases where the two can be compared, Case B is more informative than Case C, while Case A is always the most informative, we claim the following proposition.

> *Proposition 6: {m,e} is more informative than {m}, which is more informative than {e}.*

It is rather unsurprising that the combination of the indicators would be more informative than either one alone. However, the result that messages give more precise information than effort levels seems to run counter to the conventional wisdom that "actions speak louder than words." As mentioned above, the reason for this is that in this model, the Incumbents separate more in word than in deed. Messages allow Entrants to distinguish between Incumbents who successfully self-deceive (sending $m = 0$) and those who do not ($m > 0$). This reduction in the informational pooling that the Entrant perceives permits him to make more precise estimates of σ.

3.4 Discussion and extensions

This paper introduced a framework to analyse deontological morality and self-deception in a "rationalistic" setting. Self-deception was modeled (in somewhat reduced-form fashion) as an adjustment of prior beliefs so that observed evidence was closer to desired states of the world, in which deontological rules were less relevant. In this way players informed of the rule's importance (Incumbents) were able to adjust their beliefs about it – for a cost – to relax the constraint the rule imposed. The resultant model showed endogenous distortion of beliefs, with a partial-pooling equilibrium structure. Uninformed players (Entrants) used the deceptive signals that informed players sent themselves, the actions that those informed players subsequently played, or both, to form estimates of the force of the rule. The differences between these cases led to several different kinds of effect on the Entrant's behavior that are summarized in the table below.

Table 3.1 Summary of the main results.

Case	*A*	*B*	*C*
Observable	Both e and m	Only m	Only e
Correlation between Entrant action and Incumbent type	Positive	Negative	Positive
Influence felt when	$m = 0$ (high γ)	$m > 0$ (low γ)	Always
Informativeness of the information	Best. Perfect information when $m > 0$, otherwise limited to $[0,\sigma_\gamma]$.	Second-best. Superior to C when $m > 0$; otherwise no comparable	Lowest.

There are important theoretical extensions to the model. For instance, one aspect of the phenomenon of moral rules is that people attach some social status to those who follow them. There is a kind of "righteousness" that comes with (at least the appearance of) having a strong moral sense. This could be introduced in the model as a parameter for each player that increases in the posterior expectation of γ, perhaps along the lines established in Bénabou and Tirole (2006). While the details of the extension are left to further work, qualitatively, one might easily conjecture that this would lead to higher effort levels, and potentially lower equilibrium messages.

Another important theoretical extension stems from the fact that the welfare effects of the effort have so far been left entirely obscure. Moral (or other behavioural) rules drive a potential wedge between individual welfare and patterns of choice. However, the social welfare effect of a rule (or rule following) depends entirely on the inherent effects of the behaviour in question. Many (but not all!) moral rules address public goods problems, where the behaviour that maximize the level of individual welfare is not necessarily socially optimal. In these cases, getting individuals to exert more or less effort than their individually rational level may in

fact increase social welfare. This is the basis, for instance, of Kaplow and Shavell (2007), and may be a familiar feeling about many moral strictures: they are somewhat of a burden individually, but overall, society is better for their presence. The impression that actual moral rules seem to address social welfare issues raises the interest in developing the aspect of the public nature of effort in the theoretical model. The public nature would provide a benefit to effort that would in turn generate an incentive for the Incumbent to manipulate the information that the Entrant has. This feedback on information and reward would bring the model closer to the literature on persuasion games (Grossman 1981).

Bibliography

Battaglini, M., R. Bénabou, et al. (2005). "Self-control in peer groups." *Journal of Economic Theory* **123**(2): 105-134.

Bénabou, R. (2009). Groupthink: Collective delusions in organizations and markets. *NBER Working Paper*.

Bénabou, R. and J. Tirole (2006). "Incentives and prosocial behavior." *American Economic Review* **96**(5): 1652-1678.

Bénabou, R. and J. Tirole (2010). Identity, Morals and Taboos: Beliefs as Assets. *IDEI Working Papers*, Institut d'Économie Industrielle (IDEI), Toulouse: 49 p.

Bernheim, B. D. (1994). "A theory of conformity." *Journal of Political Economy* **102**(5): 841-877.

Bikhchandani, S., D. Hirshleifer, et al. (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades." *Journal of Political Economy* **100**(5): pp. 992-1026.

Blackwell, D. (1953). "Equivalent Comparison of Experiments." *Annals of Mathematical Statistics* **24**: pp. 265-272.

Crawford, V. and J. Sobel (1982). "Strategic information transmission." *Econometrica* **50**(6): pp. 1431-1451.

Dana, J., R. A. Weber, et al. (2007). "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* **33**(1): pp. 67-80.

Deffains, B. and C. Fluet (2009). *Journal of Law, Economics and Organization*, forthcoming.

DePaul, Micheal R. (1993). *Balance and Refinement: Beyond coherence methods of moral inquiry.* Routledge: New York.

Erat, S. and U. Gneezy (2009). White lies. San Diego, Rady School of Management, UC San Diego.

Farrell, J. and M. Rabin (1996). "Cheap Talk." *Journal of Economic Perspectives* **10**(3): 103-118.

Grossman, S. J. (1981). "The informational role of warranties and private disclosure about product quality." *Journal of Law and Economics* **24**(1981): pp. 461-483.

Haidt, J. (2001). "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological Review* **108**: pp. 814-834.

Haisley, E. C. and R. A. Weber (2008). Self-serving interpretations of ambiguity in other-regarding behavior. *Department of Social and Decision Sciences, Paper 105*: 30p.

Hausman, D. M. and M. S. McPherson (1993). "Taking Ethics Seriously: Economics and Contemporary Moral Philosophy." *Journal of Economic Literature* **31**(2): 671-731.

Kant, I. (2005). *Groundwork for the Metaphysics of Morals.* Peterborough, ON, Broadview Press.

Kaplow, L. and S. Shavell (2007). "Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System." *Journal of Political Economy* **115**(3): 494-514.

Kartik, N. (2009). "Strategic Communication with Lying Costs." *Review of Economic Studies* **76**(4): 1359-1395.

Mailath, G. J. (1987). "Incentive Compatibility in Signaling Games with a Continuum of Types." *Econometrica* **55**(6): 1349-1365.

Mele, A. R. (1997). "Real self-deception." *Behavioral and Brain Sciences* **20**(1997): pp. 91-136.

Shavell, S. (2002). "Law versus morality as regulators of conduct." *American Law and Economics Review* **4**(2): pp 227-257.

Simkin, M. V. and V. P. Roychowdhury (2003). "Read before you cite!" *Complex Systems* **14**: pp. 269-274.

Van Staveren, I. (2007). "Beyond utilitarianism and deontology: Ethics in economics." *Review of Political Economy* **19**(1): pp. 21-35.

White, M. D. (2004). "Can *homo economicus* follow Kant's categorial imperative?" *The Journal of Socio-Economics* **33**(2004): pp 89-106.

APPENDICE E : SELECTED PROOFS

Proof of Lemma 1:
Two cases can be distinguished: $m = 0$ and $m > 0$. Consider first $m > 0$. In this case, Incumbents follow the separating function, which is invertible; that is, a given γ-type Incumbent will send a given message m only for one particular value of σ, although a range of (γ, σ) combinations will send each message m. Because distortion increases in γ, those who send a given message m having seen a higher value of σ will also have a higher value of γ, and so will also furnish higher levels of effort. That is, given a message, every (γ, σ) combination will exert a different level of effort, so e will be sufficient to uniquely determine the Incumbent's type after observing $m > 0$. Now consider the case where $m = 0$. In this case an observed effort level e tells the Entrant that σ was below the Incumbent's threshold, σ_γ. This threshold is monotonically increasing in γ, and so the expectation $\mu(0 \mid \gamma)$ is too. By the same reasoning as above, therefore, higher-γ individuals will also have higher expected values of σ, and thus will furnish higher levels of e: there will again be an invertible relationship between e and σ_γ, allowing the Entrant to make the exact same inference as the Incumbent. This completes the proof.

Proof of Proposition 1(d)
The Proposition holds that Incumbents have a lower threshold, and hence a longer range of separation, when γ is lower however only those with $\gamma = 0$ will never pool on $m = 0$, for any value of σ, and only those with $\gamma < k$ will ever send $m > 0$.

Proof

Introducing the assumption that σ is uniformly distributed, so $\mu(0) = \sigma_y/2$, and rearranging the threshold for pooling defined in (3.11) with this information yields

$$-\gamma\sigma_y + 2k\sigma_y m(\sigma_y) - km(\sigma_y)^2 = 0 \qquad \text{(E1)}$$

This is quadratic in (m,σ) space; it partitions that spaces into an "upper" (positive) region, inside the curve, and a "lower" (negative) region outside it. If and only if the separating function prescribes a message in the upper region, the Incumbent prefers the lower distortion cost of the positive message to the lower interpretation of the zero-message.

By inspection, the threshold decreases linearly in γ, therefore the separating function will cut it at a lower value of σ.

The expression is without real roots for positive γ. Therefore any separating function which cuts the axis (as all must do) will fall outside the curve first; all types with positive γ will send $m = 0$ for some σ.

Recall that for any γ_1, the separating function passes through the point (1,1). Therefore the claim is established if that point is in the "upper region". Furthermore, this curve opens upwards in m with a local extremum (minimum value of σ) at $\sigma = m$. For parameters γ and k, given $\sigma = m = \xi$, the value at this extremum is $\xi = \gamma/k$. Since all separating functions pass through (1,1) and cut the horizontal axis at some $\sigma_0 > 0$, is guaranteed to intersect with the separating function for any Incumbent with $\gamma/k < 1$

Q.E.D.

Corollary: When $\gamma^H > k$, there are those for whom the minimum positive message is undefined for any $\sigma \leq 1$; $m = 0$ dominates any positive message in the acceptable domain for any σ and there is a positive measure of Incumbents with $\sigma_\gamma = 1$. When $\gamma^H < k$, the maximum value that σ_γ takes is less than 1, and there is a positive measure of σ for which no Incumbent will choose $m = 0$.

Proof of Lemma 2

Messages are always less than the signal by Proposition 1, which establishes the lower bound. For the upper bound, isolating γ in and plugging into (3.9) gives

$$\sigma_{max}(m) = m + \frac{\gamma}{k}\left(1 - e^{\frac{k}{\gamma}(m-1)}\right), \text{ s.t. } \gamma = km\left(2 - \frac{m}{\sigma_{max}(m)}\right) \qquad \text{(E2)}$$

Taking the total differential of (E2) and simplifying shows that it is $\sigma_{max}(m)$ is a strictly increasing function whenever $\sigma > m$. Moreover, inspection shows that $\sigma_{max}(1) = 1$. Therefore, $\sigma_{max}(m) < 1$ for all $m < 1$. This completes the proof.

Proof of Lemma 3

The lemma can be restated: for any effort level e in $[0,\gamma^H/2]$, for any σ in $[0,1]$, if there is no γ' such that $\gamma'\sigma_{\gamma'}/2 = e$ then there is a γ^* such that $\gamma^*\sigma = e$.

Proof: Note that for any e, the γ' of the Incumbent whose zero-message results in e must be higher than the maximum γ^* of any Incumbent who would exert e after a

positive signal. Therefore, $\sigma_{\gamma'}$, the maximum σ which would lead Incumbent of type γ' to send $m = 0$ (and therefore exert e) is greater than $\sigma_{\gamma*}$, the minimum σ required to get *some* Incumbent to exert e after a positive message. Furthermore, Incumbent γ' will exert e following any $\sigma < \sigma_{\gamma'}$, and Incumbents with values of γ from γ^* down to e will exert e for the appropriate values of σ from $\sigma = \sigma_{\gamma*}$ up to $\sigma = 1$. Therefore, if the salience is higher than $\sigma_{\gamma'}$, then it must be higher than some other $\sigma_{\gamma*}$ such that $\gamma^*\sigma = e$. Conversely, if $\sigma < \sigma_{\gamma*}$, low enough that *no* Incumbents would exert e after observing σ and sending a positive message, then, as long as $e < \gamma^H/2$, there must be a $\sigma_{\gamma'} > \sigma$: some Incumbent who would exert e following $m = 0$. Q.E.D.

Proof of Lemma 4

The proof will proceed by comparison with a benchmark case. This problem is a special case of a general class in which a value of interest is the average of two quantities, but where not only the quantities themselves but also the relative weighting depends on some other variable. The general format is therefore

$$V(x) = p(x)F(x) + (1 - p(x))G(x) \qquad \text{(E3)}$$

where $p(.)$, F(.) and $G(.)$ are all continuous functions. In this particular case, all the derivatives are positive, and $F(x) < G(x)$ for all x.

The derivative of (E3) can be written as

$$V'(x) = p'(x)[F(x) - G(x)] + p(x)F'(x) + (1 - p(x))G'(x) \qquad \text{(E4)}$$

We can divide this overall change into two components. Denote the first term the *weighting effect*. Shifting the weight placed on the values will exert an influence on the overall value proportional to the difference between them. The remainder of (E4) shows an *average change effect*. This reflects the intuition that if each component of a total rises, that will exert an upward influence on the sum. In the current case, the overall expected value will rise with x if and only if the average increase in F and G is greater than the increase in the weight assigned to the lower value. Thus the weighting effect is analogous to a marginal cost, while the average change effect is analogous to a marginal benefit. The value increases with x if the marginal benefit is greater than the marginal cost.

Now consider the simple benchmark example with $p = x$, $F = x/2$, and $G = (1 + x)/2$ for x in $[0,1]$. This example partitions the unit interval at x, then takes the expected value of the lower and upper regions, each weighted by its respective length. The law of total probability obviously implies here that the average will not change with x. Applying (E4) implies

$$1 \cdot \left[\frac{x}{2} - \frac{(1+x)}{2} \right] + x \frac{1}{2} + (1-x) \frac{1}{2} = 0 \qquad \text{(E5)}$$

In the case of the model presented here, the expression can be re-written:

$$\hat{\sigma}_e(e) = p(e)\frac{\sigma(2e)}{2} + (1 - p(e))\frac{-\sigma(e)\ln\sigma(e)}{(1-\sigma(e))} \tag{E6}$$

where

$$p(e) = \frac{\sigma(2e)}{\sigma(2e) + e\dfrac{1-\sigma(e)}{\sigma(e)}} \tag{E7}$$

Writing this in the format of (E4) yields

$$\begin{aligned}\hat{\sigma}_e'(e) = p'(e)\left[\frac{\sigma(2e)}{2} + \frac{\sigma(e)\ln\sigma(e)}{1-\sigma(e)}\right] \\ + p(e)\sigma'(2e) + (1-p(e))\left[-\sigma'(e)\frac{1-\sigma(e)+\ln\sigma(e)}{(1-\sigma(e))^2}\right]\end{aligned} \tag{E8}$$

with the weighting effect on the first line and the average change effect on the second. The goal is to interpret (E8) in terms of $\sigma(e)$. Define $\varphi(\sigma(e)) \equiv \sigma(2e)$, which can be shown to be an increasing, convex function defined over $0 < \sigma < k/2$, and $e(\sigma)$ is the inverse of $\sigma(e)$, an increasing, concave function. Thus (E7) becomes

$$p(\sigma) = \frac{\varphi(\sigma)}{\varphi(\sigma) + e(\sigma)\dfrac{1-\sigma}{\sigma}} \tag{E9}$$

The proof of the Lemma will be based on two claims. First (Claim a), I show that the weighting effect in (E5) is greater than that in (E8). In other words, the weighting effect for the current model is less than 0.5. Second (Claim b), I show that the average change effect in the benchmark is smaller than in the current model – i.e., that the average change effect under examination is greater than 0.5. These claims combine to show that the total must be greater than (E5), which is the desired result.

Claim a: The weighting effect is less than 0.5

Again, I take this in two steps. First, notice that for any σ in [0,1], $\varphi(\sigma)/2 > \sigma/2$, while $\sigma \cdot \ln(\sigma)/(1-\sigma) < (1+\sigma)/2$. Thus $[F-G] > -0.5$. This implies that the claim is established if $p' < 1$.

Suppressing arguments for notational clarity, the derivative of (E9) can be written

$$p' = \frac{\frac{1-\sigma}{\sigma}(\varphi' e - \varphi e') - e\frac{1}{\sigma^2}}{\left(\varphi + e\frac{1-\sigma}{\sigma}\right)^2} \qquad \text{(E10)}$$

This is less than unity if

$$\frac{1-\sigma}{\sigma}(\varphi' e - \varphi e') - \frac{e}{\sigma^2} < \varphi^2 + 2e\frac{1-\sigma}{\sigma} + e^2\frac{(1-\sigma)^2}{\sigma^2}$$

$$0 < \varphi\left[\varphi + e'\frac{1-\sigma}{\sigma}\right] + e\frac{1-\sigma}{\sigma}[2-\varphi'] + \frac{e}{\sigma^2}\left[e(1-\sigma)^2 + 1\right] \qquad \text{(E11)}$$

This is guaranteed if $\varphi' < 2$. Notice that

$$\varphi(\sigma) = \sigma(2e) = \sigma + e\sigma'$$

$$= \sigma + \frac{e}{e'} \qquad \text{(E12)}$$

by a first-order Taylor expansion, where the change in the derivative is due to the fact that the derivative of an inverse of a function is the reciprocal of the derivative of the function. This implies that

$$\varphi' = 1 + \frac{(e')^2 - e \cdot e''}{(e')^2}$$

$$= 2 - \frac{e \cdot e''}{(e')^2} < 2 \qquad \text{(E13)}$$

due to the concavity of $e(\sigma)$. This establishes the first claim.

Claim b: The average change effect is greater than 0.5.

For any p, this will be established if both F' and G' are greater than 0.5, since the average change effect is a weighting of those values. Expression (E8) reveals that F' is shown by (E13). Thus it is established as long as

$$\frac{e \cdot e''}{(e')^2} \equiv \varpi < 1.5 \qquad \text{(E14)}$$

Thus it is satisfied if the function is not "too curved". For instance, exponential curves have $\varpi = 1$, and curves of the form e^n have $\varpi = (n-1)/n < 1$. Functions of the form

$$e = \frac{1}{n}\sum_{k=1}^{n} \sigma^k, \qquad \text{(E15)}$$

which maintain the fixed points at zero and 1, violate (E14) for any σ only when n grows beyond 22. By contrast, the best fit with the function as defined in the text has $n = 6$, which generates $\varpi < 1$ for all σ; the curve also looks substantially like e, as noted below, which generates a constant $\varpi = 0.63$. For illustrative purposes, the figure below shows computational results for the different formulations. The dotted curves are plotted from expression (E15) with $n = 6$, and $e^{2.70}$, respectively; the solid line is plots the implicit relationship. The dashed curve, which plots (E15) with $n = 22$, illustrates the level of curvature required to violate (E14). This makes clear that, at least for functional forms "close to" that used here in terms of the behavioral predictions they generate, the conclusion that $F' > 0.5$ is without much risk.

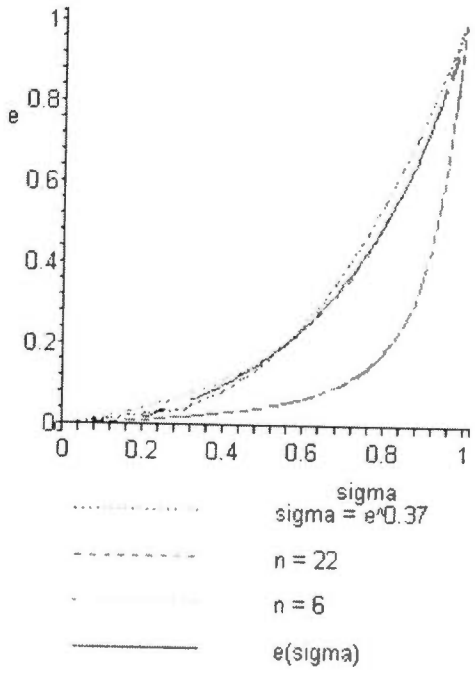


Figure E.3 Computed approximations and the threshold violating function.

Turning at last to G', the derivative of the relevant (last) term in (E8) yields

$$G' = -\frac{(1-\sigma)+\ln\sigma}{(1-\sigma)^2} \quad \text{(E16)}$$

This is a decreasing function, with a limit of 0.5 when $\sigma = 1$, and therefore greater than 0.5 for all $\sigma < 1$. This establishes the second claim, and thereby concludes the proof.

Q.E.D.