Communication and Deception in 2-Player Games*

Benjamin Bachi† Sambuddha Ghosh‡ Zvika Neeman§

October, 2016

Abstract

Real communication among people is often not 'cheap': those engaged in it may unintentionally betray their true intentions, or guess the true intentions of others, through involuntary bodily gestures. This implies that players' strategies should be described by response functions mapping gestures of the other players into actions in the game, rather than by mere actions, as in the standard formulation. This has a profound effect on the way games are played. In particular, it can account for the significant levels of cooperation and correlation observed in experimental Prisoner's Dilemma games with non-binding pre-play communication. Quite significantly, our explanation differs from existing explanations in keeping the payoffs to action profiles unchanged, and instead reexamining the notion of a strategy.

KEYWORDS: deception, communication, cheap talk, Prisoners' Dilemma.
JEL CLASSIFICATION CODES: C72, D83.

*Acknowledgements to be added.
†Max Planck Institute for Research on Collective Goods, Email: bbachi@coll.mpg.de
‡Shanghai University of Finance and Economics, Email: sghosh@mail.shufe.edu.cn
§Tel Aviv University; Email: zvika@post.tau.ac.il; URL: http://tau.ac.il/~zvika/.

1 Introduction

Dominance solvability is considered the strongest solution concept in game theory. Yet, it seems at odds with a large body of experimental evidence on the ubiquitous Prisoner's Dilemma (henceforth PD) with pre-play communication. A meta-analysis of 37 experiments from 1958 to 1992 in Sally (1995) revealed that non-binding pre-play communication increases the rate of cooperation by roughly 40%, although theory predicts cheap talk makes no difference. Second, perhaps more interestingly, Frank (1988) reports experiments showing that when subjects are allowed to interact for 30 minutes before playing the PD, they are able to predict quite accurately their opponent's action. Moreover, roughly 84% of the subjects who predict that their opponent will cooperate (defect) respond with defection, which makes play correlated rather than independent.[1] Longer communication leads to a higher probability of cooperation; in the experiment, both the level of cooperation and the accuracy of the predictions drop when players are allowed to interact for only 10 minutes.

In this paper we present a new explanation for the pattern of cooperation in the PD that hinges on a new model of communication between the players. We believe that unlike "cheap talk," *real communication* is distinguished by the fact that those engaged in it may unintentionally betray their true intentions, or learn the true intentions of others.[2] If players in a game can possibly guess other players' intentions by observing their bodily or verbal gestures as argued above, then they may obviously want to condition their play on the bodily or verbal gestures they observe. This implies that players' pure strategies should include mappings or response functions from bodily and verbal gestures of the other players into actions in the game, rather than be given by mere actions, as in the standard formulation of normal form games.

This guessing of intentions may have a profound effect on the way games are played. To see this, consider the following PD, where each player's action set is $\{C, D\}$ and payoffs are as follows:

	C	D
C	$3,3$	$0,4$
D	$4,0$	$1,1$

Take as given a set $\{(\hat{C}, \text{☺}), (\hat{D}, \text{☹}), (nice, \text{ᓚᘏᗢ})\}$ comprising pairs of response functions and their corresponding bodily gestures, with the following interpretation. The constant response $\hat{C}$ implies that the player always plays cooperate (C), regardless of what she learns about the other player;

[1] Additional experimental evidence is provided below, in the section on related literature.

[2] For example, a recent paper by DeSteno et al. (2012) reports that subjects in an experiment were able to successfully predict the untrustworthiness of other subjects with whom they communicated by four visual cues – leaning away; crossing arms in a "blocking fashion"; touching, rubbing, or grasping hands together; and touching oneself on the face, abdomen or elsewhere. Interestingly, the cues were found to be predictive only in combination. Another interesting observation is that identification of someone as less trustworthy may be unconscious in the sense that subjects may not be aware that it is due to the aforementioned visual cues.

Another experimental paper by Manzini et al. (2009) shows that players use and recognize simple cues such as "smiles, winks and handshakes [...] as a tell-tale sign of each other's trustworthiness."

a player who chooses $\hat{C}$ *involuntarily* smiles (☺). A player frowns (☹) if she chooses the constant response $\hat{D}$ that implies that she always plays defect (D), regardless of what she learns about the other player. The mysterious facial expression ☋ corresponds to the response *nice*, which reveals that the player will cooperate (C) if she observes the gesture ☋ (which indicates that the other player is also playing the response *nice*), and defect (D) otherwise.

The timing of the game is as follows:

1. Players decide simultaneously how they plan to play the game by choosing a response from the set $\{\hat{C}, \hat{D}, nice\}$.

2. Players involuntarily display the bodily gesture that is associated with their response function (plan).

3. Players play an action taking into account the bodily gesture exhibited by the other player and following the response function picked earlier.

The following matrix shows the actions played by the players as a functions of the responses they choose:

	$\hat{C}$	$\hat{D}$	*nice*
$\hat{C}$	C, C	C, D	C, D
$\hat{D}$	D, C	D, D	D, D
nice	D, C	D, D	C, C

These actions generate the following payoff matrix:

	$\hat{C}$	$\hat{D}$	*nice*
$\hat{C}$	3,3	0,4	0,4
$\hat{D}$	4,0	1,1	1,1
nice	4,0	1,1	3,3

Two response profiles give the best payoff — $(\hat{C}, \hat{C})$ and $(nice, nice)$. Only the second is an equilibrium of the game. In fact the conclusion is stronger: *nice* weakly dominates both $\hat{C}$ and $\hat{D}$, even though $\hat{D}$ is a dominant strategy in the 2×2 game that includes only the responses $\hat{C}$ and $\hat{D}$, it is not weakly dominant in the 3×3 game with communication although $(\hat{D}, \hat{D})$ is still a Nash equilibrium. As we shall explain at greater length, our explanation is novel because *it leaves payoffs to action profiles unchanged and reinterprets the notion of a strategy when explicit communication is available.*

The example shows that, by making deviations easy to detect, the involuntary revelation of intentions facilitates commitment, which expands the range of equilibrium outcomes. But players who understand that other players believe they can guess their intentions may try to actively deceive them. We now enrich this model to allow deception, which has the opposite effect of the involuntary revelation described above: it shrinks the set of equilibrium outcomes.

Deception matters because it is the "dual" of trust, which seems to be an important determinant of important macroeconomic variables such as GDP growth, inflation, and the volume of trade between countries.[3] Ample evidence of deception in the field and in the laboratory exists; see Croson (2005) for a short but careful summary of the literature.

What are some reasonable properties of deception? First, people seem to have a moral compunction about deception.[4] Second, the ability to deceive varies across persons. Finally, people also vary in their ability to detect lying (Ekman et al., 1999; Kartik et al. 2007). Our model combines the first and second features above by assuming that deception is costly, and that people may differ in their costs of deception. These costs may come from moral compunctions, or the extensive and self discipline required to deceive.

Deception is difficult to capture in a standard game theoretic framework because the fact that in equilibrium players best-respond to each other's (true) strategy implies that players cannot be fooled in equilibrium. We therefore define equilibrium deception behaviorally as transmiting a false signal as explained below. Fix a 2-player normal form game G, henceforth called the underlying game. Each response of each player is associated with a specific "natural" bodily or verbal gesture of the player that would be observed by the other players. A deceptive player can associate any bodily gesture with any response, but at a cost. Each player has a given list of responses R_i, a given list of bodily gestures $\mathcal{G}_i$, and a function $b_i : R_i \to \mathcal{G}_i$ mapping responses to the corresponding bodily gestures naturally associated with them. Each player chooses a response $r_i \in R_i$, a bodily gesture $g_i \in \mathcal{G}_i$, and a probability of deception p_i. The triple $\langle r_i, g_i, p_i \rangle$ is a strategy of i in the augmented game obtained by adding pre-play communication to the underlying game G. Player j observes player i's gesture g_i with the probability p_i and observes player i's true gesture $b_i(r_i)$ with the complementary probability $1 - p_i$. Player j cannot tell if player i's gesture is natural or deceptive. Player i is *deceptive* if she chooses a bodily or verbal gesture $g_i \neq b_i(r_i)$ and a positive probability of deception $p_i > 0$; otherwise, player i is *truthful*. Our formulation allows player i to completely deceive player j with any probability p_i, and make player j believe whatever player i wants j to believe about i's intentions at a linear cost $c_i p_i \geq 0$, arising from moral compunctions or the exertion required to control one's gestures.

Taking the costs of deception c_1 and c_2 to be commonly known, we ask how the equilibrium payoff set in the augmented game differs from that in the underlying game. We show that deception is possible when the costs associated with it are low. However, such deception does not expand the set of equilibrium payoffs: if *both* players engage in deception, they must play a Nash equilibrium of the underlying game without communication. Furthermore, if at least one of the players engages in deception, then there is another equilibrium which has no deception and induces the same actions in the underlying normal-form game. In summary, if the costs of deception

[3] LaPorta et al. (1997), Knack and Keefer (1997), Zak and Knack (2001), Guiso et al. (2009) document various correlations in support of this.

[4] Gneezy (2005) finds that experimental subjects are reluctant to tell lies that bring only small benefits to themselves, or do great harm to others. Croson (2005) observes that people are willing to punish deception even it is costly to do so. See also Fehr (2009) on why trust is not just risk-taking but rooted in commonly understood aversion to betrayal.

are sufficiently small, then the set of equilibrium outcomes of the augmented game coincides with that of the underlying game with only cheap-talk communication. Any expansion in the payoff set comes from the possibility of commitment when deception costs are high enough that there is no equilibrium deception. We prove a "folk theorem" that characterizes the set of equilibrium payoffs: if deception is sufficiently costly, all feasible and individually rational payoffs may be attained in equilibrium. The set of equilibrium payoffs shrinks as the costs of deception decrease.

Thus, we show that, perhaps contrary to intuition, a *commonly known* low cost of deception or being known to be a good liar actually puts a player at a disadvantage because it undermines the ability to commit. Lying in our model can be profitable only if a player is not believed to be a good liar. Indeed, successful equilibrium deception in which one player gains at the expense of another player can exist only if (i) the costs of deception are privately known, and (ii) the ex-ante probability that the deceitful player's cost of deception is small is itself sufficiently small. In this case, we show that there exist equilibria in which the types of the player whose cost of deception is small successfully deceive the types of the other player whose cost of deception is large, even though the latter still play best responses. The fact that all players' types play best response implies that the ex-ante probability of deceitful players must be small; otherwise, other players' types cannot be successfully fooled. The equilibria we describe are consistent with the experimental data described above.

The paper proceeds as follows. Section 2 consists of a review of related literature. In Section 3 we present the model and basic definitions. In section 4 we find what payoffs may be sustained in equilibrium of general two player games with commonly known costs of deception. In Section 5 we consider the case of privately known costs of deception.

2 Related Literature

Evidence from the laboratory is documented in Frank (1988) and Sally (1995) mentioned earlier. Recent papers (Kalay et al. 2003, Belot et al. 2010, and den Assem et al. 2010) study data from game shows where a PD game (with defection as weakly dominant) is played after non-binding communication. All papers find significant cooperation and also correlation between the players' actions.[5]

Communication in our paper is very different from cheap talk, introduced in Crawford and Sobel (1982) and described in Farrell and Rabin (1996) as "costless, non-binding, non-verifiable messages that may affect the listener's beliefs." In our model the player can convey a false impression, but at a cost. As we show below, making deception costly leads, in some games, to outcomes that cannot be generated by either cheap talk equilibria or Aumann's (1974) correlated

[5]The first paper finds that each player cooperated 42% of the time; both cooperated 21% of the time compared to 17.64% had there been no correlation, implying a correlation coefficient of 0.14. Yaari (2011) makes similar observations to these papers. See also Belot et al. (2012) for a survey of related research by economists and psychologists. The curious reader can search Youtube with the keywords 'split or steal' or 'share or shaft'; the results are both entertaining and instructive.

equilibrium. The best example is the Prisoners' Dilemma; our model permits players to achieve *full* cooperation.[6]

Other explanations for the high degree of cooperation exist. However we argue below that the explanations closest in spirit are not formulated in a game-theoretic framework, whereas other explanations are substantively different. We group together as "behavioral explanations" a number of models that argue that the actual payoffs of the game tested were very different from the payoff matrix of the standard PD presented above. Some prominent behavioral explanations are Fehr and Schmidt (1999), based on fairness; Battigalli and Dufwenberg (2007) and Miettinen (2013), on guilt; Levine (1998), on altruism; and López-Pérez (2012), on a preference for honesty.[7] Our work leaves the payoffs to action profiles unchanged, making it clear that the observed data are consistent with the PD payoffs and are not on account of the game having a different payoff matrix. Several other differences stand out. These models leave room for explaining the observed correlation between actions.[8] These do not model communication, or answer why the rate of cooperation increases with the length of communication. In our model longer communication increases 'transparency' and this serves to improve cooperation and generate correlation. Other explanations for cooperation, such as reciprocity (Axelrod, 1987), have no bite in non-repeated, one-shot games such as ours.

What is crucial for our approach is this element of one's strategies depending on others' strategies, and the delicate part of the model involves introducing this dependence into a simultaneous-move game. Specifically, suppose that players dislike lying. Is it possible for both players to promise to cooperate and then cooperate in equilibrium ? The answer depends on how a player responds to a player who did not promise to cooperate. If a player hates to lie, and therefore cooperates after promising to do so regardless of what the other player promised, then the other player would do better to not make any promises, and defect. So what is important is not so much whether players hate to lie or disappoint as such, but how they respond to a player who did not indicate a willingness to cooperate.

Therefore, it is possible to support a cooperative equilibrium with players who dislike lying only if this dislike is conditional. Namely, players dislike lying only if the other player also dislikes lying and promised to cooperate. If the other player did not make any such promise, then a player is only too happy to revoke its own promise to cooperate. A similar argument applies if instead of a preference for honesty, players suffer from guilt if they impose a lower payoff on the other player, or some other behavioral assumption. It is also possible to assume that players incur a cost not only for breaking their promises, but also for not making them. This turns the game into

[6] But see Forgó (2010) for a generalization of correlated equilibrium that admits some cooperation in PD-like games.

[7] Apart from fundamental conceptual differences from ours, some comparable behavioral explanations may be distinguished from ours through experiments; for example, sufficient guilt could lead to the play of (C, D) in the PD, although this could never arise in our model because our formulation respects individual rationality.

[8] Cheap talk can induce correlation in coordination games. It is possible to use these behavioral explanations to turn the PD game into a coordination game and then use cheap talk to achieve correlation; however other points of difference would remain.

a psychological game (Geanakoplos et al., 1989).[9] We differ from the standard model in that our signals are not about exogenous and intrinsic types but about intentions and choices.

Frank (1988) describes an informal model of commitment without any external mechanisms, where the players' emotions serve as commitment devices. Since psychological research shows that emotions are both observable and hard to fake (see Frank (1988) and references within), an agent can use them as signals. Gauthier (1986) describes an environment with two types of agents: straightforward maximizers and constrained maximizers. A straightforward maximizers simply maximizes own utility, whereas a constrained maximizer "is conditionally disposed to cooperate in ways that, followed by all, would yield nearly optimal and fair outcomes, and does cooperate in such ways when she may actually expect to benefit." An agent's type is known to everybody (or at least with positive probability). Thus, in the Prisoners' Dilemma, when a constrained maximizer meets another constrained maximizer, they will both cooperate; in any other interaction between the two players, they will both defect. These last two works resemble ours but are not posed in a formal game-theoretic framework; Binmore (1994), for example, criticized Gauthier for logical inconsistency. Neither considers the possibility of deception.

Amann and Yang (1998) extend Frank (1988) to study replicator dynamics in a Prisoner's Dilemma where players must invest a small cost of sophistication and learn their partner's trait. Here are the main differences. Unlike our work the definition of trait is rooted in the specific stage-game. Active deception is not permitted. Non-maximizing behavior, such as being trustworthy, is shown to be evolutionary stable if propensities are even partially observable.[10]

A small game-theoretic literature establishes a folk theorem when players' strategies may depend on the other players' strategies. We discuss the connection between the results of this literature and those obtained here after we present our own folk theorem (Proposition 4) in Section 4 below. We should like to mention Matsui (1989), which shows how leakage of the dynamic game strategy can force cooperation in the prisoners' dilemma.

Finally, a small literature attempts to model deception as an equilibrium phenomenon. As explained above, deception is a tricky idea to model in standard game theory because the fact that in

[9]The resemblance of our work to this is one of appearance only: in our model utilities depend, as in conventional games, only on the action profile. The 'psychological' aspect of our work pertains to how well players may hide their meta-strategies from the others, but not to how actions translate into payoffs. In contrast, in psychological games a player's utility may depend directly on his hierarchy of beliefs – for example, a player wants to take an action to 'surprise' the other player.

[10]See Guth and Yaari (1992) for an evolutionary justification for reciprocity, and Dekel et al. (2007) for a theory of endogenous preferences. Robson (1990) employs a "secret handshake" argument to argue that certain evolutionary stable strategies (ESS) may actually be destroyed by a mutant strategy that transmits a signal that only other mutants can identify, thereby allowing mutants to play differently against other mutants and those playing the ESS. Wiseman and Yilankaya, (2001) point out that this may give rise to cycles. Costs and deception do not enter Robson and its generalizations, just as evolutionary arguments are absent in ours. Also, in our setting the handshake is public, not secret. Interestingly, this literature gave rise to the so called "truth-telling hypothesis" (Frank, 1988; Ockenfels and Selten, 2000) that asserts that opportunists inadvertently look and behave differently from trustworthy people, despite their attempts at deception. Demichelis and Weibull (2008) extend the cheap-talk approach to pre-play communication by way of introducing a meaning correspondence between messages and actions, and by assuming that players have a lexicographic preference, second to material payoffs, against deviating from the meaning correspondence and engaging in deception. They show that in symmetric coordination games, a Nash equilibrium is evolutionarily stable if and only if it results in the unique Pareto efficient outcome of the underlying game.

equilibrium players best-respond to each other's (true) strategy implies that *successful deception* is essentially precluded by the notion of Nash equilibrium. As in this paper, the focus of Crawford (2003) is on active misrepresentation rather than less-than-full disclosure, and on signaling intentions rather than private information. Motivated by the allied invasion of Normandy in World War II, he considers a 2×2 sender-receiver game in which the sender has several different types: a truthful type whose action is identical to its signal, several "wily" types whose actions and signals may differ, and a "sophisticated" type that plays optimally given its beliefs. Importantly, deception cannot be sustained without the truthful and wily types, and so hinges on bounded rationality.[11] All our types play best responses.

Ettinger and Jehiel (2010) also rely on bounded rationality to model deception. Their motivating example is that of a seller of a house who reveals an undesirable fact about his property to induce the prospective buyer to believe that the house does not suffer from a much more serious defect. Agents decide based on the simplest theories compatible with the available knowledge. This predisposes their agents to the so-called Fundamental Attribution Error and allows them to be deceived. Our model does not rely on this psychological bias. Anderson and Smith (2013) studies dynamic deception.[12]

Kartik, Ottaviani and Squintanni (2007) and Kartik (2009) modify the standard sender-receiver game of Crawford and Sobel (1982) to include either costs of lying or a proportion of naive receivers who blindly follow the sender's recommendation. In these two papers the sender lies about an exogenous state; in contrast, deception in our paper is about masking one's intended play, which is endogenous. Very importantly, the models are very different because our "response function" does not have a counterpart in the above two papers. We do not restrict attention to cheap-talk games. We also show below that if there is an equilibrium in which one player deceives, then there is an alternative outcome-equivalent equilibrium without deception. This is very different from both the cheap-talk papers mentioned above.

3 The Model

Let $G = \langle N = \{1,2\}, (A_1, A_2), (\pi_1, \pi_2) \rangle$ be a two-person game in normal form, where A_i is the (finite) set of pure actions for player $i \in \{1,2\}$, and $\pi_i : A_1 \times A_2 \to \mathbb{R}$ is the payoff function for player i. The set of (mixed) action profiles of G is $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, where $\mathcal{A}_i := \triangle A_i$. A Nash equilibrium of G is a pair of mixed actions $\alpha^* \equiv (\alpha_1^*, \alpha_2^*) \in \mathcal{A}$ such that for each $i, j \in \{1,2\}$ with

[11] Hendrick and McAfee (2006) is also motivated by the Normandy example. One player attacks another and must decide how to divide the available resources between deception and strength of the attack. When signals are not very revealing about intentions, the lion's share of resources is devoted to attack. We do not have a resource allocation problem between obfuscation and attack, and our game is more general. See also Ellingsen and Östling (2010) for a more general model in which players hold level-k beliefs about other players.

[12] An informed long-lived player, whose actions are obscured from a sequence of uninformed rivals by exogenous Gaussian noise, derives utility from matching his rival's action and the state of the world. The essence is that the long-lived player faces a dynamic trade off between his immediate payoffs and the capitalized future value of his informational edge. Such a trade off plays no role in our static model; in addition, deception is endogenous in our model, while the noise in the other model is exogenous.

$i \neq j$ we have $\pi_i(\alpha_i^*, \alpha_j^*) \geq \pi_i(\alpha_i, \alpha_j^*)$ for any $\alpha_i \in \mathcal{A}_i$. For any $\alpha_i \in \mathcal{A}_i$, let $\alpha_i(a_i)$ be the probability of the action a_i.

We define an augmented game $G(\mathcal{I}, c)$ induced by G by an "interaction structure" $(\mathcal{I}, c)$. The augmented game can be thought of as the game G with pre-play communication, or the "game with real talk" induced from G. In this paper, we interpret "interaction structure" as a process that maps one's intention to play into a bodily or verbal gesture. For each i we fix a set $\mathcal{G}_i$ of bodily and verbal gestures that are perceptible by the other player, a set $R_i \subseteq \{f : \mathcal{G} \to \mathcal{A}_i\}$ of response functions from the players' bodily gestures $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$ into one's own actions $\mathcal{A}_i$,[13] a function $b_i : R_i \to \mathcal{G}_i$ that maps player i's response functions into the bodily and verbal gestures naturally associated with them, and a cost c_i of deceiving the other player about one's intentions.

A strategy of player i in the augmented game $G(\mathcal{I}; c)$ induced from G is a triple $(p_i, r_i, g_i) \in [0, 1] \times R_i \times \mathcal{G}_i$ comprising

1. a probability $p_i \in [0, 1]$;
2. a response function $r_i \in R_i$;
3. a bodily gesture $g_i \in \mathcal{G}_i$.

The probability p_i that player i devotes to deceiving player j is the level of (costly) deception. The response function $r_i \in R_i$ describes how player i intends to play the augmented game $G(\mathcal{I}; c)$; and the bodily gesture $g_i \in \mathcal{G}_i$ describes what player i wants player j to think about player i's intentions.

Definition 1. *Fix a normal-form game G and an interaction structure with known costs*

$$(\mathcal{I}, c) = ((R_1, R_2), (\mathcal{G}_1, \mathcal{G}_2), (b_1, b_2), (c_1, c_2)),$$

where $\mathcal{G}_i$ is a finite or infinite set, R_i is a subset of $\{f : \mathcal{G} \to \mathcal{A}_i\}$, $b_i : R_i \to \mathcal{G}_i$, and $c_i \geq 0$. The augmented game $G(\mathcal{I}; c)$ induced from G by the interaction structure $(\mathcal{I}, c)$ is played as follows:

1. *Two players $i = 1, 2$ choose their strategies $(p_i, r_i, g_i) \in [0, 1] \times R_i \times \mathcal{G}_i$ simultaneously.*
2. *Each player i observes a signal $\sigma_j \in \mathcal{G}_j$, where $j \neq i$, that equals the bodily gesture g_j with probability p_j and the bodily gesture $b_j(r_j)$ with probability $1 - p_j$.*
3. *Each i plays in the underlying game G according to her response function r_i depending on the signal from $\mathcal{G}_j$ she observed and the signal the other player observed about her.*
4. *Player i earns a payoff $\pi_i(r_1(\sigma_1, \sigma_2), r_2(\sigma_1, \sigma_2)) - c_i p_i$, where $c_i \geq 0$ denotes i's cost of deception.*

[13] For most of the paper, player i's response would anyway be constant over i's own gestures and would depend only on j's gestures. For this reason, we suppress the notation of $\mathcal{G}_i$ below, and describe i's response function as $r_i : \mathcal{G}_j \to \mathcal{A}_i$. We re-invoke the more general form when we explicitly refer to it in the proofs of Propositions 4 and 6 below.

Remark 1. Among others, the sets of response functions R_i are given exogenously as part of the description of the augmented game. If R_1 and R_2 include all and only pure actions of the underlying game G as constant responses of the augmented game $G(\mathcal{I}; c)$ and there is a one-to-one mapping from R_i to $\mathcal{G}_i$, then the augmented game coincides with the underlying game G if there is no possibility of deception. The sets R_1 and R_2 and the corresponding $\mathcal{G}_i$'s can also be either smaller or larger. For example, R_i and its corresponding $\mathcal{G}_i$ can be a singleton set containing just one response function for each player, i.e. $R_i = \{r_i\}$ where $r_i(g_j) = a_i \in A_i$ for all $g_j \in \mathcal{G}_j$. At the other extreme, and perhaps also more naturally, R_i and its corresponding $\mathcal{G}_i$ may include all responses from G_j to $\mathcal{A}_i$ that can be described in finite sentences (using Gödel encoding, see Peters and Szentes (2012) for details);[14] we henceforth assume that the set of gestures and mappings from gestures to actions are both rich in the sense that it includes all the response functions that are necessary to establish our results, unless otherwise stated. While the entire force of this assumption is not needed for our results, it saves us the tedium of clarifying exactly which response functions are included in the information structure $\mathcal{I}$.

For example, in the PD game that is described in the introduction, the sets R_i and $\mathcal{G}_i$ include three elements each but in principle, these sets could be much larger (or smaller) and include any response function that can be described by a finite sentence.

Remark 2. Meaningful communication requires that the set of bodily and verbal gestures $(\mathcal{G}_1, \mathcal{G}_2)$ be sufficiently rich and the functions (b_1, b_2) not be constant functions. As will become clear below, we do not impose any restrictions on these sets and functions beyond the following minimal one: to support, say, the responses r_1 and r_2 in equilibrium the sets $\mathcal{G}_1$ and $\mathcal{G}_2$ must each contain at least one gesture that is naturally associated only with r_1 and r_2 respectively, and at least one other gesture that would be naturally associated through the functions b_1 and b_2, respectively, with all other responses.

Remark 3. Note that in the definition above player i's payoff $\pi_i(r_1(\sigma_1, \sigma_2), r_2(\sigma_1, \sigma_2)) - c_i p_i$ depends on both the signal from his opponent and the signal he transmits. It is usually enough to let r_i depend only on σ_j. The general formulation will be used only to generate correlated actions. We can think of the bodily gestures as including verbal messages that have no intrinsic meaning but

[14] An important observation is that we cannot let $R_i = \left\{f : \mathcal{G}_j \longrightarrow \mathcal{A}_i\right\}$ with a one-to-one mapping b_i from R_i to $\mathcal{G}_i$ because this set would be too big. For example, suppose that each player has two pure actions in the underlying game. If player j has two different response functions in R_j and two gestures that identify these response functions, then $R_i = \left\{f : \mathcal{G}_j \longrightarrow \mathcal{A}_i\right\}$ implies that player i needs to have $2^2 = 4$ different (pure) response functions in R_i and four gestures that identify these response functions, which in turn implies that player j needs to have $2^4 = 16$ different (pure) response functions in R_i and sixteen gestures that identify these response functions. A contradiction. This means that the sets R_i, R_j and $\mathcal{G}_i, \mathcal{G}_j$ need to be specified in such a way that is mutually consistent.

A common set of response function that is used often in the field of mathematical logic where a similar issue of how to encode sentences (in our case strategies) arises is the set of finite sentences, which is countable. The set of response functions that are defined on a countable domain is uncountable, and so the restriction to finite sentences implies that this construction imposes a restriction on the set of response functions, which implies that $R_i \subsetneqq \left\{f : \mathcal{G}_j \longrightarrow \mathcal{A}_i\right\}$ and $R_j \subsetneqq \left\{f : \mathcal{G}_i \longrightarrow \mathcal{A}_j\right\}$.

serve to produce a jointly controlled lottery over action profiles; this is why we need r_i to depend on both σ_i and σ_j. Other than for this purpose r_i need not depend on σ_i. Hence we shall often simplify and write $r_i(\sigma_j)$ instead of $r_i(\sigma)$.

Remark 4. Uncertain costs are easily incorporated into the above formulation. For each player i we introduce a set $C_i \subset \mathbb{R}_+ := [0, \infty)$ with a distribution function F_i on it. Before the strategy is chosen each player knows only his own cost; the distribution functions are part of the model. Each type $c_i \in C_i$ then chooses a strategy that gives the maximum expected utility, taking expectations over the cost c_j of the other player using the function F_j.[15] When costs are drawn from C_i player i's reaction function is a collection

$$\{r_{i,c_i} : \mathcal{G} \to \mathcal{A}_i\}_{c_i \in C_i}.$$

A polar case of this model is the one with known costs, where $C_i = \{c_i\}$, and F_i is the Dirac function $\delta(c_i)$. It is useful to think of the relation between these two models as that between static games of incomplete and complete information respectively.

4 Equilibrium: Known Costs

This section studies the model with known costs of deception. In part this will serve as a benchmark to clarify the role of uncertainty in sustaining deception. Once the sets of response functions R_i, the gestures $\mathcal{G}_i$, the mappings b_i, and the costs of deception c_i are specified for $i \in \{1, 2\}$, the augmented game $G(\mathcal{I}; c)$ reduces to a normal-form game with strategy sets $\mathcal{S}_i = [0, 1] \times R_i \times G_i$ and payoff functions (with a slight abuse of notation):

$$\begin{aligned} \widehat{\pi}_i(s_i, s_j) \quad := \quad & \pi_i \left[r_i(b_j(r_j)), r_j(b_i(r_i)) \right] (1 - p_i)(1 - p_j) + \pi_i \left[r_i(b_j(r_j)), r_j(g_i) \right] p_i(1 - p_j) \\ & + \pi_i \left[r_i(g_j), r_j(b_i(r_i)) \right] (1 - p_i) p_j + \pi_i \left[r_i(g_j), r_j(g_i) \right] p_j p_i - c_i p_i, \end{aligned}$$

where $s_i = (p_i, r_i, g_i) \in \mathcal{S}_i$ for $i \in \{1, 2\}$.

Definition 2. *A Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ is a pair of strategies $(s_1^*, s_2^*) \in \mathcal{S}_1 \times \mathcal{S}_2$ such that for $i, j \in \{1, 2\}$ with $i \neq j$ we have $\widehat{\pi}_i(s_i^*, s_j^*) \geq \widehat{\pi}_i(s_i, s_j^*)$ for any $s_i \in \mathcal{S}_i$.*

Thus, the augmented game $G(\mathcal{I}; c)$ admits the existence of a Nash equilibrium under the usual conditions on strategies and payoffs under which strategic form games admit the existence of a

[15] A slightly more general model would allow the pair (c_1, c_2) to be drawn from $C_1 \times C_2$ using a joint distribution F that is not necessarily the product of marginals F_1 and F_2. Note that experiments are distinguished by, among other things, the length of the communication. The model thus far does not incorporate the reasonable restriction that longer periods of communication lead to better chances of "detection" when players attempt to deceive. We could do so using a signal θ_i correlated with c_i; while c_i is private information this allows players a better peek into how honest or transparent the other party is.

Nash equilibrium. In particular, the augmented game $G(\mathcal{I}; c)$ typically "inherits" all the Nash equilibria of the game G, as the following result states formally.

Proposition 1. *If the augmented game $G(\mathcal{I}; c)$ contains all strategies of the underlying game G as constant mappings, then for any Nash equilibrium (α_1, α_2) of G, the deception-free strategies $(0, r_1, b_1(r_1))$ and $(0, r_2, b_2(r_2))$ where $r_i(\sigma_j) = \alpha_i$ for all $\sigma_j \in \mathcal{G}_j$, $i \neq j$, constitute a Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ yielding the same outcome (α_1, α_2) as in the underlying game G.*

Proof. If players play the strategies $(0, r_1, b_1(r_1))$ *and* $(0, r_2, b_2(r_2))$ in the augmented game then player j's action is independent of player i's response function. This implies that player i's induced action in the underlying game G must be a best response to player j's induced action. The conclusion follows since (α_1, α_2) is a Nash equilibrium of G. □

The next three examples illustrate various ways in which players employ communication and deception to affect their payoffs in the underlying game. The first shows that communication together with the cost of deception allows a player to increase his payoff by proving a commitment to an induced strategy that is not a best response in the underlying game.

Example 1 (Unconditional Commitment to an Action). Consider the normal-form game below.

	L	R
U	$3,3$	$1,4$
D	$0,1$	$2,2$

Strategy L is dominated by strategy R for Player 2 and so the only Nash equilibrium of the game is (D, R) with payoffs $(2,2)$. But, if the cost of deception for Player 2 is high enough, say $c_2 > 1$, and $b_2(\widehat{L}) \neq b_2(r_2)$ for every $r_2 \neq \widehat{L}$, the strategies $(0, r_1^*, b_1(r_1^*))$ and $(0, \widehat{L}, b_2(\widehat{L}))$ where

$$\widehat{L}(\sigma_1) = \left\{ \begin{array}{ll} L & \text{for any } \sigma_1 \in \mathcal{G}_1 \end{array} \right. \quad ; \quad r_1^*(\sigma_2) = \left\{ \begin{array}{ll} U & \text{if } \sigma_2 = b_2(\widehat{L}) \\ D & \text{if } \sigma_2 \neq b_2(\widehat{L}) \end{array} \right.$$

are an equilibrium of the augmented game in which player 2 "commits" to playing L so that play of (U, L) with payoffs $(3,3)$ can be sustained. □

In the above example a player benefits by committing to a fixed action. The next example is more nuanced as he benefits by committing not to a fixed action but to punish the other player if she does not play the action desired by him.

Example 2 (Conditional Commitment to Punish). Consider the normal-form game below.

	L	R
U	$3,3$	$0,0$
D	$2,4$	$1,0$

Strategy R is dominated by strategy L for Player 2 and the only Nash equilibrium of the game is (U, L), with the payoff profile $(3, 3)$. But, if player 1's cost of deception is sufficiently high, say $c_1 > 1$, and $b_1(\widehat{D}) \neq b_1(r_1)$ for every $r_1 \neq \widehat{D}$, then the strategies $(0, \widehat{D}, b_1(\widehat{D}))$ and $(0, r_2^*, b_2(r_2^*))$ where

$$\widehat{D}(\sigma_2) = \left\{ \begin{array}{ll} D & \text{for any } \sigma_2 \in \mathcal{G}_2 \end{array} \right. \quad ; \quad r_2^*(\sigma_1) = \left\{ \begin{array}{ll} L & \text{if } \sigma_1 = b_1(\widehat{D}) \\ R & \text{if } \sigma_1 \neq b_1(\widehat{D}) \end{array} \right.$$

constitute an equilibrium of the augmented game.

The equilibrium is sustained by player 2's "commitment" to penalize player 1 by playing R if player 1 deviates from playing D. If the cost of deception is sufficiently high, then it is too costly for player 1 to fool player 2 into believing that player 1 is playing D when he plays the induced action U, which is better for him. Thus, it is in the best interest of player 1 to "surrender" to player 2's threat and to play D.[16] □

Intuition might suggest that, since players anyway best respond to each other's true strategies in equilibrium, deception cannot arise in equilibrium. The next example shows that this intuition fails in our setting.

Definition 3. *Deception occurs in equilibrium if (i) $g_i \neq b_i(r_i)$, i.e. at least one player i's bodily gesture is different from the natural bodily gesture associated with her chosen mapping, and (ii) $p_i > 0$.*

Our definition of deception is "behavioral" in that it does not require deception to be successful in fooling the other player.

Example 3. (Deception Equilibrium) Consider the following normal-form game.

[16] The equilibrium in this example is sustained by player 2's "threat" to play the dominated action R if player 1 does not play in the way that player 2 wants her to. This raises the question, What is the set of equilibria payoffs in the augmented game when players are constrained to only use "credible" strategies?

A natural definition of a credible threat seems to be the following. A player's strategy is credible if it plays a best response to whatever the player learns about the other player. There are two points at which this best response can be evaluated. The first point is at the start of the game. After all, in spite of the intuitive interpretation of our game as a sequential game, our model is of a simultaneous one shot game. The game is analyzed as such, using a standard Nash equilibrium. Thus, a strategy which best responds to the opponent's strategy is credible by definition. Consequently, any deviation to another strategy which plays a best response to what a player learns about the other player may trigger the other player to switch to another action, which may end up making the player worse off.

The second point at which the issue of best response can be raised is at the underlying game itself. In that game, taking the induced action of the other player as given, is a player's action a best response ? While such a requirement may seem desirable, note that it precludes the possibility of commitment, and would therefore rule out the (cooperate, cooperate) outcome induced by the (nice, nice) equilibrium. Moreover, in a game such as battle of the sexes, such a notion of credibility implies that each player is able to induce its favorite equilibrium as an outcome (by deviating to a strategy that plays this action with certainty and counting on the other player to best respond), which means that there is no equilibrium with credible strategies for this game.

Finally, if the cost of deception is low, then players can send arbitrary gestures that are independent of their strategies. This robs the idea of credibility of much of its force, because why should a player take seriously a gesture that has possibly nothing to do with the player's actual strategy.

	C	N
C	$1,1$	$0,0$
N	$0,0$	$0,0$

If players' deception costs are not too high and $b_i(\widehat{N}) \neq b_i(r_i)$ for every $r_i \neq \widehat{N}$, $i \in \{1,2\}$, then the pair of strategies $((1, r_C, b_1(\widehat{N})), (1, r_C, b_2(\widehat{N})))$, where $\widehat{N}$ is the constant mapping that plays N regardless and

$$r_C(\sigma_{-i}) = \begin{cases} C & \text{if } \sigma_{-i} = b_{-i}(\widehat{N}) \\ N & \text{if } \sigma_{-i} \neq b_{-i}(\widehat{N}), \end{cases}$$

are an equilibrium of the augmented game; each player engages in costly deception in equilibrium to lead the other player into believing that he will play $\widehat{N}$, only to play the mapping r_C. Observe that in this equilibrium, both players benefit from deception because it allows them to play the cooperative outcome (C, C). However, this benefit does not come at the expense of the other player. There is something silly about this equilibrium because players can instead play the cooperative outcome (C, C) more transparently without incurring the costs of deception, which make the equilibrium inefficient.[17] □

The next proposition shows that Example 3 illustrates a general result. Examples 1 and 2 showed that in our model with costly deception a Nash equilibrium in the augmented game does not necessarily induce a pair of (possibly mixed) actions in the underlying game that constitute a Nash equilibrium. Neither of those examples involved deception. We now pose the main questions. Can deception expand the range of equilibrium payoffs? Can it benefit one player at the expense of another? The former question is answered in Propositions 2 and 3. The latter is answered in Example 4, and in Section 5.

Proposition 2. *An equilibrium of the augmented game in which* both *players engage in equilibrium deception induces a Nash equilibrium of the underlying game G for generic deception costs.*[18]

[17] This equilibrium is similar to the equilibrium in Demichelis and Weibull's (2008) Example 1 in that players say they will do one thing and require the other player to say the same thing, only to do something else; the key difference is that their equilibrium is efficient.

[18] The term "generic" refers to all but one specific value of the deception cost for each player. The following example demonstrates the necessity of this qualification. Let the underlying normal-form game G be given by the matrix below.

	L	R
T	$0,1$	$0,2$
M	$0,2$	$0,0$
B	$0,0$	$0,0$

Suppose that the costs of deception are $c_1 = c_2 = 1$. Consider the strategies $\left(0, r_1^*, b_1\left(r_1^*\right)\right)$ and $(p_2, \widehat{L}, b_2(\widehat{R}))$ where

$$r_1^*(\sigma_2) = \begin{cases} T & \text{if } \sigma_2 = b_2(\widehat{L}) \\ M & \text{if } \sigma_2 = b_2(\widehat{R}) \\ B & \text{otherwise} \end{cases}$$

Thus, the players' actions in an equilibrium of the augmented game in which both players deceive each other are identical to their actions in some Nash equilibrium of the underlying game G. It should be noted, however, that from an ex-ante perspective, Proposition 2 is also consistent with the players mixing over different Nash equilibria of the underlying game.

Proof. Let s be a Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ where each $s_i = (p_i, r_i, g_i)$ is such that player i deceives player j, i.e. $p_i > 0$ and $g_i \neq b_i(r_i)$. For $i = 1, 2$, let $\alpha_i(r, p_j, g_j)$ denote the expected action induced from player i in this augmented equilibrium. Since the cost of deception is linear in the probability of deception, either $p_i = 1$ or player i is indifferent between the strategies (p_i, r_i, g_i) and $(1, r_i, g_i)$ in $G(\mathcal{I}; c)$. If $p_i = 1$ then a deviation by i to a different response r_i' cannot be detected by player j, i.e. player j's induced action in the underlying game α_j is independent of r_i. If r_i is such that $\alpha_i \notin BR_i(\alpha_j)$ then player i can raise his payoff by deviating to a response function that induces an action that is a best-response to α_j (this is where the independence of α_j and the response of i is used.) Therefore player i's induced action in the underlying game G must be a best response to player j's induced action in the underlying game G. This logic does not apply when i is indifferent between the strategies (p_i, r_i, g_i) and $(1, r_i, g_i)$ — this occurs for only one specific value of player i's cost of deception c_i and is thus nongeneric. If $p_j > 0$ then α_j is also a best response to α_i (except for at most one value of c_j) and thus the pair of induced actions constitute a Nash equilibrium in the underlying game G. □

The proof of Proposition 2 hinges on the costs being linear, which ensures that if any deception is optimal then full deception is optimal (at least for generic costs). Under convex costs intermediate deception might be optimal.

Recall our question whether deception can benefit one player at the expense of another. Example 4 below shows that this is possible, but where the *deceived* player is the one who benefits at the expense of the deceptive player, compared to any equilibrium of the underlying game with only cheap talk communication. This leaves open the question of whether equilibrium deception can benefit the deceptive player at the expense of the deceived player. We return to this question in Section 5 below.

Example 4. (Deception Hurts the Deceiver) Consider the normal-form game below.

	L	M	R
U	$1, 3$	$0, 3$	$1, 4$
D	$0, 1$	$-1, 1$	$2, 2$

Strategies L and M are dominated by strategy R for Player 2 and so the only Nash equilibrium of the game below is (D, R) with payoffs $(2, 2)$. If the cost of deception for Player 2 is high enough,

and the mappings $\widehat{L}$ and $\widehat{R}$ denote the constant mappings that play L and R, respectively, regardless of what is learned about the other player.

The strategies $(0, r_1^*, b_1(r_1^*))$ and $(p_2, \widehat{L}, b_2(\widehat{R}))$ form a Nash equilibrium of the augmented game for every $p_2 \in [0, 1]$. We need not worry about player 1. As for player 2, his induced payoffs is 1, which is maximal given r_1^* because player 2 can get a payoff of 2 only when he deceives player 1 but this costs him 1 for any "unit" of deception.

that for player 1 is low enough, $b_1(\widehat{D}) \neq b_1(r_1)$ for every $r_1 \neq \widehat{D}$, and $b_2(\widehat{LM}) \neq b_2(r_2)$ for every $r_2 \neq \widehat{LM}$, then the strategies $(1, \widehat{UD}, b_1(\widehat{D}))$ and $(0, \widehat{LM}, b_2(\widehat{LM}))$ defined by

$$\widehat{LM}(\sigma_1) = \begin{cases} L & \text{if } \sigma_1 = b_1(\widehat{D}) \\ M & \text{if } \sigma_1 \neq b_1(\widehat{D}) \end{cases} \quad \text{and}$$

$$\widehat{UD}(\sigma_2) = \begin{cases} U & \text{if } \sigma_2 = b_2(\widehat{LM}) \\ D & \text{if } \sigma_2 \neq b_2(\widehat{LM}) \end{cases} \; ; \; \widehat{D}(\sigma_2) = \; D \quad \text{for any } \sigma_2 \in \mathcal{G}_2$$

constitute an equilibrium of the augmented game. In this equilibrium player 2 "commits" to playing the action L but only if player 1 commits to playing $\widehat{D}$, and "threatens" to play the action M, which player 1 dislikes, otherwise. Player 1 deceives player 2 into believing he plays the constant mapping $\widehat{D}$ while actually playing the mapping $\widehat{UD}$, which is a best response to player 2's induced strategy in the underlying game above. The equilibrium outcome is (U, L) with payoffs $(1, 3)$, respectively. □

Earlier examples showed that making deception costly allowed commitment and expanded the set of equilibrium payoffs. This expansion was achieved without deception. The reliance on deception in Example 4 led to a payoff profile that is weakly dominated by another equilibrium payoff profile of the augmented game. The next proposition shows that this is true in general — if players' costs of deception are commonly known then deception is "redundant" in that it does not expand the range of equilibrium outcomes, and "undesirable" in that it gives worse payoffs than an equilibrium of the underlying game.

Proposition 3. *For an augmented game $G(\mathcal{I}; c)$ with generic costs and any equilibrium s^* thereof that has positive deception, i.e. $p_1^* + p_2^* > 0$, there exists a deception-free equilibrium of $G(\mathcal{I}; c)$ that induces the same actions in the underlying game G, generating weakly higher payoffs to both players.*

The proof constructs an equilibrium where players' minmax strategies in the underlying game G are used as disciplining devices. Player i gets his minmax value w_i when j plays her minmaxing action profile $\psi_j \in \mathcal{A}_j$, i.e.

$$w_i = \min_{\alpha_j \in \mathcal{A}_j} \max_{a_i \in A_i} \pi_i(a_i, \alpha_j) = \max_{a_i \in A_i} \pi_i(\alpha_i, \psi_j).$$

Proof. Let $s^* := (p_i^*, r_i^*, g_i^*)_{i=1,2}$ be a Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ that induces each $i \in \{1, 2\}$ to play the actions

$$\alpha_i^* := (1 - p_j^*) r_i^*(\sigma_j = b_j(r_j^*)) + p_j^* r_i^*(\sigma_j = g_j^*)$$

in the underlying game G, and generate payoffs $v_i^* = \pi_i(\alpha_i^*, \alpha_j^*) - p_i^* c_i$. By our richness assumption, for each $k \in \{1, 2\}$, there exist gestures $b_k(r_k^{**})$ such that $b_k(r_k^{**}) \neq b_k(r_k)$ for any $r_k \neq r_k^{**}$.

Hence the following reaction functions are well-defined:

$$r_1^{**}(\sigma_2) = \begin{cases} \alpha_1^* & \text{if } \sigma_2 = b_2\left(r_2^{**}\right) \\ \psi_1 & \text{otherwise} \end{cases};$$
$$r_2^{**}(\sigma_1) = \begin{cases} \alpha_2^* & \text{if } \sigma_1 = b_1\left(r_1^{**}\right) \\ \psi_2 & \text{otherwise} \end{cases}.$$

Clearly, if $s_1^{**} = (0, r_1^{**}, b_1(r_1^{**}))$ and $s_2^{**} = (0, r_2^{**}, b_2(r_2^{**}))$ are played, then the induced actions are the same as in the original equilibrium. Payoffs may be higher because (s_1^{**}, s_2^{**}) involves no deception.

We shall show that s^{**} is also a Nash equilibrium in $G(\mathcal{I}; c)$. If players follow s^{**} the payoff profile is $\pi(\alpha_i^*, \alpha_j^*)$. The best deviation of player i that does not involve deception is to the strategy $(0, r_i^{\text{maxmin}}, b_i(r_i^{\text{maxmin}}))$, where r_i^{maxmin} is a constant reaction function that maps to player i's maxmin action (which could be mixed). By *Sion's minmax theorem* (see Sion, 1958) the players' payoffs under s^{**} is no worse than their minmax payoff, which is what a player would get if he deviates, and the fact that s^* is an equilibrium implies that the payoff of i from s^* exceeds this, i.e. $v_i^* = \pi_i(\alpha_i^*, \alpha_j^*) - p_i^* c_i \geq w_i$, and therefore $\pi_i(\alpha_i^*, \alpha_j^*) \geq w_i$. Hence, by construction of r_j^{**} a unilateral deviation by i from s_i^{**} without using deception (i.e. with $p_i = 0$) leads to a payoff smaller than or equal to the minmax and is hence not profitable.

Now consider deviations by i to strategies with deception. Let $\overline{BR}_{i,\alpha_j^*} \in R_i$ denote the constant mapping that plays the best response BR_{i,α_j^*} against the induced action α_j^* irrespective of the signal σ_j. Thus, the most profitable deviation from $(0, r_i^{**}, b_i(r_i^{**}))$ is $(1, \overline{BR}_{i,\alpha_j^*}, b_i(r_i^{**}))$, which gives player i the payoff

$$\hat{\pi}_i(\overline{BR}_{i,\alpha_j^*}, r_j^{**}) = \pi_i(BR_{i,\alpha_j^*}, \alpha_j^*) - c_i$$

against $(0, r_j^{**}, b_j(r_j^{**}))$.

Player i could have played $(1, BR_{i,r_j^*(b(r_i^*))}, b(r_i^*))$ or $(1, BR_{i,r_j^*(g_i^*)}, g_i^*)$ against (p_j^*, r_j^*, g_j^*) in the original equilibrium (s_1^*, s_2^*). Thus,

$$v_i^* \geq \pi_i(BR_{i,r_j^*(b(r_i^*))}, r_j^*(b(r_i^*))) - c_i; \text{ and } v_i^* \geq \pi_i(BR_{i,r_j^*(g_i^*)}, r_j^*(g_i^*)) - c_i.$$

Because α_j^* is a mixture of $r_j^*(b(r_i^*))$ and $r_j^*(g_i^*))$ with probabilities $p_i^*$and $1 - p_i^*$, respectively, it follows that

$$\begin{aligned} v_i^* &\geq p_i^* \pi_i(BR_{i,r_j^*(g_i^*)}, r_j^*(g_i^*)) + (1 - p_i^*)\pi_i(BR_{i,r_j^*(b(r_i^*))}, r_j^*(b(r_i^*))) - c_i \\ &\geq p_i^* \pi_i(BR_{i,\alpha_j^*}, r_j^*(g_i^*)) + (1 - p_i^*)p_i^* \pi_i(BR_{i,\alpha_j^*}, r_j^*(b(r_i^*))) \\ &= \pi_i(BR_{i,\alpha_j^*}, \alpha_j^*) - c_i. \end{aligned}$$

□

Is it the case then that players cannot benefit from being able to easily deceive other players? As we show in the next section, the answer to this question is negative. Players are in fact

able to benefit from deception at the expense of the other player, but this hinges on asymmetric information about the players' costs of deception.

We conclude this section with a characterization of the set of equilibrium payoffs as a function of the costs of deception c_1 and c_2.

Definition 4. *A payoff v_i for player i is said to be individually rational if $v_i \geq w_i$.*

Definition 5. *A payoff profile (v_1, v_2) of a strategic form game G is feasible if there exists a possibly correlated action profile α such that*

$$(v_1, v_2) = \sum_{a \in A} \alpha(a_1, a_2)\pi(a_1, a_2).$$

Proposition 4. *Fix any game G in normal form. For any $\epsilon > 0$, there exists a cost of deception c^* and an interaction structure $(\mathcal{I}, c)$ such that any individually rational and feasible payoff profile v of G is within ϵ of an equilibrium payoff of $G(\mathcal{I}; c)$ provided that $c_1, c_2 \geq c^*$.*

Proof. The proof consists of two steps. We first show that it is possible to obtain any individually rational *independently mixed* action profile as an equilibrium payoff of the augmented game, and then show that it is possible to approximate any payoffs in the convex hull of the payoffs generated by the set of individually rational and independently mixed action profiles.

Let

$$v = (v_1, v_2) = \sum_{a_1 \in A_1} \sum_{a_2 \in A_2} \alpha_1(a_1)\alpha_2(a_2)\pi(a_1, a_2)$$

be an individually rational and independently mixed feasible payoff profile. Let F^{IIR} be the set of IR payoff profiles that are also independently mixed, and using these define

$$\mathcal{G}_i = \{g_x\}_{x \in F^{IIR}} \cup \{N\},$$

where the signal $\sigma_i = g_x$ conveys the message that i is playing the assigned role in obtaining the payoff x, and N is a gesture that indicates that i has picked a response function other than the ones that sustain some payoff x in F^{IIR}. Define a response profile $r^v := (r_1^v, r_2^v)$ as follows:

$$r_i^v(\sigma_j) = \begin{cases} \alpha_i & \text{if } \sigma_j = g_v \\ \psi_i & \text{otherwise.} \end{cases}$$

Let R_1 and R_2 be any pair of mutually consistent sets of response functions such that each $r_i^v \in R_i$ for all individually rational and feasible payoffs v. For each i, let $b_i(r_i) = g_v$ if and only if $r_i = r_i^v$. Let this structure be denoted by $\mathcal{I}$.

The strategy profile $s_v^* := (0, r_i^v, b_i(r_i^v))_{i=1,2}$ has payoff v; we show that it is a Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ if

$$c_i \geq c^* := \max_{(a_1,a_2),(a_1',a_2') \in A_1 \times A_2} \max_{j=1,2} \left\{ \pi_j(a_1, a_2) - \pi_j(a_1', a_2') \right\}. \tag{4.1}$$

If player 1 deviates to any other strategy with $p_1 = 0$, player 2 will play ψ_2 against him, giving player 1 a payoff of no more than w_1. However, since v is individually rational, $v_1 \geq w_1$. If player 1 chooses $p_1 > 0$, then his payoff is at most

$$(1 - p_1)w_1 + p_1 \left\{ \max_{a_1, a_2 \in A_1 \times A_2} \pi_1(a_1, a_2) - c_1 \right\},$$

which falls short of the equilibrium payoff π_1 if inequality (4.1) holds. The same argument holds for player 2. Hence the augmented strategy profile s^* is a Nash equilibrium of the augmented game $G(\mathcal{I}; c)$ for $c_i \geq c^*$.

Suppose now that α and β are both Nash equilibrium payoffs of the augmented game and that both are obtainable without deception as described above. We show that for every $q \in (0, 1)$ and every $\epsilon > 0$, it is also possible to sustain payoffs that are ϵ close to $q\alpha + (1 - q)\beta$. The construction is based on the idea of jointly controlled players' lotteries. Since the dyadic rationals are dense in the real line it is enough to restrict attention to q being a dyadic rational, i.e.

$$q = \frac{\tau^1}{2} + \frac{\tau^2}{4} + \ldots + \frac{\tau^L}{2^L},$$

where $\tau = (\tau^1, \ldots, \tau^L) \in \{0, 1\}^L$ for some integer L that is chosen to obtain the desired level of approximation.[19]

We expand the sets of gestures to

$$\tilde{\mathcal{G}}_i := \mathcal{G}_i \times \{\mathtt{a}, \mathtt{b}\}^L, \text{ where } \mathcal{G}_i = \{g_x\}_{x \in F^{IIR}} \cup \{N\},$$

where $g_x \neq g_y$ if $x \neq y$, and $N \neq g_x$ for all $x \in F^{IIR}$. The idea is that the bodily gesture can be decomposed into two components, the first of which signals whether or not i intends to play according to the suggested reaction function, and the second is a verbal message $m_i = (m_i^t)_{t=1}^L \in \{\mathtt{a}, \mathtt{b}\}^L$ that is transmitted verbatim to the listener (player j). Players use the jointly controlled lottery described below to coordinate play. Define another variable $m = (m^1, \ldots, m^L) \in \{0, 1\}^L$ such that $m^t = 1$ if $m_i^t = m_j^t$ and $m^t = 0$ if $m_i^t \neq m_j^t$. Calculate

$$q(m_1, m_2) = \frac{m^1}{2} + \frac{m^2}{4} + \ldots + \frac{m^L}{2^L}.$$

Player i does not engage in deception, and sends the message m_i. If $q(m_1, m_2) < q$, then player i plays a response that generates the payoffs α, and otherwise he plays a response that generates the payoffs β. Here we invoke the fact that in general a player's response may depend both on his own as well as the other player's gesture. Namely, there are many strategies in which the player plays the same exact response function but with different verbal gestures. Notice that if player j mixes over all the messages in $\{\mathtt{a}, \mathtt{b}\}^L$ uniformly, then for any $t \in \{1, ..., L\}$ the probability that

[19] The purpose of ϵ is that even if q is not a dyadic rational there is one within a distance ϵ of it. With messages unbounded in length, i.e., $L = \infty$ above, we can obtain the target payoffs exactly rather than approximately.

$m^t = 1$ is 0.5 regardless of what message player i chooses. So that it is a Nash equilibrium of the augmented game for each players to mix over all the messages in $\{\mathtt{a}, \mathtt{b}\}^L$, and in this equilibrium of the augmented game, the payoffs α are obtained with probability q. □

Tennenholtz (2004) proves a similar folk theorem for the case of a game played by computer programs that may condition their strategy on the other computer program with which they are matched to play the game. Kalai et al. (2010) offers a similar model and a folk theorem for the case where each player chooses a "commitment device" (rather than a computer program) that may depend on the commitment devices chosen by other players. Peters and Szentes (2009) explores games where each player writes a contract that obliges a player to respond with a specified action depending on the opponent's contract. They prove a similar folk theorem to Kalai et al. and further show that this result does not hold in an environment with incomplete information.[20] None of these papers allows for the possibility of deception. Another small point of difference is that our model achieves payoffs approximately rather than exactly because we use messages in a finite set $\{a, b\}^L$ to approximate a real number.

Proposition 4 shows that our model produces the same equilibrium payoffs as Tennenholtz (2004) and Kalai et al. (2010) if the costs of deception are high. But if the costs of deception are low, then our model produces the same equilibrium payoffs as probability distributions over Nash equilibria (intermediate deception costs produce equilibrium outcomes that lies between these two extreme cases).

Let $N[G]$ denote the set of Nash equilibrium action profiles of a strategic game G with cheap talk. Denote the set of equilibrium actions (in the underlying game) of $G(\mathcal{I}; c)$ by $N[G(\mathcal{I}; c)]$. If $c_1, c_2 \geq c^*$, then the previous proposition establishes a type of "folk theorem". The next two propositions describe how the set of equilibrium action profiles, and therefore equilibrium payoffs, varies with the costs of deception c_1 and c_2. Proposition 5 shows that the set of equilibrium action and payoff profiles increases monotonically with the costs of deception, where $\subset$ allows set equality. Proposition 6 shows that if deception is nearly costless ($c_1, c_2 \leq \epsilon$ for some ϵ close to zero), then the equilibrium payoffs of the augmented game $G(\mathcal{I}; c)$ approximate those in the convex hull of the Nash equilibrium payoffs of the underlying game G. This preserves continuity because the set $co(N[G])$ is known to equal the set of equilibria of the underlying game plus cheap talk (see, e.g., Aumann and Hart, 2003).

Proposition 5. *For any strategic form game G and for any interaction structure $\mathcal{I}$,*

$$c_1 \leq c_1',\ c_2 \leq c_2' \Rightarrow N[G(\mathcal{I}; c)] \subset N[G(\mathcal{I}; c')].$$

Proof. Let $G(\mathcal{I}; c)$ be an augmented game with known costs $(c_1, c_2) \in \mathbb{R}^2$. By Proposition 3 for any Nash equilibrium in $G(\mathcal{I}; c)$ there exists a deception-free equilibrium that induces the same

[20] See Forges (2013) for a generalization of Kalai et al.'s model to incomplete information.

actions and yields the same or higher payoffs. This last equilibrium still holds when costs are increased to $(c_1', c_2') \geqq (c_1, c_2)$: the same deviations are available in both cases, and the higher costs of deception make each deviation strictly less profitable with (c_1', c_2').

When there is no deception in equilibrium, the payoffs are independent of deception costs, and therefore any possible equilibrium payoff vector under (c_1, c_2) remains possible under (c_1', c_2'). □

Proposition 6. *If deception costs are sufficiently close to zero, then the equilibrium outcomes of any augmented game $G(\mathcal{I}; c)$ are contained in a set that approximates the convex hull of Nash equilibrium payoffs of G in the following sense: if α is in the convex hull of Nash equilibria of the game G and $\epsilon > 0$, then there exist finite sets $\mathcal{G}_1$ and $\mathcal{G}_2$ of gestures such that the set of Nash equilibrium payoffs of the augmented game $G(\mathcal{I}; c)$ includes points within ϵ of α. Furthermore, any equilibrium of $G(\mathcal{I}; c)$ is within ϵ of the convex hull of the Nash equilibria of G.*

Proof. The proof employs the device of jointly controlled lotteries, and is similar to the one used in the second part of the proof of Proposition 4. From Proposition 1 we know that $N[G] \subset N[G(\mathcal{I}; c)]$. Fix $\epsilon > 0$. It then suffices to show that if α and β are both Nash equilibria of the game G and $q \in (0, 1)$, we can find strategy sets $\mathcal{G}_i$ such that $q\alpha + (1-q)\beta \in N[G(\mathcal{I}; c)]$. Since the dyadic rationals are dense in the real line it is enough to restrict attention to q being a dyadic rational, i.e.

$$q = \frac{\tau^1}{2} + \frac{\tau^2}{4} + \ldots + \frac{\tau^L}{2^L},$$

where $\tau = (\tau^1, \ldots, \tau^L) \in \{0, 1\}^L$ for some integer L. Let $\mathcal{G}_i = \{\mathtt{a}, \mathtt{b}\}^L$. The players can coordinate on the equilibrium to be played through a jointly controlled lottery that is constructed as follows. The bodily gesture of player i includes the message $m_i = (m_i^t)_{t=1}^L \in \{\mathtt{a}, \mathtt{b}\}^L$. Define another variable $m = (m^1, \ldots, m^L) \in \{0, 1\}^L$ such that $m^t = 1$ if $m_i^t = m_j^t$ and $m^t = 0$ if $m_i^t \neq m_j^t$. Calculate

$$q(m_1, m_2) = \frac{m^1}{2} + \frac{m^2}{4} + \ldots + \frac{m^L}{2^L}.$$

If $q(m_1, m_2) < q$, then player i plays the constant response with action α, no deception and sends the message m_i, and otherwise he plays the constant response with action β, no deception, and sends the message m_i. Notice that if player j randomizes over all the messages in $\mathcal{G}_j$, then for any $t \in \{1, ..., L\}$ the probability that $m^t = 1$ is equal to 0.5 regardless of what gesture player i chooses. So that it is a Nash equilibrium of the augmented game for both players to randomize over all the messages in $\mathcal{G}_1$, $\mathcal{G}_2$, respectively, and in this equilibrium of the augmented game, the Nash equilibrium α of the underlying game is played with probability q.

To conclude the proof, we need to show that no other payoff can be sustained in an equilibrium of the augmented game. This follows from the fact that if the augmented game induces anything which is not a Nash equilibrium of the underlying game, then at least one of the players can deviate, while deceiving the other player into believing that she still plays the original equi-

librium, and switch to playing a best response in the underlying game. If the cost of deception is sufficiently small, this deviation would yield a strictly higher payoff to the player, which would destabilize the equilibrium of the augmented game. □

In summary, Proposition 4 shows that the commitment that can be achieved through communication in which players involuntarily reveal their strategies expands the set of equilibria and generates a folk theorem. Propositions 5 and 6 show that deception has the power to undermine this commitment: the ability to commit is monotonically increasing with the cost of deception; and easy deception completely undermines communication and commitment.

5 Uncertain Costs: Successful Equilibrium Deception

In this section we show by example that if the cost of deception is privately known by the players, or equivalently, if the cost of deception is heterogenous and unobservable and players are matched to play the game in pairs, then

- deception can be successfully practiced in equilibrium, and
- players who engage in deception benefit at the expense of players who are duped.

Example 5. Consider again the example of the Prisoner's Dilemma with the following payoffs.

	C	D
C	$3,3$	$0,4$
D	$4,0$	$1,1$

Suppose that the cost of deception is independently drawn from the uniform distribution on the interval $[0,2]$. The cost of deception is the private information of the players so that each player knows its own cost of deception and believes the cost of the other player to be uniform on the interval $[0,2]$. The following strategies are an equilibrium of the augmented game that is based on the Prisoner's Dilemma:

$$s_i(c_i) = \begin{cases} (1, \widehat{D}, b_i(nice)) & \text{if } c_i < 1 \\ (0, nice, b_i(nice)) & \text{if } c_i \geq 1, \end{cases}$$

where $\widehat{D}$ denotes the constant mapping that plays *Defect* regardless of what is learned about the other player, and *nice* is defined as in the introduction to be "if the other player plays *nice* then *Cooperate* and if the other player picks any other response function, play *Defect*."

In this equilibrium, a player with a high cost of deception ($c_i \geq 1$) commits to playing *nice*, which, since the other player plays *nice* and *Defect* with equal probabilities, yields the payoff $\frac{1}{2} \cdot 3 + \frac{1}{2} \cdot 0 = \frac{3}{2}$. This is higher than or equal to the payoff that the player can get from playing *Defect*,

which is 1; or from the payoff that the player can get by deceiving the other player, which is no more than $\frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 1 - 1 = \frac{3}{2}$ because with probability one half the other player would play *nice*, be deceived and therefore cooperate, but with probability one half the other player would anyway play *Defect*, and the player needs to deduct the cost of deception, which is at least one.

A player with a low cost of deception ($c_i < 1$) successfully deceives players with a high cost of deception into playing cooperate, and therefore obtains an equilibrium payoff of $\frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 1 - c_i = \frac{5}{2} - c_i$, which is larger than the payoff that the player would obtain if it just played *nice*, which is $\frac{3}{2}$ as shown above. □

This example shows how the model developed in this paper can account for the experimental evidence described in the introduction. Namely, the opportunity to communicate increases the extent of cooperation; but nevertheless, some players still succeed in inducing the other player to play cooperatively against their own defection.

The observation that is more difficult to account for is the players' correlated play, or the fact that players are more likely to both play cooperatively than can be accounted for by independent mixing. Indeed, the probabilities with which the players play each strategic combination in this example are:

	C	D
C	$\frac{1}{4}$	$\frac{1}{4}$
D	$\frac{1}{4}$	$\frac{1}{4}$

which indicates independent mixing.

However, correlated play can easily be accounted for by our model under one of two assumptions. First, if the players' costs of deception are correlated, which seems sensible enough because all it requires is that players with a high cost of deception would be slightly more inclined to believe that the other player is also likely to have a high cost of deception and players with a low cost of deception would be slightly more inclined to believe that the other player is also likely to have a low cost of deception. Second, we may enrich the model to allow the circumstances of the conversation to influence both players ability to infer the strategy of the other player. For example, if one player is more extroverted and communicative it might make it easier for each to spot the intentions of the other.

The next example illustrates how our model can rationalize the famous example of the Trojan horse.

Example 6. The story of the Trojan horse is taken from Homer's *Illiad*. The Greeks and Trojans fought bitterly for ten years with neither side being able to defeat the other and claim victory. After ten years, the Greeks pretended to retreat and left behind them a large wooden horse with a few Greek warriors hidden in its belly. The Trojans debated whether the horse is a Greek trick that should be burned on the spot or brought into Troy as a symbol of their victory. They decided to

bring the horse into Troy. When night fell, and the unsuspecting Trojans were all asleep, the Greek warriors came out of the horse's belly and opened Troy's gates to the returning Greek army, who exploited this opportunity and destroyed Troy.

We describe the situation as a 2×2 game as follows:

Trojans:Greeks	*not*	*trick*
in	$2,0$	$-2,2$
burn	$1,0$	$1,-1$

The Greeks either attempt to fool the Trojans or not, and the Trojans either take the horse into the city or burn it. We assume that the Greek's privately known cost of deception is independently drawn from the uniform distribution on the interval $[0,8]$. The following pair of strategies is an equilibrium of the augmented game that is based on this deception game:

$$s_G(c_i) = \begin{cases} \left(1, \widehat{trick},\, b_G(\widehat{not})\right) & \text{if } c_G < 2 \\ \left(0, \widehat{not},\, b_G(\widehat{not})\right) & \text{if } c_G \geq 2 \end{cases}$$

and$s_T = (0, r_T(\sigma_G), b_T(r_T))$ where $r_T(\sigma_G)$ is given by

$$r_T(\sigma_G) = \begin{cases} in & \text{if } \sigma_G = b_G(\widehat{not}) \\ burn & \text{if } \sigma_G \neq b_G(\widehat{not}) \end{cases}.$$

In this equilibrium, the Greeks are wily and Troy is destroyed with probability 0.25. □

The two last examples show that successful equilibrium deception depends on the proportion of cheaters, or the probability of deception, not being too large. This holds more generally as well.

The formal description of the game with uncertain costs is similar to the case of known costs. Given a game G in strategic form denote the interaction structure by $(\mathcal{I}, C_1, C_2, \mu)$ where $\mathcal{I}$ is as before, $C_i \subset \mathbb{R}_+$ for $i \in \{1,2\}$, and μ is a measure on $C_1 \times C_2$ obtained as the product of independent measures μ_1 and μ_2 on C_1 and C_2 respectively. We say that a measure f^+ on $\mathbb{R}_+$ first order stochastically dominates another measure f^- if

$$f^-[0,x] \geq f^+[0,x] \; \forall x \in \mathbb{R}_+.$$

We write this as $f^+ \succcurlyeq f^-$, where the distribution f^+ represents higher deception costs.

As noted earlier, in a game of incomplete information, a strategy s_i of player i is a collection of tuples, one per cost type:

$$\{s_i(c_i) = (p_i(c_i), r_i(c_i), g_i(c_i))\}_{c_i \in C_i}, \text{ where } p_i(c_i) \in [0,1], r_i(c_i) \in R_i, g_i(c_i) \in \mathcal{G}_i.$$

An *outcome function* is a mapping $d : C_1 \times C_2 \to \mathcal{A}$, where $d(c_1, c_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ is the action profiles played when the type profile (c_1, c_2) is drawn. Let $N[G(\mathcal{I}; C_1, C_2, \mu)] \subset \mathcal{A}^{C_1 \times C_2}$ denote the set of

all outcome functions that arise in Bayesian-Nash equilibria (BNE) of the game G augmented by the interaction structure $(\mathcal{I}; C_1, C_2, \mu)$.

It seems natural enough to conjecture the following analogue of our monotonicity result (Proposition 5) for the case of known costs of deception: For any strategic game G and any interaction structure $(\mathcal{I}, C_1, C_2, \mu)$ with privately known costs of deception,

$$(\mu_i^+ \succcurlyeq \mu_i^- \forall i) \Rightarrow N[G(\mathcal{I}; C_1, C_2, \mu^-)] \subset N[G(\mathcal{I}; C_1, C_2, \mu^+)].$$

In other words, the conjecture is that an outcome function continues to be sustainable in equilibrium when the cost distribution goes up (in the sense of stochastic dominance).

However, the reasoning that works for known costs fails to deliver a proof of this conjecture with unknown costs. To see why, consider moving from μ^- to μ^+, satisfying first order stochastic dominance as defined above. Suppose that all types of players i and j continue to behave as they did before. Fix a cost type c_i of player i. In general the change from μ_j^- to μ_j^+ means that the distribution of actions of j changes, on account of a strictly larger proportion of players no longer finding it optimal to deceive. This could give type c_i a new incentive to deviate. The key difference between the known and unknown cost cases is the fact that in the former each player can reason if the other player is cheating or not, whereas in the latter some cost types may find it profitable to deceive, while others do not; in general, both types would continue to exist as we shift costs upwards, leading to the problem explained above. In some sense it is more useful to think of the known-cost result in the following way: the move from c_1 to $c_1' > c_1$ is an upward shift of the support of the (Dirac) distribution, rather than stochastic dominance.

As asking for monotonicity to hold on a type-by-type basis seems too demanding, we turn to the distribution of actions averaged over types. For any player i, the strategy profile s assigns probability $\alpha_i^*[s](a_i, c_i)$ to the action a_i conditional on the cost type c_i being drawn; this may be calculated as in the known-cost case. The induced distribution of s is represented by the probability distribution $\beta[s] = \beta_1[s] \times \beta_2[s]$. Each $\beta_i[s]$ is a probability distribution over the corresponding A_i:

$$\beta_i[s] : A_i \to [0,1] \text{ and } \sum_{a_i \in A_i} \beta_i[s](a_i) = 1,$$

where the player i's average probability of playing $a_i \in A_i$ is

$$\beta_i[s](a_i) := \int_{C_i} \alpha_i^*[s](a_i, c_i)\, d\mu_i(c_i) \in [0,1].$$

We now state a modified 'monotonicity result'. The complexity of the proof contrasts with the corresponding result for the case of known costs.

Proposition 7. *Take any strategic game G and any interaction structure $(\mathcal{I}, C_1, C_2, \mu)$ with privately known deception costs. Let $\mu_i^+ \succcurlyeq \mu_i^- \forall i$. If s^- is an equilibrium of $G(\mathcal{I}; C_1, C_2, \mu^-)$, then there exists an equilibrium s^+ of $G(\mathcal{I}; C_1, C_2, \mu^+)$, such that $\beta[s^-] = \beta[s^+]$.*

Proof. Find the cutoff cost type above which s_i^- does not involve deception, i.e.

$$c_i^{(1)} := \sup\{c_i \in C_i | p_i(c_i) > 0\}.$$

For any $c_i > c_i^{(1)}$, we thus have $p_i = 0$; since types are private information and utility functions are the same for all types conditional on $p_i = 0$, it is without loss of generality to assume that all these types play the same action in the equilibrium s^-, i.e.

$$\alpha_i^*[s^-](\cdot, c_i) = \alpha_{i,H}^- \; \forall c_i > c_i^{(1)},$$

where the H in the subscript is meant to suggest that these are relatively high cost types of player i. Similarly, from the utility functions it follows that all types with costs $c_i < c_i^{(1)}$ will choose $p_i(c_i) = 1$ and play a best response to $\beta_j[s^-]$; let us assume without loss of generality that they play the same distribution $\alpha_{i,L}^-$. Letting $\mu_i^-[c_i^{(1)}, \infty) \cap C_i =: \mu_{i,H}^-$ denote the probability of types who do not deceive in the equilibrium s^-. Hence

$$\beta_j[s^-] = (1 - \mu_{i,H}^-)\alpha_{i,L}^- + \mu_{i,H}^- \alpha_{i,H}^-.$$

Let $\mu_i^+[c_i^{(1)}, \infty) \cap C_i =: \mu_{i,H}^+$. Since $\mu_i^+ \succcurlyeq \mu_i^-$, we have $\mu_{i,H}^- \leq \mu_{i,H}^+$. Define a new distribution $\alpha_{i,H}^+$ using the following equality

$$\beta_j[s^+] = (1 - \mu_{i,H}^+)\alpha_{i,L}^- + \mu_{i,H}^+ \alpha_{i,H}^+.$$

Clearly the inequality $\mu_{i,H}^- \leq \mu_{i,H}^+$ implies that $\alpha_{i,H}^+$ is a convex combination of $\alpha_{i,H}^-$ and $\alpha_{i,L}^-$. Following this same procedure we can modify the strategy of player j to maintain the same distribution over actions even after the change in cost distributions: $\beta_j[s^+] = \beta_j[s^-]$.

Having constructed the strategy profile s^+ from s^-, we now check that s^+ is an equilibrium of $G(\mathcal{I}; C_1, C_2, \mu^+)$. This is enough because by construction we have $\beta[s^-] = \beta[s^+]$.

- First consider the incentive of types $c_i > c_i^{(1)}$. Since $\alpha_{i,H}^+$ is a convex combination of $\alpha_{i,H}^-$ and $\alpha_{i,L}^-$, and the latter $\alpha_{i,L}^-$ is a best response to $\beta_j[s^-]$, the action $\alpha_{i,H}^+$ is at least as good a response to $\beta_j[s^-]$ as $\alpha_{i,H}^-$; since a player of type $c_i > c_i^{(1)}$ had no incentive to deceive under the earlier distribution he will have no incentive to do so now if $\beta_j[s^-] = \beta_j[s^+]$ (improving the payoff to honesty cannot lead to perverse incentives).

- Types with $c_i < c_i^{(1)}$ are playing a best response conditional on $p_i = 1$. Since deception is costly, all that remains is to check that they do not want to deviate to $\alpha_{i,H}^+$, which gives a higher payoff than their earlier honest option $\alpha_{i,L}^+$. If some of them "become honest", we need to iteratively modify the actions of the honest types. In other words, we find a new cutoff $c_i^{(2)} < c_i^{(1)}$ such that cost types above $c_i^{(2)}$ find it profitable to not deceive. If at some step this process stops, i.e.

$$c_i^{(k+1)} = c_i^{(k)}, \; c_j^{(k+1)} = c_j^{(k)},$$

then the corresponding strategy profile constitutes an equilibrium. If not, the decreasing sequence $(c_i^{(k)})$ must have a limit. The corresponding limit of the strategies must exist as well; this limiting distribution is the required equilibrium of $G(\mathcal{I}; C_1, C_2, \mu^+)$ with the same distribution of actions as the equilibrium s^- of $G(\mathcal{I}; C_1, C_2, \mu^-)$. □

Finally, even with uncertain costs of deception, the set of *average* equilibrium actions (and hence payoffs) varies from mixtures over Nash in the *underlying* game with cheap talk (where no deception is possible) to the case where a folk theorem is possible. As with known costs of deception, the former case occurs when players are sufficiently likely to have a low enough cost of deception, while the former occurs when distribution of costs puts sufficiently high probability on costs above a threshold. The monotonicity result Proposition 7 tells us that once we have the folk theorem, increasing costs according to first-order stochastic dominance maintains the folk theorem. It is worth noting that these results pertain to 'average' distributions, not type-by-type distributions. In other words, when we obtain a folk theorem, types with low deception costs will deceive and continue to best respond (in expectation) to the the other player, but the equilibrium is sustained by a large mass of players who do not find it hard to deceive. We omit formal versions as they are similar to those with known costs of deception.

6 Conclusion

This paper proposes a theory of equilibrium deception that draws on psychology and experimental evidence. One contribution is to note that real communication includes but is not limited to cheap talk. In other words, what we say may have consequences for how we act and the subjective satisfaction or utility we derive from actions. The standard notion of a strategy then appears too restrictive a concept. Indeed, it is not at the level of payoffs but at the level of strategies that we augment the standard model of a one-shot bilateral game, and propose a solution concept where deception co-exists with rationality, both in non-trivial forms.

We find two consequences of the model worth noting in relation to the problem of sustaining cooperation in societies and communities. Indeed this question has driven much research on other-regarding payoffs and in evolutionary game theory.

1. First, for deception to happen in equilibrium it must be somewhat but not too costly: if it is too easy to deceive then the 'rational actor' component of our model will take over and ensure that agents incorporate the possibility of deception and play something 'safe'; our formal results show that in this case 'safe' means Nash equilibria of the underlying game, augmented with correlation generated by standard cheap talk, because these are deception-proof. We augment rather than alter the standard model: as explained, our model and equilibrium concept reduce to the standard one with costless deception. One can draw a parallel with the way in which theories such as 'fairness' have broadened the reach of the standard model, but using payoffs rather than strategies and communication to do so.

2. The second connection is with 'monotonicity': roughly, a more honest society can sustain the same average distribution of actions that a less honest one can. A twist awaits us when costs are drawn using a (non-degenerate) distribution: the earlier outcome may not be sustainable with *each type* behaving the same way, but is sustainable when the average distribution of actions remains the same but the 'honest' types obtain a higher payoff than they received in the less honest society. The quotes around honest are meant to highlight that in our model honesty is the outcome of rational choice, as indeed is active deception.

References

[1] Amann and Yang (1998) "Sophistication and the persistence of cooperation," *Journal of Economic Behavior & Organization* 37, 91-105.

[2] Anderson, A. and L. Smith (2013) "Dynamic deception." *American Economic Review*, forthcoming.

[3] Aumann, R. (1974). "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics* 1, 67-96.

[4] Aumann, R. and S. Hart (2003). "Long cheap talk." *Econometrica* 71, 1619-1660.

[5] Axelrod, R. (1987) "The evolution of strategies in the iterated prisoner's dilemma," in *The Dynamics of Norms*, eds. C. Bicchieri, R. Jeffrey, and B. Skyrms, Cambridge Studies in Probability, Induction, and Decision Theory, Cambridge University Press.

[6] Battigalli, P. and M. Dufwenberg (2007) "Guilt in games," *The American Economic Review*, 170-176.

[7] Belot, M., V. Bhaskar, J. Van de Ven, G. (2010). "Promises and cooperation: evidence from a TV game show." *Journal of Economic Behavior & Organization* 73, 396-405.

[8] Belot, M., V. Bhaskar, J. Van de Ven, G. (2012). "Can observers predict trustworthiness ?" *Review of Economics and Statistics* 94, 246-59.

[9] Binmore, K. (1994). *Playing Fair: Game Theory and the Social Contract,* MIT Press, Cambridge, Massachusetts.

[10] Camerer, Colin F., Spezio, Michael, and Wang, Joseph Tao-yi (2010). "Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *American Economic Review* 100, 984-1007.

[11] Crawford, Vincent P. (2003) "Lying for strategic advantage: rational and boundedly rational misrepresentation of intentions," *American Economic Review* 93, 133-149.

[12] Croson, Rachel T.A. (2005) "Deception in economics experiments."*Deception in Markets: An Economic Analysis*, Palgrave Macmillan, Ch 5, 113-130.

[13] Dekel, E., J. Ely, and O. Yilnakaya, "The evolution of preferences," *Review of Economic Studies* 74, 685-704.

[14] Demichelis, S. and J. Weibull (2008) "Language, meaning, and games: a model of communication, coordination, and evolution" *American Economic Review* 98, 1292–1311.

[15] den Assem, V., M. J., Dennie Van Dolder, and R. H. Thaler. (2012) "Split or steal? cooperative behavior when the stakes are large" *Management Science* 58, 2-20.

[16] DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L, & Lee, J. (2012). "Detecting the trustworthiness of novel partners in economic exchange," *Psychological Science* 23, 1549-56.

[17] Ekman, P., M. O'Sullivan, and M. Frank (1999) "A few can catch a liar." *Psychological Science* 10, 263-66.

[18] Ellingsen, T. and R. Östling (2010) "When does communication improve coordination ?" *The American Economic Review* 100, 1695-1724.

[19] Ettinger, D. and P. Jehiel (2010) "A theory of deception" *American Economic Journal: Microeconomics* 2, 1-20.

[20] Farrell, J. (1988) "Communication, coordination and Nash equilibrium" *Economics Letters* 27, 209-214.

[21] Farrell, J. and M. Rabin (1996) "Cheap talk," *Journal of Economic Perspectives* 10, 103-118.

[22] Fehr, E. (2009) "On the economics and biology of trust." *Journal of European Economic Association* 7, 235-266.

[23] Fehr, E. and K. M. Schmidt (1999) "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics* 114, 817-868.

[24] Forges, F. (2013) "A folk theorem for Bayesian games with commitment," *Games and Economic Behavior* 78, 64-71.

[25] Forgó, F. (2010) "A generalization of correlated equilibrium: a new protocol," *Mathematical Social Sciences* 60, 186-190.

[26] Frank, R. H. (1987). "If homo economicus could choose his own utility function, would he want one with a conscience?," *American Economic Review* 77, 593-604.

[27] Frank, R. H. (1988). *Passions within Reason: the strategic role of the emotions.* New York: Norton.

[28] Gauthier, D. (1986). "Morals by agreement," Oxford: Clarendon Press.

[29] Geanakoplos, J., D. Pearce, and E. Stacchetti (1989) "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60-80.

[30] Gneezy, U. (2005) "Deception: The role of consequences," *American Economic Review* 95, 384-394.

[31] Guiso, L., P. Sapienza, and L. Zingales (2009). "Cultural biases in economic exchange?" *Quarterly Journal of Economics* 124, 1095-1131.

[32] Guth, W., and M. Yaari. (1992) "Explaining reciprocal behavior in simple strategic games: an evolutionary approach", Chapter 2 in: Witt, U. (Ed.), *Explaining Process and Change: Approaches to Evolutionary Economics.* Ann Arbor: University of Michigan Press.

[33] Hendricks, K. and P. R. McAfee, (2006). "Feints." *Journal of Economics and Management Strategy* 15, 431-56.

[34] Kalai, A. T., E. Kalai, E. Lehrer, and D. Samet (2010). "A commitment folk theorem," *Games and Economic Behavior* 69, 127-137.

[35] Kalay A., A. Kalay, and A. Kalay (2003). "Friends or foes? empirical test of a simple one-period Nash equilibrium," *mimeo*.

[36] Kartik, N. (2009) "Strategic communication with lying costs," *The Review of Economic Studies* 76, 1359-1395.

[37] Kartik, N., M. Ottaviani, and F. Squintani (2007) "Credulity, lies, and costly talk," *Journal of Economic Theory* 134, 93-116.

[38] Knack, S., and P. Keefer (1997). "Does social capital have an economic payoff? a cross-country investigation." *Quarterly Journal of Economics* 112, 1251–1288.

[39] La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny (1997) "Trust in large organizations," *American Economic Review* 87, 333-338.

[40] Levine, D. (1998) "Modeling altruism and spite in experiments," *Review of Economic Dynamics* 1, 593-622.

[41] López Pérez, R. (2012) "The power of words: a model of honesty and fairness," *Journal of Economic Psychology* 33, 642-658.

[42] Manzini, P., Sadrieh A., and N. J. Vriend (2009) "On smiles, winks and handshakes as coordination device," *The Economic Journal* 119, 826-854.

[43] Matsui, A. (1989) "Information leakage forces cooperation," *Games and Economic Behavior* 1, 94-115.

[44] Miettinen, T. (2012) "Promises and conventions: an approach to pre-play agreements," *Games and Economic Behavior* 80, 68-84.

[45] Ockenfels, A. and R. Selten (2000) "An experiment on the hypothesis of involuntary truth-signaling in bargaining" *Games and Economic Behavior* 33, 90-116.

[46] Peters, M., and B. Szentes (2012). "Definable and contractible contracts," *Econometrica* 80, 363-411.

[47] Robson, A. (1990) "Efficiency in evolutionary games: Darwin, Nash and the secret handshake." *Journal of Theoretical Biology* 144, 379-396.

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

[49] Sion, M. (1958). "On general minimax theorems," Pacific J. Math. 8, 171-176.

[50] Tennenholtz, M., (2004) "Program equilibrium," *Games and Economic Behavior* 49, 363-373.

[51] Wiseman, T. and O. Yilankaya (2001) "Cooperation, Secret Handshakes, and Imitation in the Prisoners' Dilemma" *Games and Economic Behavior* 37, 216-242.

[52] Yaari, M., (2011). "Coping with correlation," mimeo.

[53] Zak, P. J., and S. Knack (2001). "Trust and growth." *Economic Journal* 111, 295–321.