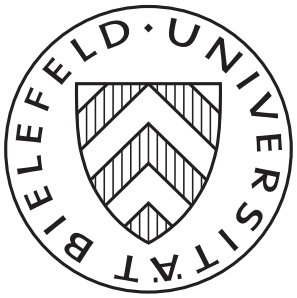
Center for
Mathematical Economics
Working Papers

551

January 2016

Cognitive Empathy in Conflict Situations

Florian Gauer and Christoph Kuzmics



Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
http://www.imw.uni-bielefeld.de/wp/
ISSN: 0931-6558

Cognitive Empathy in Conflict Situations*

Florian Gauer† and Christoph Kuzmics‡

January 12, 2016

Abstract

Two individuals are involved in a conflict situation in which preferences are ex ante uncertain. While they eventually learn their own preferences, they have to pay a small cost if they want to learn their opponent's preferences. We show that, for sufficiently small positive costs of information acquisition, in any Bayesian Nash equilibrium of the resulting game of incomplete information the probability of getting informed about the opponent's preferences is bounded away from zero and one.

Keywords: Incomplete Information, Information Acquisition, Theory of Mind, Conflict, Imperfect Empathy

JEL Codes: C72, C73, D03, D74, D82, D83

*We thank seminar participants at the Universities of Bielefeld, Graz, and Oxford, and at the CERGE-EI in Prague for useful comments and suggestions.

†Center for Mathematical Economics (IMW) and Bielefeld Graduate School of Economics and Management (BiGSEM), Bielefeld University, P.O. Box 100131, D-33501 Bielefeld, Germany. Email: fgauer86@gmail.com, Phone: +49 521 106-4918. This research was carried out within the International Research Training Group "Economic Behavior and Interaction Models" (EBIM) financed by the German Research Foundation (DFG) under contract GRK 1134/2.

‡Institute of Economics, University of Graz, Austria. Email: christoph.kuzmics@uni-graz.at, Phone: +43 316 380-7111.

> If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.
>
> — Sun Tzu, The Art of War, approximately 500BC

1 Introduction

It is probably rare in a conflict situation that we know the exact cardinal preferences of our opponent.[1] Consider, for instance, a penalty kick in soccer. This is as close as one can imagine to a pure conflict (i.e. zero-sum) situation. The kicker wants to score, the goalkeeper wants to prevent that. Imagine then that the goalkeeper incurred, earlier in the game, a slight injury, a bruising on her left side, which might induce her to have a slight additional preference for jumping to the right. The question we are interested in in this paper is whether or not the other player, the kicker, would, at some small cost, like to find out about this slight injury and its consequences for her opponent's preferences.

This cost can be in terms of effort or even money going into the actual acquisition or purchase of this piece of information. Alternatively it could also be in terms of mental costs. If the kicker saw the slight injury, will she reason through its preference consequences for her opponent? The latter interpretation leads us to the term "cognitive empathy" in our title, as defined in psychology as the process of understanding another person's perspective (see e.g. Davis, 1983), which can be traced back to at least Köhler (1929), Piaget (1932), and Mead (1934).[2] Building the possibility of empathy acquisition (or, respectively, information acquisition) into

[1] In the quote from Sun Tzu stated above, it is difficult to know what he meant with "knowing yourself" and "knowing your enemy". The last sentence of the quote, for instance, seems to suggest that it is in fact impossible that two generals know neither themselves nor their enemy, as presumably we cannot have that both "succumb" in the battle between them. What we take from the quote, however, is the idea implicitly suggested there that "knowing your enemy" is something one can acquire.

[2] This is in contrast to "affective empathy" which is defined as a person's emotional response to the emotional state of others (see again Davis, 1983) and the two are not necessarily related. Shamay-Tsoory et al. (2009) find that different areas of the human brain are responsible for "cognitive" and "affective" empathy. Rogers et al. (2007) find that people with Asperger syndrom lack "cognitive" but not "affective" empathy.

a conflict game with incomplete information, we are then interested in the following questions. To which extent do players acquire empathy in equilibrium? How does the possibility of empathy acquisition affect players' action choices in the game? Finally, how do the answers to these questions depend on the value of the cost of empathy acquisition?

To answer these questions we build a simple model. There are two players and two actions for each player. Each player can be one of a finite number of different preference types. The distribution over all preference types is commonly known (to avoid confounding our results with higher-order belief effects). Both players, before learning their own types (this is for convenience), simultaneously decide whether or not to pay a small amount of cost $c \geq 0$ in order to learn the opponent's type. Players do not observe their opponent's choice of empathy acquisition. After learning their own and, if appropriate, their opponent's type, players then choose, as a function of what they know, a (possibly mixed) action. We focus on two-player two-action Bayesian conflict games. These are such that if the types of players were common knowledge then any such complete information "realized type game" must have a unique Nash equilibrium and that Nash equilibrium must be in completely mixed strategies. We investigate Bayesian Nash equilibria of these games.

For such games we show that for sufficiently low positive costs of empathy acquisition the probability of empathy acquisition is strictly bounded away from zero and one in any Bayesian Nash equilibrium of this game.[3] These bounds do not depend on the costs of empathy acquisition beyond the requirement that these costs are sufficiently small. In other words, in any equilibrium of this game, players randomize strictly between acquiring empathy and not acquiring it. It turns out that even if the cost is zero, the game, besides a "full empathy equilibrium", still has such an equilibrium with strict mixing between acquiring empathy and not acquiring it.

There are at least two different interpretations of our model. One, along the lines as suggested above, is such that players are highly rational but have some small costs of reasoning about their respective opponent's preferences. This model could then be about the two individuals engaged

[3]In fact, for a player's equilibrium probability of empathy acquisition to be strictly greater than zero, her opponent must have (at least two) distinct payoff types.

in the penalty kick, but could also be about firms engaging in conflict or indeed, as in the quote by Sun Tzu above, military generals engaged in war.

We prefer to think of this model, however, in its evolutionary interpretation. That is there is mother nature (or evolution) who works on everyone of her subjects independently and has their material interests at her heart. Nature knows that her subjects will be involved in all sorts of conflict situations throughout their life. Nature then decides whether or not she should spend some small amount of fitness cost to endow her subject with cognitive empathy, which would allow her subject to always learn (in fact, to always know) the opponent's preferences. Whether or not the subject has cognitive empathy is not observable by her opponent. Our results then imply that nature (guiding play to Bayesian Nash equilibrium) endows some but not all of her subjects with cognitive empathy even if the costs of doing so are essentially zero.

1.1 Related Literature

This paper is related to the literature on the evolution of preferences for strategic interaction, initiated by the so-called "indirect evolutionary approach" of Güth and Yaari (1992) and Güth (1995). Individuals who are randomly matched to engage in a given form of strategic interaction are first given a utility function by mother nature. Mother nature works on every player separately and aims to maximize this player's material preferences. Players evaluate outcomes of play with the preferences given to them by mother nature. There are two kinds of results in this literature. Assuming that individuals (automatically) observe their opponents' preferences, in many settings non-material preferences arise as mother nature's optimal choice (see e.g. Koçkesen et al., 2000a,b; Heifetz et al., 2007a,b; Dekel et al., 2007; Herold and Kuzmics, 2009). On the other hand, assuming that individuals cannot observe their opponents' preferences, essentially only allows material preferences as mother nature's optimal choice (see e.g. Ely and Yilankaya, 2001; Ok and Vega-Redondo, 2001). This induced Robson and Samuelson (2010) to wish that the potential observability of

preferences is also made subject to evolutionary forces.[4] Some work in that direction has recently been begun by Heller and Mohlin (2015a,b).[5] Our model can be seen as to tackle the question of the evolution of observability of preferences without modelling the evolution of preferences.

Another such model is given in Robalino and Robson (2012, 2015). In their model, individuals are interacting in ever changing environments. An individual with "theory of mind" is able to use past experiences of opponent play to predict more quickly how her opponent will play. Thus, even if it is somewhat costly, there is a strict benefit from having a "theory of mind". One could argue that the incomplete information (about opponents' preferences) in our model is somewhat akin to the ever changing environment in Robalino and Robson (2015). Our model has no explicit learning. One could perhaps argue it is implicit in our use of Bayesian Nash equilibrium. Our example of a non-conflict game provides a similar result as that in Robalino and Robson (2015) in that any Bayesian Nash equilibrium must exhibit "full" cognitive empathy, i.e. with probability one. When we focus on conflict games alone, we find a starkly contrasting result in that any Bayesian Nash equilibrium must exhibit "partial" cognitive empathy, i.e. the probability of acquiring empathy is bounded from below as well as from above, even when costs of acquiring empathy tend to zero.

Aumann and Maschler (1972), provide an example of a complete information bimatrix game, due to John Harsanyi, to discuss the relative normative appeal of maxmin and Nash equilibrium strategies. The game is a two-player two-action game and not quite zero sum with a unique Nash equilibrium which is in completely mixed strategies. In this game, Nash equilibrium strategies and maxmin strategies differ for both players. Yet the expected payoff to a given player in the Nash equilibrium is the same as the expected payoff that this player can guarantee herself by playing her

[4]Similarly (Samuelson, 2001, p. 228) states: "Together, these papers highlight the dependence of indirect evolutionary models on observable preferences, posing a challenge to the indirect evolutionary approach that can be met only by allowing the question of preference observability to be endogenously determined within the model."

[5]The former is a model in which, while individual preferences evolve, so do individuals' abilities to deceive their opponents. The latter asks the question whether cooperation can be a stable outcome of the evolution of preferences in the prisoners' dilemma when players can observe and condition their play on some of their opponent's past actions (in encounters with other people).

maxmin strategy. Pruzhansky (2011) provides a large class of complete information bimatrix games that have this feature. If this is the case, would one not, for this class, recommend players to use their maxmin strategies? In our model, in which players have uncertainty about their opponent's preferences, and therefore in some sense greater uncertainty about their opponent's strategy, one might think that the appeal of maxmin strategies is even greater. Yet, in our model there may be a strict benefit from deviating from maxmin strategies.

The literature on level-k thinking (see e.g. Stahl and Wilson, 1994, 1995; Nagel, 1995; Ho et al., 1998; Costa-Gomes et al., 2001; Crawford, 2003; Costa-Gomes and Crawford, 2006; Crawford and Iriberri, 2007) typically finds that individuals engaged in game theory experiments do not all reason in the same way as individuals seem to have different "theories of mind". In that sense, our paper can be loosely interpreted as a model to understand why there may be individuals of different levels of strategic thinking.

There is a literature on information acquisition in oligopoly models as in e.g. Li et al. (1987), Hwang (1993), Hauk and Hurkens (2001), Dimitrova and Schlee (2003), and Jansen (2008), where firms can acquire information about the uncertain market demand before engaging in oligopoly competition. Market demand enters all agents' profit functions, whereas in our model the information a player might acquire is exclusively about the opponent's preferences. More general models in which players acquire information about an uncertain parameter affecting all players' preferences are given in Hellwig and Veldkamp (2009), Myatt and Wallace (2012), and Amir and Lazzati (2014), as well as in Persico (2000) and Bergemann et al. (2009) in a mechanism design context.

Solan and Yariv (2004) consider a sequential model of two-player two-action interaction in which one player chooses a (possibly mixed) action first, then a second player can buy, at some cost, information about the first player's (realized) action before finally then also choosing an action herself. The second player can also choose the precision of the information purchased. The structure of the game is common knowledge. In particular the first player is fully aware that she might be spied upon. Thus "spying" in their model is about the opponent's already determined action with complete information regarding payoffs, whereas in our model "spying" (or

cognitive empathy as we call it) is about the opponent's preferences.

Closest is perhaps Mengel (2012), who studies a model in which individuals play many games and ex ante do not know which game they are playing. Individuals can partition the set of games in any way they like, with the understanding that any two games in the same partition element cannot be distinguished. The individual can condition her action only on the partition element. Adopting a partition comes at some cost, called reasoning costs, and finer partitions are more costly than coarser ones. One difference between Mengel (2012) and what we do here is, therefore, that in our model players always learn their own payoff type, while in Mengel (2012) individuals do not necessarily even learn their own payoff type. Another difference is in the choice of solution concept, we study Bayesian Nash equilibria while Mengel (2012) studies asymptotically stable strategy profiles under some evolutionary process. Both these differences are probably only superficial. The real difference between the two papers is the class of games they study within their respective models. Our main results deal with the case of conflict games. Mengel (2012) does not explicitly study this class. Therefore, the nature of our results is also different.[6]

The rest of the paper is organized as follows. Section 2 states the model. Section 3 provides the main result, further characterizes equilibrium strategy profiles, and provides a uniqueness result. Finally, Section 4 concludes with a discussion of further properties of equilibria in Bayesian conflict games as well as a discussion of possible variations of the model.

[6]The main results in Mengel (2012) are that strict Nash equilibria, while (evolutionarily) stable if the game is commonly known, can be made unstable under learning across games; that weakly dominated strategies, while unstable if the game is commonly known, can be stable under learning across games; and that, if all games have distinct Nash equilibrium supports, learning across games under small reasoning costs leads to individuals holding the finest partition with probability one. Our paper is silent on all these results as our conflict games do not have strict Nash equilibria, do not have weakly dominated strategies, and are such that all (what we call realized type) games are such that their Nash equilibria all have full support. All our results, thus, add to the results in Mengel (2012). One could probably translate our main result into the language of Mengel (2012) as follows. If having the finest partition in the model of Mengel (2012) is essentially the same as acquiring cognitive empathy in our model, then our result, that in conflict games we expect proper mixing between acquiring empathy and not acquiring it, suggests that, in conflict games, learning across games as in Mengel (2012) would lead to individuals properly mixing between different partitions, including the finest as well as the coarsest.

2 The Model

There are two players $p \in \{B, R\}$, "blue" and "red". Each player p can have one of a finite number n^p of possible (payoff) types $\theta^p \in \Theta^p$. There are commonly known full support probability distributions over types given by $\mu^p : \Theta^p \to (0, 1]$ for both players $p \in \{B, R\}$. Abusing notation slightly we sometimes write μ^{θ^p} instead of $\mu^p(\theta^p)$. The types of the two players are then drawn from the respective distribution statistically independently from each other. Every type of every player has the same finite set of possible actions at her disposal, given by $A = \{a_1, ..., a_m\}$.[7] Payoffs to player $p \in \{B, R\}$ are then given by the utility function $u^{\theta^p} : A \times A \to \mathbb{R}$, where the first argument depicts the action taken by player p and the second the one taken by her opponent $-p$. Note that different types have different utility functions and that utility functions do only depend on the chosen action pair and not directly on the opponent's type.

Before players learn their own type, i.e. at the complete ex-ante stage, each of them can independently and secretly invest a cost of $c \geq 0$ in order to acquire cognitive empathy. This cost is then simply subtracted from the player's payoff. A player who acquires empathy then, at the interim stage, learns not only her own type but also the type of her opponent. These player types are then called *informed*. Note, however, that an informed type is not able to observe her opponent's choice of empathy acquisition. We further assume that there is only *no empathy* or *full empathy*. When we speak of a player having *partial empathy* we mean that this player randomizes between no and full empathy.[8] A player who does not acquire empathy learns, at the interim stage, only her own type. The corresponding player types are then called *uninformed*.

A *strategy* of player $p \in \{B, R\}$ is then given by a pair $\left(\rho^p, (\sigma^{\theta^p})_{\theta^p \in \Theta^p}\right)$ where $\rho^p \in [0, 1]$ is the *probability of empathy (or information) acquisition*,

[7] In principle, one could consider action sets of different cardinality for both players. The paper, however, focuses on what we call conflict games. A crucial feature of conflict games is that its "realized type games" (defined below) have a unique equilibrium and that equilibrium is in completely mixed strategies. One can verify that this implies that the two players must have the same number of actions.

[8] Throughout the paper, partial empathy usually comprises the case that the corresponding player acquires empathy with probability zero while it always excludes empathy acquisition with probability one.

and $\sigma^{\theta^p} : \Theta^{-p} \cup \{\emptyset\} \to \Delta(A)$, the *action strategy*, is the (mixed) action to be played by player p of type $\theta^p \in \Theta^p$ against any opponent of known type $\theta^{-p} \in \Theta^{-p}$, when informed, and of unknown type (which is indicated by the player receiving the uninformative "signal" $\emptyset$), when uninformed.

Our solution concept is Bayesian Nash equilibrium. One interpretation of equilibrium play is that it is the outcome of a long and slow evolutionary process. It is well known (see e.g. Nachbar, 1990) that if any strategy profile is the outcome of a reasonable evolutionary process, it must be an equilibrium. As our main result holds for all equilibria of the game, it is therefore true for all candidates of an evolutionary stable outcome.[9]

The paper almost exclusively focusses on what we call *Bayesian conflict games*.[10] For any pair of types $\theta^B \in \Theta^B$ and $\theta^R \in \Theta^R$ we define the *realized type game* as the complete information game that would result if it were common knowledge among the two players that they are of exactly these two types. The Bayesian game is then a *Bayesian conflict game* if every possible realized type game has a unique Nash equilibrium and if this Nash equilibrium is in completely mixed strategies.[11]

3 Results

We first show that for positive costs of empathy acquisition there cannot be an equilibrium of a conflict game in which both players choose to acquire empathy with probability one.

Proposition 1. *Consider a Bayesian conflict game. If costs of empathy acquisition are positive, then no strategy profile with full empathy, i.e. with* $(\rho^B, \rho^R) = (1, 1)$*, can be a Bayesian Nash equilibrium. On the contrary, if costs are zero, there is such a full empathy equilibrium.*

[9] It is also well known that not all games have evolutionary stable outcomes. There can, for instance, be cycles in behavior. Such cycles then tend to cycle around equilibria (see e.g. Hofbauer and Sigmund, 1998, Chapter 7.6).

[10] Section 4.3 provides an example of a non-conflict game.

[11] In our main theorem and propositions we write "Bayesian conflict game", to ensure that a reader who only browses the paper understands that the conflict games studied in this paper have incomplete information. Everywhere else in the paper we simply write "conflict game" with the understanding that we are nevertheless dealing with a Bayesian conflict game. Analogously, we refer to Bayesian Nash equilibria of Bayesian conflict games simply as equilibria of conflict games.

Proof of Proposition 1. Suppose a conflict game has an equilibrium with $(\rho^B, \rho^R) = (1, 1)$. Then whenever two types $\theta^B \in \Theta^B$ and $\theta^R \in \Theta^R$ meet, it is common knowledge that this is the case and, as this happens with positive probability, they must play a Nash equilibrium of the corresponding realized type game. Any realized type game by definition has a unique Nash equilibrium and this Nash equilibrium is in completely mixed strategies. Thus, every type of every player in every situation is always indifferent between all her pure actions. Hence, when costs are positive, any player would be better off not acquiring empathy, thus saving $c > 0$, and playing any (mixed) action. Arriving at a contradiction, we therefore have the proof for $c > 0$. Observe however that this saving opportunity disappears for $c = 0$ meaning that in this case the above strategy profile is indeed an equilibrium of the conflict game. □

Note that Proposition 1 leaves open the possibility that one (and only one) player acquires empathy with probability one. Turning to a population interpretation of (mixed) equilibrium (as in evolutionary game theory), Proposition 1 states that we expect at least a fraction of the population for at least one player position to not have cognitive empathy in equilibrium.

Suppose we consider symmetric conflict games, such as a Bayesian version of the well known rock-scissors-paper game. Suppose we are interested in the single population evolutionary model. That is, there is one population of individuals from which repeatedly two are randomly drawn to play the game. Then the appropriate solution concept is symmetric Bayesian Nash equilibrium and Proposition 1 then implies that this single population has a fraction of individuals without cognitive empathy.

Proposition 1 leaves open the possibility that, as costs of empathy acquisition tend to zero, the equilibrium probability of empathy acquisition tends to one. To see that this is at least not generally true, we turn to our main result and our analysis of *two-action Bayesian conflict games*, conflict games in which each player has two actions available. In such games one player, termed B (or blue) throughout the paper, always wants to coordinate actions while the other, termed R (or red) throughout the paper, wants to mis-coordinate actions. The Bayesian uncertainty is then only about the intensity of these preferences. One could thus alternatively

describe a two-action conflict game as a non-zero sum version of matching pennies with incomplete information.

The following theorem is the main result of this paper. It establishes that in any equilibrium of a two-action conflict game for any of the two players the probability of empathy acquisition is bounded away from zero (if the considered player's opponent has at least two distinct types) and, even more importantly, bounded away from one for all sufficiently small positive costs. In order to state this theorem we require one additional piece of notation. In a two-action conflict game, for any player $p \in \{B, R\}$ of any type $\theta^p \in \Theta^p$ denote by $x(\theta^p)$ the probability of action H that, if played by the opponent, makes θ^p indifferent between the two actions.[12] One could call $x(\theta^p)$ the *indifference probability* of type θ^p. By assumption we have $x(\theta^p) \in (0,1)$ for all $\theta^p \in \Theta^p$ and $p \in \{B, R\}$. Further, denote by θ^p_{max} (θ^p_{min}) a type with maximal (minimal) indifference probability $x(\theta^p)$.

Theorem 1. *Consider a two-action Bayesian conflict game. There exists $C > 0$ such that for all $p \in \{B, R\}$ we have in any Bayesian Nash equilibrium that*

(i) $\rho^p \geq x(\theta^{-p}_{max}) - x(\theta^{-p}_{min})$ *if* $c \in [0, C)$ *and*

(ii) $\rho^p < \max\left\{x(\theta^{-p}_{max}), 1 - x(\theta^{-p}_{min})\right\}$ *if* $c \in (0, C)$.

The proof of this theorem is delegated to the appendix. The proof rests on two lemmas that are of some independent interest. We now state these lemmas, one after the other, give their respective proof (or a sketch thereof with the full proof in the appendix), and then sketch how they combine with some additional work to establish that equilibrium empathy acquisition probabilities are bounded away from zero and one.

We first show that in equilibrium any uninformed player type must be indifferent between both actions. Just as we do this for the indifference probabilities, we omit the subscript H for ease of notation when considering action strategies in two-action conflict games from here on.

[12] For a player $p \in \{B, R\}$ we call two types $\theta^p_1, \theta^p_2 \in \Theta^p$ *distinct* if $x(\theta^p_1) \neq x(\theta^p_2)$.

Lemma 1. *Consider a two-action Bayesian conflict game. Then there exists $C > 0$ such that for all $c \in [0, C)$, $p \in \{B, R\}$ and $\theta^p \in \Theta^p$*

$$\sum_{\theta^{-p} \in \Theta^{-p}} \mu^{\theta^{-p}} \left(\rho^{-p} \sigma^{\theta^{-p}}(\theta^p) + (1 - \rho^{-p}) \sigma^{\theta^{-p}}(\emptyset) \right) = x(\theta^p) \tag{1}$$

in any Bayesian Nash equilibrium.

Sketch of Proof of Lemma 1. Suppose there is a player, w.l.o.g. blue, of some type that is uninformed and not indifferent between her two actions. Suppose, w.l.o.g. that she prefers action H. As she is uninformed, she is facing a (mixed) action that is a convex combination of all (mixed) actions of all opponent (red player) types. As she prefers H against this mixture, and as she prefers to coordinate actions, this mixture must place a relatively high probability on H. But as this mixture is a convex combination of mixed actions of all red types there must be one red type who also plays H with higher probability. Thus, the same blue player type, when informed and facing that red type, also plays H. But then the red player, the miscoordination player, of this type, when informed and playing against the considered blue type, must play T as she is facing the pure action H. This can then be argued to imply on the one hand that the red player is not acquiring empathy with high probability and on the other hand that she is not playing close to T when of the considered type and uninformed. But then, as costs are small, she should deviate to acquire empathy with probability one and to play T when meeting this given blue type. □

The second result we need is that, for positive costs, in equilibrium for each of the two players there must be a type who, when informed cannot be indifferent between both actions against all opponent types.

Lemma 2. *Consider a two-action Bayesian conflict game.*[13] *If $c > 0$, then for any Bayesian Nash equilibrium and $p \in \{B, R\}$ with $\rho^p > 0$ there must*

[13] The reader may feel that we use an overabundance of different types in the statement of this lemma. This is, unfortunately, necessary. There are three different types for each player, denoted $\bar{\theta}^p$, $\hat{\theta}^p$, and $\tilde{\theta}^p$. It is important to realize that generally it may well be that all three types on each side are different from each other. In the case where there are only two types for one player, then some of these three types naturally must coincide. This additional structure allows us to prove more in such cases. See Proposition 3.

exist $\bar{\theta}^p \in \Theta^p$ *and* $\hat{\theta}^{-p}, \tilde{\theta}^{-p} \in \Theta^{-p}$ *such that*

$$\rho^{-p}\sigma^{\hat{\theta}^{-p}}(\bar{\theta}^p) + (1-\rho^{-p})\sigma^{\hat{\theta}^{-p}}(\emptyset) > x(\bar{\theta}^p), \tag{2a}$$

$$\rho^{-p}\sigma^{\tilde{\theta}^{-p}}(\bar{\theta}^p) + (1-\rho^{-p})\sigma^{\tilde{\theta}^{-p}}(\emptyset) < x(\bar{\theta}^p). \tag{2b}$$

For $p = B$ $\Big(p = R\Big)$ *this induces* $\sigma^{\bar{\theta}^B}(\hat{\theta}^R) = 1$ *and* $\sigma^{\bar{\theta}^B}(\tilde{\theta}^R) = 0$ $\Big(\sigma^{\bar{\theta}^R}(\hat{\theta}^B) = 0$ *and* $\sigma^{\bar{\theta}^R}(\tilde{\theta}^B) = 1\Big)$.

Proof of Lemma 2. This proof is similar to that of Proposition 1. Suppose a player $p \in \{B, R\}$ does acquire empathy with some positive probability $\rho^p > 0$ in equilibrium while costs are positive, i.e. $c > 0$. Now assume that every type θ^p of player p, when informed, is indifferent between the two actions H and T against any opponent type θ^{-p}. Then player p could benefit strictly from deviating to acquiring empathy with probability zero (thus, saving costs $c > 0$ with probability $\rho^p > 0$) and playing any (mixed) action (not losing anything because of the complete indifference). Arriving at a contradiction, we therefore have that there must be at least one player type $\bar{\theta}^p$ strictly preferring H or T against some opponent type here. Together with Lemma 1 this concludes the proof. □

For the special case that the opponent $-p$ only has one possible type, Lemma 2 has the obvious implication that for any positive cost of empathy acquisition player p does not acquire empathy in any equilibrium.

Consider first Part (i) of Theorem 1, which states that there is a specific lower bound on the equilibrium probabilities of empathy acquisition.

Sketch of Proof of Theorem 1(i). The key to this part is Lemma 1. It states that every type of any player, when uninformed, must be indifferent between both actions as long as costs are sufficiently small. Consider, w.l.o.g., the red player and assume that $x(\theta^R_{max}) > x(\theta^R_{min})$ (otherwise the lower bound is trivially satisfied). Now both player types θ^R_{max} and θ^R_{min} must be indifferent between both actions when uninformed. These two red types, however, face the same distribution over actions if the blue player's probability of empathy acquisition is zero. Why? If the blue player did not acquire empathy, she cannot recognize the red player's type and cannot condition her action strategy on that information. On the other hand, the two (extreme) red types cannot be both indifferent between the two actions

if they are facing the same distribution. Thus, arriving at a contradiction, it must be that the blue player acquires empathy with positive probability. In fact the exact lower bound can be obtained by taking the difference between Equation(s) (1) for the extreme types θ^R_{max} and θ^R_{min}. □

The key statement in Part (ii) of the theorem is that it establishes an upper bound, strictly below one, for each player's equilibrium probability of empathy acquisition. What this upper bound is, is less important. In the appendix we, in fact, prove two results that imply the existence of an upper bound strictly below one. One is as stated in Theorem 1(ii), the other is stated in the appendix as Theorem 1(ii)'. The respective statements are similar but neither implies the other. The former is more elegant in its expression, the latter is more intuitive in its proof. Therefore, we choose to present the sketch of proof for Theorem 1(ii)' here.

Sketch of Proof of Theorem 1(ii)'. W.l.o.g. we focus on the blue player (the coordination player). Assume that the blue player acquires empathy with a probability greater than the stated bound and close to one. By Lemma 2 we know that, in equilibrium, there must be a type of blue player who, when informed, strictly prefers to play H against some type of red player. This type of red player, when informed herself and meeting the given type of blue player, then faces with high likelihood an informed blue player who plays H. Her best response then (being the mis-coordination player) is to play T against this blue type. But as the informed blue type's equilibrium action against this red type is H, two things must be true about the red player. First, she cannot be too informed, i.e. her probability of acquiring empathy must be low, and second, when she is of the considered type and uninformed, she must play H with a high probability. But then, as the cost of empathy acquisition is small, the red player could strictly benefit from deviating to acquire empathy and then, when she is of this type, play T against this blue type. We thus arrive at a contradiction. □

While we do not know whether there are equilibria in general two-action conflict games in which a player's probability of empathy acquisition is close to the upper bound stated in Theorem 1, for small costs, there is always an equilibrium in which the lower bound is achieved.

Proposition 2. *For every two-action Bayesian conflict game there exists $C > 0$ such that for all $c \in [0, C)$ it has a Bayesian Nash equilibrium with $\rho^p = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min})$ for both $p \in \{B, R\}$.*

The proof of this proposition is given in the appendix. The key to reaching the lower bound for the probability of empathy acquisition is to let all informed types of player $p \in \{B, R\}$ play H against opponent type θ^{-p}_{max} and T against type θ^{-p}_{min}. Taking into account Lemma 1, this immediately pins down the equilibrium probability of empathy acquisition of player p and it is exactly the lower bound. The equilibrium is then further constructed by letting uninformed types mix in a way that makes the opponent $-p$ indifferent between acquiring empathy and not doing so.

With more than two types for one player and at least two for the other, this kind of equilibrium can be constructed in different ways. In general, this gives rise to a continuum of equilibria that differ in terms of players' action strategies but not in terms of their information strategies. In what follows, any representative of this class is called a *partial empathy equilibrium*. Note, however, that a player acquires empathy with probability zero in such an equilibrium if her opponent does not have distinct types.

We do not know whether general two-action conflict games with positive costs of empathy acquisition can actually have equilibria in which a player's probability of empathy acquisition is strictly greater than this lower bound. For any such game with either two types for both players or a single type for one player we can show, however, that the partial empathy equilibrium considered in the proof of Proposition 2 is indeed the only equilibrium.

Proposition 3. *For every two-action Bayesian conflict game with only one type for one player and more than one type for the other player or with exactly two types for both players there exists $C > 0$ such that for all $c \in (0, C)$ and $p \in \{B, R\}$ the probability of empathy acquisition is $\rho^p = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min})$ in any Bayesian Nash equilibrium. In these cases the Bayesian Nash equilibrium is uniquely determined by the strategy profile considered in the proof of Proposition 2 if there is a unique maximal type θ^p_{max} and a unique minimal type θ^p_{min} for both players $p \in \{B, R\}$.*[14]

[14] In the case in which one player has a single type the equilibrium is in fact unique only in terms of action strategies played with positive probability in equilibrium.

The proof is again given in the appendix. The key for this proposition is to realize that with a limited number of types for at least one player we can pin down behavior of all types of this player fairly quickly with the help of Theorem 1. Things turn out to be much more complex if both players have many types, as then all we know is, for instance, that there is one type of player red who plays H against some type of player blue, but we do not know which types these are. If there are only two types on both sides, for instance, then by starting with one type who does something specific against one opponent type all other types' behavior follows.

4 Discussion and Conclusion

In this paper we study two-player conflict situations with ex-ante uncertainty over (the exact) opponent preferences for both players. We allow players, before learning their own payoff type, to acquire cognitive empathy at some (small) cost. Cognitive empathy enables a player to learn the preferences of her opponent in all situations. There are at least two ways we can interpret this model. The first interpretation is that there are indeed two strategic opponents (the two soccer players from the introduction, two firms, two military generals, etc.) who are involved in a conflict situation and who can acquire information about their opponent's ex-ante unknown preferences. Given this interpretation, we find that in equilibrium these strategic players do not fully acquire information about their opponent's preferences, even if the cost of doing so is vanishingly small. A second interpretation is that there are many individuals who are often and randomly engaged in pairwise conflict situations and mother nature can endow these individuals (each individual separately) with cognitive empathy, i.e. with the ability to understand opponents' preferences, at some positive cost (e.g. by providing an additional brain function). Under the assumption that nature then guides play to an evolutionary stable state, which must be a Bayesian Nash equilibrium of this game, our results imply that nature endows some but not all of her subjects with cognitive empathy, even if the costs of doing so are essentially zero.

Our model is simple and sparse and many alterations and additions are conceivable. In what follows we discuss additional consequences of our

results as well as some possible modifications of our model.

4.1 Empathy Acquisition at Zero Costs

In this subsection we provide a corollary to (the proof of) Proposition 2 for the special case of zero costs of empathy acquisition that allows us to provide additional intuition for our main result.

Corollary 1. *Any two-action Bayesian conflict game with $c = 0$ has a Bayesian Nash equilibrium with partial empathy, i.e. $\rho^p \in [0, 1)$, and*

$$\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1 - \rho^p)\sigma^{\theta^p}(\emptyset) = x(\theta^{-p})$$

for all $p \in \{B, R\}, \theta^p \in \Theta^p, \theta^{-p} \in \Theta^{-p}$.

This corollary implies that an outside observer who can observe the two players' types would observe, for any pair of types, a frequency of actions exactly as given by the Nash equilibrium of the realized type game (in which that game is common knowledge) and, therefore, the same frequency of actions as in the full empathy equilibrium. In other words, even though, when the two types meet, players are far from having common knowledge that the two are of these particular types, they nevertheless, on average, manage to play "as if" they had common knowledge of this fact.[15] If an outside observer, however, were to be able to track individuals behavior throughout many interactions, then this observer could distinguish, by means of the observed frequency of behavior, those individuals that are empathic from those that are not. This observer could then also distinguish whether the partial or full empathy equilibrium is played.

4.2 Equilibrium Payoffs

In this subsection we turn to a discussion of equilibrium payoffs in two-action conflict games with the (costly) possibility of empathy acquisition. Note that, when we talk about the payoff to an (informed) type, we mean the payoff without taking into account the costs of empathy acquisition

[15] This insight could be useful if one were to attempt to generalize our result to more than two actions.

that players have to bear. In contrast, these costs are included when we consider players' ex-ante expected payoffs. In what follows, the full empathy equilibrium under zero costs, which we know from Proposition 1, is referred to as the benchmark case. Consider first the case of small costs.

Corollary 2. *Consider a two-action Bayesian conflict game. There exists* $C > 0$ *such that for all* $c \in (0, C)$ *in any Bayesian Nash equilibrium*

(i) every player obtains an ex-ante expected payoff equal to her ex-ante expected payoff in the benchmark case,

(ii) every uninformed type for each player obtains the same expected payoff as she does in the benchmark case, and

(iii) for each player acquiring empathy with positive probability there is at least one type who, when informed, obtains a strictly higher expected payoff than she obtains in the benchmark case.

This set of statements is a corollary to Lemma 1, Theorem 1, and Proposition 1. Part (i) follows from the fact that all types, when uninformed, are indifferent between both actions in any equilibrium by Lemma 1 and all players acquiring empathy with positive probability are ex-ante indifferent between acquiring empathy and not doing so by Theorem 1(ii). Part (ii) follows from Lemma 1 alone. Part (iii) follows from Part (i) and the fact that players have to bear a cost of $c > 0$ to acquire empathy.

This corollary states that there is a sense in which in two-action conflict games, for all sufficiently low cost levels, all equilibria are ex ante payoff equivalent (if we consider payoffs net of costs). As costs are positive, this implies that some types of players must, when informed, expect higher payoffs than they expect when they are uninformed. One can show by example that this need not be the case for all types and may even be true for only one type.

Before we turn to the case of large costs, it is fruitful to partition the class of conflict games into two subclasses. These are inspired by Pruzhansky (2011). For every type $\theta^p \in \Theta^p$ of a player $p \in \{B, R\}$ define the *type-induced zero-sum game* as the complete information game in which player p has preferences given by her type, i.e. given by u^{θ^p}, and her

opponent has preferences $-u^{\theta^p}$. We call a type *immunizable* (or *robustly immunizable*) if the type induced zero-sum game has no strictly dominated mixed action strategy (or no weakly dominated mixed action strategy) for both players.[16] We call a conflict game *immunizable* if every type of every player is immunizable. If in a conflict game there is at least one non-immunizable type, then this game is called *non-immunizable*.

We choose the label "immunizable" because of a result due to Pruzhansky (2011, p. 355). He shows that in any complete information game with two immunizable players both players have "equalizer" strategies. If a player adopts an "equalizer" strategy, she gets the same expected payoff regardless of the action taken by the opponent. He then shows in his Lemma 1, that in any complete information game with immunizable players on both sides every equalizer strategy is a maxmin strategy. Moreover, he then shows in his Lemma 2 that in such games equalizer strategies guarantee the player the Nash equilibrium payoff. This generalizes the insight found by Aumann and Maschler (1972) in their example.

Remark 1. *Consider a two-action Bayesian conflict game with costs of empathy acquisition so high that any strategy including empathy acquisition is dominated by one without empathy acquisition. If this game is immunizable, then in any Bayesian Nash equilibrium*

(i) every player of every (necessarily uninformed) type obtains an expected payoff that is at least as large as in the benchmark case, and

(ii) every player $p \in \{B, R\}$ with at least two robustly immunizable types $\theta_1^p, \theta_2^p \in \Theta^p$ with $x(\theta_1^p) \neq x(\theta_2^p)$ obtains an ex-ante expected payoff strictly larger than her ex-ante expected payoff in the benchmark case.

If this game is non-immunizable, then in any Bayesian Nash equilibrium

(iii) every player of every (necessarily uninformed) non-immunizable type obtains an expected payoff that strictly exceeds her maxmin payoff.

In a Bayesian Nash equilibrium of a non-immunizable game there can be

[16] Note that in a conflict game no type has a dominated action strategy. A type in a conflict game is therefore (robustly) immunizable if her fictitious zero-sum opponent in the type-induced zero-sum game has no strictly (weakly) dominated strategy.

(iv) (necessarily uninformed) types of a player who obtain an expected payoff that is strictly larger, resp. strictly lower, than her expected payoff in the benchmark case, and even

(v) players with an ex-ante expected payoff that is strictly larger, resp. strictly lower, than her ex-ante expected payoff in the benchmark case.

Part (i) of the remark follows from the result of Pruzhansky (2011) that in such games any type of any player's maxmin payoff is equal to her Nash equilibrium payoff in any realized type game. The latter payoff is the payoff this type of player obtains in the benchmark case. As she can always guarantee herself this payoff by playing her maxmin action strategy, she can certainly never receive less in any equilibrium for any cost level. Moreover, as players are uninformed here, each type faces the same average opponent action strategy. Under the additional assumption of Part (ii) this means that in any equilibrium at least one of the two robustly immunizable types must have incentives to play a pure action strategy which makes her strictly better off than in the benchmark case. This, together with Part (i), then proves Part (ii). Part (iii) of the remark follows from the observation that in a two-action conflict game, to prevent a non-immunizable player type from obtaining strictly more than the maxmin payoff, the opponent needs to play a pure action. However, one can show that in any equilibrium of such a game the opponent, on average, does not use a pure action strategy. Therefore every such player type must receive a payoff that is strictly larger than her maxmin payoff. Finally, to see Parts (iv) and (v) of the remark, consider the following example.

Example 1. *Consider the two-action Bayesian conflict game with action set $A = \{H, T\}$, type sets $\Theta^B = \{\theta_1^B, \theta_2^B\}$ and $\Theta^R = \{\theta_1^R, \theta_2^R\}$, probability distributions over types $\mu^B = \mu^R = \left(\frac{1}{2}, \frac{1}{2}\right)$, and payoffs given in Figure 1 with $a, b \in \mathbb{R}$ (where player B chooses rows and R chooses columns).*

In this example, types θ_1^B and θ_1^R are (robustly) immunizable, while types θ_2^B and θ_2^R are non-immunizable. The indifference probabilities are given by $x(\theta_1^B) = \frac{1}{2}$, $x(\theta_2^B) = \frac{2}{3}$, $x(\theta_1^R) = \frac{1}{2}$, and $x(\theta_2^R) = \frac{4}{5}$.

One can verify that the following is an equilibrium of this game under large costs. Obviously, we need to have $\rho^B = \rho^R = 0$, i.e. no empathy is acquired. Furthermore, let $\sigma^{\theta_1^B}(\emptyset) = \sigma^{\theta_2^R}(\emptyset) = 1$ and $\sigma^{\theta_2^B}(\emptyset) = \sigma^{\theta_1^R}(\emptyset) = 0$.

$u^{\theta_1^B}$:

	H	T
H	1	-1
T	-1	1

$u^{\theta_1^R}$:

	H	T
H	-1	1
T	1	-1

$u^{\theta_2^B}$:

	H	T
H	3	-1
T	2	1

$u^{\theta_2^R}$:

	H	T
H	-2	$-\frac{3}{2}$
T	1	-1

Figure 1: Payoffs of the conflict game in Example 1.

One can then verify that type θ_2^B receives an equilibrium payoff of $\frac{3}{2}$ while in any realized type game her payoff in the unique Nash equilibrium would be $\frac{5}{3}$. Her payoff in the considered equilibrium of the conflict game under large costs is thus strictly lower than her payoff in the benchmark case. On the other hand, type θ_2^R receives an equilibrium payoff of $-\frac{1}{2}$ which is strictly larger than her payoff of $-\frac{7}{5}$ which she obtains in any realized type game and, thus, in the benchmark case. As in the considered equilibrium all other types expect the same payoff (of zero) as in the benchmark case, player B receives a lower ex-ante expected payoff here than in the benchmark case, while for player R the opposite is true.

4.3 A Non-Conflict Example

In this subsection we provide, as a point of contrast to our main results, a non-conflict example.

Example 2. *Consider a symmetric setup in which both players $p \in \{B, R\}$ can have one of three types $\Theta^B = \Theta^R = \{\theta_1, \theta_2, \theta_3\}$ chosen uniformly (i.e. $\mu^\theta = \frac{1}{3}$ for all $\theta \in \Theta^p$) for the two players. Both players can choose between two actions H and T. Type θ_1 finds action H strictly dominant, type θ_3 finds action T strictly dominant, and type θ_2 has pure coordination preferences. These payoffs, in matrix form, are given in Figure 2.*

u^{θ_1}:

	H	T
H	1	1
T	0	0

u^{θ_2}:

	H	T
H	1	0
T	0	1

u^{θ_3}:

	H	T
H	0	0
T	1	1

Figure 2: Payoffs of the non-conflict game in Example 2

For costs of empathy acquisition sufficiently low ($c < \frac{1}{9}$) this game has no equilibrium in which a player acquires empathy with probability less than one. Suppose a player (say blue) attaches positive probability to not acquiring empathy. Red makes her choice of action dependent on her own type with dominant action types playing their dominant actions. Now consider the uninformed coordination type of blue. The best she can do is to play a best response to the given (mixed) action of the coordination type of red. W.l.o.g. let this best response action be H. The uninformed coordination type of blue then receives a payoff of zero against the red type with dominant action T. For blue switching to acquiring empathy with probability one and playing T against the T dominant action type of red is then beneficial if $c < \frac{1}{9}$.[17]

4.4 The Timing of Decisions

Given the evolutionary interpretation of our model in which nature's subjects play many conflict games with often different preferences and opponents throughout their life, it seems appropriate that nature makes the decision about empathy acquisition at the very beginning. Also for the other interpretation, in which players are consciously strategic about their choice of information acquisition, it can make sense to have the information acquisition decision before knowing the exact nature of the conflict situation. A soccer team may study the opposing goalkeeper for the eventuality of a penalty kick before knowing which of their own players will actually take the penalty kick. A military general might want to spy on her opponent's strengths (and thus preferences) before knowing the future strength of her own troop or on which terrain, in which place, at which state of the war etc. the actual battle will take place. One could imagine a firm to acquire information about another firm's cost structure before knowing the exact demand function in markets the two firms compete in.

Yet, there are certainly cases, in which the reverse timing, in which players consider acquiring the information about their opponent's preferences

[17] Suppose individuals choose whether or not to acquire empathy after they learn their own type. The two dominant action types do not acquire empathy now, but for c small enough (now $c < \frac{1}{3}$) coordination types acquire empathy with probability one in any equilibrium of this game.

only after they know their own preferences, is more plausible.

We do not believe that reversing the timing in our model would change the main messages of our paper. We did not go through this model thoroughly, but only looked at two examples. First, as mentioned in Footnote 17, for the given non-conflict example (see Example 2), the main insight there does not change even if we reverse the timing. We have also looked at the reverse-timing model for a two-action two-type conflict example. While small details change, the main result, that for small positive costs of empathy acquisition any equilibrium has partial cognitive empathy, seems to remain unchanged.[18] In fact, in the equilibrium in this example, all types acquire partial empathy: the probability of empathy acquisition is, as in our main result, bounded from below and above.

4.5 Degrees of Cognitive Empathy

Another issue, especially for the evolutionary interpretation of our model, is this. If nature has to make her decision on cognitive empathy at the beginning once and for all possible situations, then these "all possible situations" should probably cover more than just conflict games. And, if these situations include, for instance, the three possible types (for both players) as given in our non-conflict example, then for small costs nature would always endow her subjects with full empathy. One could now state that it is then a question of which is smaller, the cost of empathy acquisition or the probability of these three types, but this is not where we want to go in this discussion. Instead, we think that a better model in such cases would be one in which nature can give her subjects degrees of empathy. For instance, nature could give us enough cognitive empathy to always check whether or not our opponent has a dominant action strategy, but if our opponent does not, nature may not give us more cognitive empathy to differentiate our opponent's preferences further. The result would then be as in our model.

A related consideration is the following. Consider, for convenience, two-action conflict games with two types per player. For these games Proposition 3 implies that a player's probability of empathy acquisition is exactly

[18] To be precise, we used one such example and Gambit by McKelvey et al. (2014) and found exactly one equilibrium. We have not attempted to prove that this equilibrium is unique but we conjecture that it is.

given by the difference of the two indifference probabilities of her two opponent types. This means that the more similar her two opponent types are, the more similar are their indifference probabilities and the less empathy is acquired by her in equilibrium. This is also true for the lower bound established for the probability of empathy acquisition in our main theorem. In particular, this implies that the more different kinds of situations a person faces, i.e. the bigger the possible difference between the possible opponent types, the more empathy is acquired. If this goes as far as to include even dominant strategy types, she will acquire full empathy in equilibrium.

But if we insist in considering a large set of possible situations, we believe our model of full or no empathy acquisition is too simple. A more appropriate model in this case would be one of "rational (in)attention" as in the decision theoretic models of Sims (2003, 2006); Matêjka and McKay (2012, 2015). Adapting these models to our strategic interaction setting could be done by allowing players to buy signals about their opponent's preferences of any precision with costs increasing in the information content of these signals. Another model would be to allow individuals to acquire multiple signals of whatever precision, one after the other, about their opponent's preferences, before making their final action decision. While we do not think that the main insight of our paper would change in such a model, especially of the latter variety, such a model might nevertheless add substantial additional insights.

A Proofs

Throughout this section we again abuse notation of action strategies in two-action conflict games slightly by denoting by $\sigma^{\theta^p}(\cdot) \in [0,1]$ the probability of H chosen by player p of type θ^p. For ease of notation, when it comes to the arguments of utility functions u^{θ^p}, we also only mention the probabilities of action H. And finally, let $\mathcal{U}^{\theta^p}_{Info}$ denote the (interim) expected payoff of a type $\theta^p \in \Theta^p$ of player p if she acquired empathy and before she learns her opponent's type. Similarly, $\mathcal{U}^{\theta^p}_{N}$ denotes the expected payoff of a type θ^p of player p who did not acquire empathy.

A.1 Proof of Lemma 1

For $p \in \{B, R\}$, $\theta^p \in \Theta^p$ we define

$$\begin{aligned}
C^{B_H}(\theta^B, \theta^R) &:= u^{\theta^R}(0,1) - u^{\theta^R}(x(\theta^B), 1),\\
C^{B_T}(\theta^B, \theta^R) &:= u^{\theta^R}(1,0) - u^{\theta^R}(x(\theta^B), 0),\\
C^{R_H}(\theta^B, \theta^R) &:= u^{\theta^B}(1,1) - u^{\theta^B}(x(\theta^R), 1), \text{ and}\\
C^{R_T}(\theta^B, \theta^R) &:= u^{\theta^B}(0,0) - u^{\theta^B}(x(\theta^R), 0).
\end{aligned}$$

Notice that $C^{B\cdot}(\theta^B, \theta^R) > 0$ $\left(C^{R\cdot}(\theta^B, \theta^R) > 0\right)$ for all $\theta^B \in \Theta^B, \theta^R \in \Theta^R$ as player R wants to mis-coordinate (as player B wants to coordinate). Let

$$C := \min_{a \in \{B_H, B_T, R_H, R_T\}} \min_{\theta^B, \theta^R} \mu^{\theta^B} \mu^{\theta^R} C^a(\theta^B, \theta^R).$$

W.l.o.g. we consider player $p = B$ and assume that we have

$$\sum_{\theta^R} \mu^{\theta^R} \left(\rho^R \sigma^{\theta^R}(\bar{\theta}^B) + (1 - \rho^R)\sigma^{\theta^R}(\emptyset)\right) > x(\bar{\theta}^B)$$

for some $\bar{\theta}^B \in \Theta^B$.[19] Since player B wants to coordinate actions, this implies $\sigma^{\bar{\theta}^B}(\emptyset) = 1$ (if $\rho^B < 1$). Furthermore, if a probability weighted sum of terms exceeds $x(\bar{\theta}^B)$ then at least one term must exceed $x(\bar{\theta}^B)$ as well. Thus, there must exist a type $\bar{\theta}^R$ such that

$$\rho^R \sigma^{\bar{\theta}^R}(\bar{\theta}^B) + (1 - \rho^R)\sigma^{\bar{\theta}^R}(\emptyset) > x(\bar{\theta}^B). \tag{3}$$

In turn, this implies $\sigma^{\bar{\theta}^B}(\bar{\theta}^R) = 1$ (if $\rho^B > 0$), meaning that

$$\rho^B \sigma^{\bar{\theta}^B}(\bar{\theta}^R) + (1 - \rho^B)\sigma^{\bar{\theta}^B}(\emptyset) = 1 > x(\bar{\theta}^R).$$

Moreover, it is obvious that this equality and inequality also hold for $\rho^B = 0$ and $\rho^B = 1$. As player R wants to mis-coordinate, this implies $\sigma^{\bar{\theta}^R}(\bar{\theta}^B) = 0$ (if $\rho^R > 0$). Inserting the latter into Inequality (3) gives $(1 - \rho^R)\sigma^{\bar{\theta}^R}(\emptyset) > x(\bar{\theta}^B)$. Again, it is obvious that this inequality is satisfied for $\rho^R = 0$ as well. It follows from this that $1 - \rho^R > x(\bar{\theta}^B)$ and $\sigma^{\bar{\theta}^R}(\emptyset) > x(\bar{\theta}^B)$. Hence,

[19] Observe that the subsequent line of argument is almost identical for the reversed inequality as well as for $p = R$. Thus, we can omit these cases.

for $c \in [0, C)$ player R can improve her payoff by deviating to a strategy with $\hat{\rho}^R = 1$ and obtaining an additional payoff of at least

$$\begin{aligned}&(1-\rho^R)\Big(\mu^{\bar{\theta}^R}\mu^{\bar{\theta}^B}\Big(u^{\bar{\theta}^R}(0,1) - u^{\bar{\theta}^R}(\sigma^{\bar{\theta}^R}(\emptyset),1)\Big) - c\Big)\\ >&x(\bar{\theta}^B)\Big(\mu^{\bar{\theta}^R}\mu^{\bar{\theta}^B}\Big(u^{\bar{\theta}^R}(0,1) - u^{\bar{\theta}^R}(x(\bar{\theta}^B),1)\Big) - c\Big) > 0.\end{aligned}$$

We thus arrive at a contradiction.

A.2 Proof of Theorem 1

An additional technical lemma is needed in order to prove Theorem 1.

Lemma 3. *Consider $\alpha, \beta', \beta'', \gamma \in \mathbb{R}$ where $\beta' - \beta'' \leq \alpha$. Then (at least) one of the following three conditions must be satisfied:*

$$\alpha + (1-\alpha)\gamma = \beta' \text{ and } (1-\alpha)\gamma = \beta'', \tag{4a}$$

$$\alpha + (1-\alpha)\gamma > \beta' \text{ or} \tag{4b}$$

$$(1-\alpha)\gamma < \beta''. \tag{4c}$$

Proof of Lemma 3. Suppose none of the three conditions is satisfied. Then

$$\alpha + (1-\alpha)\gamma < (\leq)\beta' \text{ and } (1-\alpha)\gamma \geq (>)\beta''.$$

In either case, subtracting the second from the first inequality yields $\alpha < \beta' - \beta'' \leq \alpha$, a contradiction. □

We can now turn to the proof of Theorem 1.

Part (i): Lower Bound
Lemma 1 and Equation (1) imply for $p \in \{B, R\}$

$$\rho^p\Big(\underbrace{\sum_{\theta^p}\mu^{\theta^p}\sigma^{\theta^p}(\theta^{-p}_{max})}_{\leq 1} - \underbrace{\sum_{\theta^p}\mu^{\theta^p}\sigma^{\theta^p}(\theta^{-p}_{min})}_{\geq 0}\Big) = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min}).$$

Hence, we have that $\rho^p \geq x(\theta^{-p}_{max}) - x(\theta^{-p}_{min})$.

Part (ii): Upper Bound

Consider player p. If $\rho^p = 0$ then the statement is trivially satisfied. Thus, suppose that $\rho^p > 0$. We need to distinguish two different cases.

Case 1: $\rho^{-p} = 0$

Given the lower bound we proved in Part (i), we then must have that $x(\theta^p_{max}) = x(\theta^p_{min})$. Lemma 2 then implies that there are two opponent types $\hat{\theta}^{-p}$ and $\tilde{\theta}^{-p}$ such that $\sigma^{\hat{\theta}^{-p}}(\emptyset) > x(\theta^p)$ and $\sigma^{\tilde{\theta}^{-p}}(\emptyset) < x(\theta^p)$ for all θ^p. For $p = B$ $\Big(p = R\Big)$ this induces $\sigma^{\theta^p}(\hat{\theta}^{-p}) = 1$ and $\sigma^{\theta^p}(\tilde{\theta}^{-p}) = 0$ $\Big(\sigma^{\theta^p}(\hat{\theta}^{-p}) = 0$ and $\sigma^{\theta^p}(\tilde{\theta}^{-p}) = 1\Big)$ for all $\theta^p \in \Theta^p$. Applying Lemma 1 yields

$$\rho^p \Big(\underbrace{\sum_{\theta^p} \mu^{\theta^p} \sigma^{\theta^p}(\hat{\theta}^{-p})}_{=1\ (=0)} - \underbrace{\sum_{\theta^p} \mu^{\theta^p} \sigma^{\theta^p}(\tilde{\theta}^{-p})}_{=0\ (=1)} \Big) = x(\hat{\theta}^{-p}) - x(\tilde{\theta}^{-p}).$$

Taking into account Part (i) this gives

$$\begin{aligned} \rho^p = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min}) &< \min\Big\{x(\theta^{-p}_{max}), 1 - x(\theta^{-p}_{min})\Big\} \\ &\leq \max\Big\{x(\theta^{-p}_{max}), 1 - x(\theta^{-p}_{min})\Big\}. \end{aligned} \tag{5}$$

Case 2: $\rho^{-p} > 0$

The reasoning is very similar for both players and w.l.o.g. we consider the case $p = B$. Again, Lemma 2 implies that there is a type $\bar{\theta}^B$ and that there are two opponent types $\hat{\theta}^R$ and $\tilde{\theta}^R$ such that

$$\begin{aligned} \alpha + (1-\alpha)\gamma &= \rho^B \sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset), \\ (1-\alpha)\gamma &= \rho^B \sigma^{\bar{\theta}^B}(\tilde{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) \end{aligned}$$

with $\alpha = \rho^B$ and $\gamma = \sigma^{\bar{\theta}^B}(\emptyset)$. As we have already seen that $x(\theta^R_{max}) - x(\theta^R_{min})$ is a lower bound for ρ^B, according to Lemma 3 one of the following three subcases must apply:

Subcase 2(a): $\rho^B \sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) = x(\hat{\theta}^R)$ and
$\rho^B \sigma^{\bar{\theta}^B}(\tilde{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) = x(\tilde{\theta}^R)$

This subcase is straightforward. We simply have

$$\rho^B = x(\hat{\theta}^R) - x(\tilde{\theta}^R) \leq x(\theta^R_{max}) - x(\theta^R_{min}) < \max\Big\{x(\theta^R_{max}), 1 - x(\theta^R_{min})\Big\}.$$

Subcase 2(b): $\rho^B \sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) > x(\hat{\theta}^R)$
This subcase implies that $\sigma^{\hat{\theta}^R}(\bar{\theta}^B) = 0$. Moreover, by Lemma 1 there must exist $\breve{\theta}^B \neq \bar{\theta}^B$ such that

$$\rho^B \sigma^{\breve{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\breve{\theta}^B}(\emptyset) < x(\hat{\theta}^R). \tag{6}$$

This induces $\sigma^{\hat{\theta}^R}(\breve{\theta}^B) = 1$. Furthermore, due to Inequality (2a) we have

$$x(\bar{\theta}^B) < \rho^R \underbrace{\sigma^{\hat{\theta}^R}(\bar{\theta}^B)}_{=0} + (1-\rho^R)\sigma^{\hat{\theta}^R}(\emptyset) = (1-\rho^R)\sigma^{\hat{\theta}^R}(\emptyset).$$

Applying Lemma 3 again – here with $\alpha = \rho^R$, $\beta' = x(\breve{\theta}^B)$, $\beta'' = x(\bar{\theta}^B)$, and $\gamma = \sigma^{\hat{\theta}^R}(\emptyset)$ – then gives

$$x(\breve{\theta}^B) < \rho^R + (1-\rho^R)\sigma^{\hat{\theta}^R}(\emptyset) = \rho^R \sigma^{\hat{\theta}^R}(\breve{\theta}^B) + (1-\rho^R)\sigma^{\hat{\theta}^R}(\emptyset). \tag{7}$$

This induces $\sigma^{\breve{\theta}^B}(\hat{\theta}^R) = 1$, which put into Inequality (6), yields $\rho^B + (1-\rho^B)\sigma^{\breve{\theta}^B}(\emptyset) < x(\hat{\theta}^R)$ and $\rho^B < x(\hat{\theta}^R) \leq \max\left\{x(\theta^R_{max}), 1 - x(\theta^R_{min})\right\}$.

Subcase 2(c): $\rho^B \sigma^{\bar{\theta}^B}(\tilde{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) < x(\tilde{\theta}^R)$
This subcase proceeds analogously to the previous one and is omitted.

A.3 An Alternative to Theorem 1

Theorem 1(ii)'. *Consider a two-action Bayesian conflict game. For all $\epsilon > 0$ there exists $C > 0$ such that for all $p \in \{B, R\}$ and $c \in (0, C)$ we have $\rho^p < \min\left\{x(\theta^{-p}_{max}), 1 - x(\theta^{-p}_{min})\right\} + \epsilon$ in any Bayesian Nash equilibrium.*

Proof of Theorem 1(ii)'. From Inequality (5) in Case 1 of the proof of Part (ii) of Theorem 1 we already know for $p \in \{B, R\}$ that $\rho^p < \min\left\{x(\theta^{-p}_{max}), 1 - x(\theta^{-p}_{min})\right\}$ in any equilibrium with $\rho^{-p} = 0$. Therefore, we only need to consider the case $\rho^{-p} > 0$.

For $p \in \{B, R\}$, $\theta^p \in \Theta^p$ we define

$$\begin{aligned}
C^{B_H}(\epsilon, \theta^B, \theta^R) :=& u^{\theta^R}(0, x(\theta^R) + \epsilon) - u^{\theta^R}(x(\theta^B), x(\theta^R) + \epsilon),\\
C^{B_T}(\epsilon, \theta^B, \theta^R) :=& u^{\theta^R}(1, x(\theta^R) - \epsilon) - u^{\theta^R}(x(\theta^B), x(\theta^R) - \epsilon),\\
C^{R_H}(\epsilon, \theta^B, \theta^R) :=& u^{\theta^B}(1, x(\theta^B) + \epsilon) - u^{\theta^B}(x(\theta^R), x(\theta^B) + \epsilon), \text{ and}
\end{aligned}$$

$$C^{R_T}(\epsilon, \theta^B, \theta^R) := u^{\theta^B}(0, x(\theta^B) - \epsilon) - u^{\theta^B}(x(\theta^R), x(\theta^B) - \epsilon).$$

Notice that, as in the proof of Lemma 1, we have $C^{B\cdot}(\epsilon, \theta^B, \theta^R) > 0$ $\Big(C^{R\cdot}(\epsilon, \theta^B, \theta^R) > 0\Big)$ for all $\theta^B \in \Theta^B, \theta^R \in \Theta^R$ as player R wants to mis-coordinate (as player B wants to coordinate). Based on this let

$$C(\epsilon) := \min_{a \in \{B_H, B_T, R_H, R_T\}} \min_{\theta^B, \theta^R} \mu^{\theta^B} \mu^{\theta^R} C^a(\epsilon, \theta^B, \theta^R).$$

Now assume that the statement of the theorem does not to hold. Then there must exist $c \in (0, C(\epsilon))$ such that

(a) $\rho^p \geq x(\theta^{-p}_{max}) + \epsilon$ or

(b) $\rho^p \geq 1 - x(\theta^{-p}_{min}) + \epsilon$

for some $p \in \{B, R\}$ in an equilibrium. Again, the reasoning is almost identical for both players and w.l.o.g. we consider $p = B$.

Case (a): $\rho^B \geq x(\theta^R_{max}) + \epsilon$
By Lemma 2 there exist types $\bar{\theta}^B \in \Theta^B, \hat{\theta}^R \in \Theta^R$ such that $\sigma^{\bar{\theta}^B}(\hat{\theta}^R) = 1$. We then have

$$\rho^B \sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1 - \rho^B)\sigma^{\bar{\theta}^B}(\emptyset) \geq x(\hat{\theta}^R) + \epsilon > x(\hat{\theta}^R). \tag{8}$$

This implies that $\sigma^{\hat{\theta}^R}(\bar{\theta}^B) = 0$ as player R wants to mis-coordinate and as $\rho^R > 0$. Inserting this into Inequality (2a) gives $(1 - \rho^R)\sigma^{\hat{\theta}^R}(\emptyset) > x(\bar{\theta}^B)$. From this we deduce that $1 - \rho^R > x(\bar{\theta}^B)$ and $\sigma^{\hat{\theta}^R}(\emptyset) > x(\bar{\theta}^B)$. Now consider an alternative strategy for player R with $\breve{\rho}^R = 1$ and $\sigma^{\theta^R}(\theta^B)$ a best response for all $\theta^B \in \Theta^B, \theta^R \in \Theta^R$. By Inequality (8) deviating to this strategy player R would obtain an additional payoff of at least

$$\begin{aligned}
&(1 - \rho^R)\Big(\mu^{\bar{\theta}^B}\mu^{\hat{\theta}^R}\Big(u^{\hat{\theta}^R}(0, \rho^B\sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1 - \rho^B)\sigma^{\bar{\theta}^B}(\emptyset)) \\
&\qquad\qquad - u^{\hat{\theta}^R}(\sigma^{\hat{\theta}^R}(\emptyset), \rho^B\sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1 - \rho^B)\sigma^{\bar{\theta}^B}(\emptyset))\Big) - c\Big) \\
\geq &(1 - \rho^R)\Big(\mu^{\bar{\theta}^B}\mu^{\hat{\theta}^R}\Big(u^{\hat{\theta}^R}(0, x(\hat{\theta}^R) + \epsilon) - u^{\hat{\theta}^R}(\sigma^{\hat{\theta}^R}(\emptyset), x(\hat{\theta}^R) + \epsilon)\Big) - c\Big) \\
> &x(\bar{\theta}^B)\Big(\mu^{\bar{\theta}^B}\mu^{\hat{\theta}^R}\Big(u^{\hat{\theta}^R}(0, x(\hat{\theta}^R) + \epsilon) - u^{\hat{\theta}^R}(x(\bar{\theta}^B), x(\hat{\theta}^R) + \epsilon)\Big) - c\Big) > 0
\end{aligned}$$

as we have $c \in (0, C(\epsilon))$. We, thus, arrive at a contradiction.

Case (b): $\rho^B \geq 1 - x(\theta^R_{min}) + \epsilon$
The proof is analogous to that of Case (a) and omitted. □

A.4 Proof of Proposition 2

The proof is by construction. We identify a particular strategy profile $\Big(\rho^p, (\sigma^{\theta^p})_{\theta^p \in \Theta^p}\Big)_{p \in \{B,R\}}$ with the desired property and then show that it is an equilibrium. Let

$$\rho^p = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min}), \tag{9a}$$

$$\sigma^{\theta^p}(\emptyset) = \frac{1}{1-\rho^p} x(\theta^{-p}_{min}) \quad \forall \theta^p \in \Theta^p \backslash \{\theta^p_{max}, \theta^p_{min}\} \tag{9b}$$

$$\sigma^{\theta^p}(\theta^{-p}) = \begin{cases} \frac{1}{\rho^p}\Big(x(\theta^{-p}) - x(\theta^{-p}_{min})\Big) & \text{if} \quad \rho^p > 0 \\ 0 & \text{if} \quad \rho^p = 0 \end{cases} \quad \forall \theta^p \in \Theta^p, \theta^{-p} \in \Theta^{-p}. \tag{9c}$$

Note that $\sigma^{\theta^p}(\theta^{-p}_{max}) = 1$ and $\sigma^{\theta^p}(\theta^{-p}_{min}) = 0$ for all $p \in \{B, R\}$ and $\theta^p \in \Theta^p$ if $x(\theta^{-p}_{max}) > x(\theta^{-p}_{min})$.[20] The strategy profile is, thus, fully specified except for the behavior of uninformed extreme types. In the case that $x(\theta^p_{max}) > x(\theta^p_{min})$ let $\sigma^{\theta^p_{max}}(\emptyset)$ and $\sigma^{\theta^p_{min}}(\emptyset)$ be chosen to satisfy

$$\sum_{\theta^{-p}} \mu^{\theta^{-p}} \Big(u^{\theta^{-p}}(1, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)) - u^{\theta^{-p}}(0, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)) \Big) = \frac{c}{\mu^{\theta^p_{max}}} \tag{10}$$

and

$$\frac{\mu^{\theta^p_{max}}}{\mu^{\theta^p_{max}} + \mu^{\theta^p_{min}}} \sigma^{\theta^p_{max}}(\emptyset) + \frac{\mu^{\theta^p_{min}}}{\mu^{\theta^p_{max}} + \mu^{\theta^p_{min}}} \sigma^{\theta^p_{min}}(\emptyset) = \frac{1}{1-\rho^p} x(\theta^{-p}_{min}). \tag{11}$$

In the case that $x(\theta^p_{max}) = x(\theta^p_{min})$ however let

$$\sigma^{\theta^p_{max}}(\emptyset) = \sigma^{\theta^p_{min}}(\emptyset) = \frac{1}{1-\rho^p} x(\theta^{-p}_{min}). \tag{12}$$

[20] This means that in case that they are informed, both players of any type play pure action strategies against extreme type opponents.

For the remainder of the proof we distinguish these two cases.

Case 1: $x(\theta^p_{max}) > x(\theta^p_{min})$

Before we move on to prove that the considered strategy profile is indeed an equilibrium in this case, we need to make sure that it is well-defined. For this we need to show that Equation (10) has a feasible solution for $c = 0$ and $c > 0$ sufficiently small. Consider

$$\sigma^{\theta^p_{max}}(\emptyset) = \frac{1}{1-\rho^p}\Big(x(\theta^{-p}_{min}) + \epsilon^p\Big) = \frac{x(\theta^{-p}_{min}) + \epsilon^p}{1 - x(\theta^{-p}_{max}) + x(\theta^{-p}_{min})},$$

where $\epsilon^p \in \mathbb{R}$. For $c = 0$ let $\epsilon^p = 0$. We then have $\sigma^{\theta^p_{max}}(\emptyset) \in (0,1)$ and

$$\text{LHS of (10)} = \sum_{\theta^{-p}} \mu^{\theta^{-p}}\Big(u^{\theta^{-p}}(1, x(\theta^{-p})) - u^{\theta^{-p}}(0, x(\theta^{-p}))\Big) = 0 = \text{RHS of (10)}$$

since player $-p$ of type θ^{-p} is indifferent between both actions if the opponent plays $x(\theta^{-p})$. Equation (11) then implies $\sigma^{\theta^p_{min}}(\emptyset) = \sigma^{\theta^p_{max}}(\emptyset)$.

Now consider $c > 0$. Notice first that the left-hand side of (10) is a linear function in ϵ^p which is strictly decreasing (increasing) for $p = B$ ($p = R$). To see this, consider temporarily and w.l.o.g. $-p = B$ and some type θ^B whose payoffs are represented by the matrix

	H	T
H	$u_{H,H}$	$u_{H,T}$
T	$u_{T,H}$	$u_{T,T}$

where $u_{H,H}, u_{H,T}, u_{T,H}, u_{T,T} \in \mathbb{R}$. As player B wants to coordinate actions, we must have $u_{H,H} > u_{T,H}$ and $u_{T,T} > u_{H,T}$. Further, we calculate $x(\theta^B) = \frac{u_{T,T} - u_{H,T}}{u_{H,H} - u_{T,H} + u_{T,T} - u_{H,T}}$. Our claim follows immediately as this gives

$$\begin{aligned}
&u^{\theta^B}(1, x(\theta^B) + \epsilon^R) - u^{\theta^B}(0, x(\theta^B) + \epsilon^R) \\
=&u_{H,H}(x(\theta^B) + \epsilon^R) + u_{H,T}(1 - x(\theta^B) - \epsilon^R) - \\
-&u_{T,H}(x(\theta^B) + \epsilon^R) - u_{T,T}(1 - x(\theta^B) - \epsilon^R) \\
=&(u_{H,H} - u_{T,H} + u_{T,T} - u_{H,T})(x(\theta^B) + \epsilon^R) - (u_{T,T} - u_{H,T}) \\
=&\underbrace{(u_{H,H} - u_{T,H} + u_{T,T} - u_{H,T})}_{>0}\epsilon^R.
\end{aligned}$$

So, generally speaking, we have that for every $c > 0$ sufficiently small there exists a unique $\epsilon^B < 0$ ($\epsilon^R > 0$) such that both Equations (10) and (11) are fulfilled and $\sigma^{\theta^p_{max}}(\emptyset), \sigma^{\theta^p_{min}}(\emptyset) \in [0,1]$.

We now turn to proving that the proposed strategy profile is indeed an equilibrium. Suppose that in the conflict game both players B and R are playing a strategy as considered above. Then player $-p \in \{B, R\}$ cannot improve by deviating if the following conditions are satisfied:

- $\sigma^{\theta^{-p}}(\theta^p)$ is a best response to $\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)$ for all $\theta^p \in \Theta^p, \theta^{-p} \in \Theta^{-p}$,
- $\sigma^{\theta^{-p}}(\emptyset)$ is a best response to $\sum_{\theta^p} \mu^{\theta^p} \Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big)$ for all $\theta^{-p} \in \Theta^{-p}$,
- $\sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}^{\theta^{-p}}_{Info} = \sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}^{\theta^{-p}}_{N} + c$.

In the following let $c = 0$ or $c > 0$ sufficiently small as mentioned above. Further let $p = B$ ($p = R$). Consider first the action strategies that types of player $-p$ face when they are informed. We calculate for $\theta^{-p} \in \Theta^{-p}$, $\theta^p \in \Theta^p \backslash \{\theta^p_{max}, \theta^p_{min}\}$:

$$\begin{aligned}
\rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset) &= x(\theta^{-p}) + \epsilon^p &&\leq (\geq) x(\theta^{-p}),\\
\rho^p \sigma^{\theta^p_{min}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{min}}(\emptyset) &= x(\theta^{-p}) - \tfrac{\mu^{\theta^p_{max}}}{\mu^{\theta^p_{min}}}\epsilon^p &&\geq (\leq) x(\theta^{-p}),\\
\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset) &&&= \quad x(\theta^{-p}).
\end{aligned} \tag{13}$$

Hence, $\sigma^{\theta^{-p}}(\theta^p_{max}) = 1$ and $\sigma^{\theta^{-p}}(\theta^p_{min}) = 0$ are indeed best responses for all $p \in \{B, R\}$, $\theta^{-p} \in \Theta^{-p}$. Against all other types $\theta^p \in \Theta^p \backslash \{\theta^p_{max}, \theta^p_{min}\}$, any informed type $\theta^{-p} \in \Theta^{-p}$ is indifferent between both actions.

Beyond that, any uninformed player type $\theta^{-p} \in \Theta^{-p}$ faces

$$\begin{aligned}
&\sum_{\theta^p} \mu^{\theta^p} \Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big)\\
=&\mu^{\theta^p_{max}} \Big(x(\theta^{-p}) + \epsilon^p\Big) + \mu^{\theta^p_{min}} \Big(x(\theta^{-p}) - \frac{\mu^{\theta^p_{max}}}{\mu^{\theta^p_{min}}}\epsilon^p\Big) + \sum_{\theta^p \notin \{\theta^p_{max}, \theta^p_{min}\}} \mu^{\theta^p} x(\theta^{-p})\\
=&x(\theta^{-p})
\end{aligned}$$

and is therefore indifferent between both actions.

Finally, we have to examine the expected payoffs. For an uninformed player type $\theta^{-p} \in \Theta^{-p}$ we have

$$\begin{aligned}\mathcal{U}_N^{\theta^{-p}} &= u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\emptyset), \sum_{\theta^p} \mu^{\theta^p}\Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big)\Big) \\ &= u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\emptyset), x(\theta^{-p})\Big).\end{aligned}$$

If θ^{-p} is informed, then her expected payoff (ex costs) is given by

$$\begin{aligned}&\mathcal{U}_{Info}^{\theta^{-p}} = \sum_{\theta^p} \mu^{\theta^p} u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\theta^p), \rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big) \\ &= \mu^{\theta^p_{max}} u^{\theta^{-p}}\Big(1, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)\Big) + \\ &\mu^{\theta^p_{min}} u^{\theta^{-p}}\Bigg(0, \frac{1}{\mu^{\theta^p_{min}}}\Big(x(\theta^{-p}) - \sum_{\theta^p \neq \theta^p_{min}} \mu^{\theta^p}\Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big)\Big)\Bigg) + \\ &\sum_{\theta^p \notin \{\theta^p_{max}, \theta^p_{min}\}} \mu^{\theta^p} u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\theta^p), x(\theta^{-p})\Big) \\ &= \mu^{\theta^p_{max}}\Big(u^{\theta^{-p}}\Big(1, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)\Big) - \\ &u^{\theta^{-p}}\Big(0, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)\Big)\Big) + u^{\theta^{-p}}\Big(0, x(\theta^{-p})\Big)\end{aligned}$$

Notice that according to (13) we have $\mathcal{U}_{Info}^{\theta^{-p}} \geq u^{\theta^{-p}}\Big(0, x(\theta^{-p})\Big) = \mathcal{U}_N^{\theta^{-p}}$ for all $\theta^{-p} \in \Theta^{-p}$. Taken together we get

$$\begin{aligned}(10) \Leftrightarrow &\sum_{\theta^{-p}} \mu^{\theta^{-p}}\Bigg(\mu^{\theta^p_{max}}\Big(u^{\theta^{-p}}\Big(1, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)\Big) \\ &- u^{\theta^{-p}}\Big(0, \rho^p \sigma^{\theta^p_{max}}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p_{max}}(\emptyset)\Big)\Big) + u^{\theta^{-p}}\Big(0, x(\theta^{-p})\Big)\Bigg) \\ &= \sum_{\theta^{-p}} \mu^{\theta^{-p}} u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\emptyset), x(\theta^{-p})\Big) + c \\ \Leftrightarrow &\sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}_{Info}^{\theta^{-p}} = \sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}_N^{\theta^{-p}} + c.\end{aligned}$$

This means that player $-p$ is indeed indifferent between acquiring empathy and not acquiring it. Thus, we established that player $-p$ has no incentives to deviate from the considered strategy in this case.

Case 2: $x(\theta^p_{max}) = x(\theta^p_{min})$

Suppose again that both players B and R are playing a strategy as

considered above. As according to Equation (9a) we have $\rho^{-p} = 0$ in this case, player $-p$ cannot improve by deviating if the following conditions are satisfied:

- $\sigma^{\theta^{-p}}(\emptyset)$ is a best response to $\sum_{\theta^p} \mu^{\theta^p} \Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big)$ for all $\theta^{-p} \in \Theta^{-p}$,
- $\sum_{\theta^{-p},\theta^p} \mu^{\theta^{-p}} \mu^{\theta^p} u^{\theta^{-p}} \Big(s^{\theta^{-p}}(\theta^p), \rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big) \leq \sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}_N^{\theta^{-p}} + c$ for all $\Big(s^{\theta^{-p}}(\theta^p)\Big)_{\theta^{-p},\theta^p} \in \Delta(A)^{n^{-p} \times n^p}$.

Taking into account Equations (9) and (12), concerning the first condition we simply have

$$\sum_{\theta^p} \mu^{\theta^p} \Big(\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big) = \sum_{\theta^p} \mu^{\theta^p} x(\theta^{-p}) = x(\theta^{-p}).$$

Hence, this condition is obviously fulfilled as any uninformed type θ^{-p} is indifferent between both actions.

The second condition states that the ex-ante expected payoff of player $-p$ from not acquiring empathy must be greater than or equal to the maximal payoff (minus costs) she could get instead from acquiring empathy and playing freely choosable action strategies which can be conditioned on the opponent's type. For all θ^{-p}, $\Big(s^{\theta^{-p}}(\theta^p)\Big)_{\theta^p}$ we have

$$\sum_{\theta^p} \mu^{\theta^p} u^{\theta^{-p}} \Big(s^{\theta^{-p}}(\theta^p), \underbrace{\rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)}_{=x(\theta^{-p})}\Big) = u^{\theta^{-p}}\Big(\cdot, x(\theta^{-p})\Big).$$

On the contrary, type θ^{-p} receives

$$\mathcal{U}_N^{\theta^{-p}} = u^{\theta^{-p}}\Big(\sigma^{\theta^{-p}}(\emptyset), x(\theta^{-p})\Big)$$

if she is uninformed. Thus, we have

$$\sum_{\theta^{-p},\theta^p} \mu^{\theta^{-p}} \mu^{\theta^p} u^{\theta^{-p}} \Big(s^{\theta^{-p}}(\theta^p), \rho^p \sigma^{\theta^p}(\theta^{-p}) + (1-\rho^p)\sigma^{\theta^p}(\emptyset)\Big) = \sum_{\theta^{-p}} \mu^{\theta^{-p}} \mathcal{U}_N^{\theta^{-p}}$$

for all $\Big(s^{\theta^{-p}}(\theta^p)\Big)_{\theta^{-p},\theta^p} \in \Delta(A)^{n^{-p} \times n^p}$. This concludes Case 2 and the proof as a whole.

A.5 Proof of Proposition 3

Recall the proof of Theorem 1. In Case 1 of Part (ii) we already established that we must have

$$\rho^p = x(\theta^{-p}_{max}) - x(\theta^{-p}_{min}) \tag{14}$$

if $\rho^{-p} = 0$ for $p \in \{B, R\}$. Notice that $\rho^p > 0$ then implies $x(\theta^{-p}_{max}) > x(\theta^{-p}_{min})$. Taking into account Theorem 1(i) in this situation we also have that $0 = \rho^{-p} \geq x(\theta^{p}_{max}) - x(\theta^{p}_{min}) \geq 0$, and thus $\rho^{-p} = x(\theta^{p}_{max}) - x(\theta^{p}_{min})$. In what follows we distinguish the two cases considered in the proposition.

Part 1: $n^B = 1$ and $n^R > 1$ ($n^B > 1$ and $n^R = 1$, respectively)
W.l.o.g. consider the case $n^B = 1$ (such that $\Theta^B = \{\theta^B\}$) and $n^R > 1$ and let $c \in (0, C)$ sufficiently small. Assume that $\rho^R > 0$ in an equilibrium. Then according to Lemma 2 there must exist $\bar{\theta}^R$ and $\hat{\theta}^B, \tilde{\theta}^B$ fulfilling Inequalities (2). This however implies $\hat{\theta}^B \neq \tilde{\theta}^B$ which is a contradiction as we have $n^B = 1$. Thus, we must have $\rho^R = 0$ which (together with the above considerations) establishes uniqueness of the empathy levels for this part of the proof.

By assumption we have that $x(\theta^R_{max}) > x(\theta^R) > x(\theta^R_{min})$ for all $\theta^R \in \Theta^R \backslash \{\theta^R_{max}, \theta^R_{min}\}$. We now show that the equilibrium considered in the proof of Proposition 2 is unique up to variations of the action strategies $\sigma^{\theta^R}(\theta^B)$ which are played with probability $\rho^R = 0$. Notice first that according to Lemma 1 we must have

$$\rho^B \sigma^{\theta^B}(\theta^R) + (1 - \rho^B)\sigma^{\theta^B}(\emptyset) = x(\theta^R) \tag{15}$$

for all $\theta^R \in \Theta^R$. Taking into account Equation (14) this gives

$$\begin{aligned}(x(\theta^R_{max}) - x(\theta^R_{min}))(\sigma^{\theta^B}(\theta^R_{max}) - \sigma^{\theta^B}(\theta^R_{min})) &= \rho^B(\sigma^{\theta^B}(\theta^R_{max}) - \sigma^{\theta^B}(\theta^R_{min})) \\ &= x(\theta^R_{max}) - x(\theta^R_{min}).\end{aligned}$$

Hence, we must have $\sigma^{\theta^B}(\theta^R_{max}) = 1$ and $\sigma^{\theta^B}(\theta^R_{min}) = 0$. Again according

to Lemma 1 this implies that

$$\sigma^{\theta^B}(\theta^R) = \sigma^{\theta^B}(\theta^R) - \sigma^{\theta^B}(\theta^R_{min}) = \frac{1}{\rho^B}(x(\theta^R) - x(\theta^R_{min})) \;\Leftrightarrow\; (9c)$$

for all $\theta^R \in \Theta^R$. Moreover, by Equation (15) this induces

$$\sigma^{\theta^B}(\emptyset) = \frac{1}{1-\rho^B}x(\theta^R_{min}) \;\Leftrightarrow\; (9b).$$

As $x(\theta^R_{max}) > x(\theta^R) > x(\theta^R_{min})$ we have $\sigma^{\theta^B}(\theta^R) \in (0,1)$ for all $\theta^R \in \Theta^R \backslash \{\theta^R_{max}, \theta^R_{min}\}$. This means that θ^B must be indifferent against any opponent type $\theta^R \in \Theta^R \backslash \{\theta^R_{max}, \theta^R_{min}\}$ if she is informed. Thus, we must have $\sigma^{\theta^R}(\emptyset) = x(\theta^B)$ for all $\theta^R \in \Theta^R \backslash \{\theta^R_{max}, \theta^R_{min}\}$. Equation (1) of Lemma 1 then transforms to

$$\begin{aligned} & \sum_{\theta^R} \mu^{\theta^R}\sigma^{\theta^R}(\emptyset) = x(\theta^B) \\ \Leftrightarrow \quad & \frac{\mu^{\theta^R_{max}}}{\mu^{\theta^R_{max}} + \mu^{\theta^R_{min}}}\sigma^{\theta^R_{max}}(\emptyset) + \frac{\mu^{\theta^R_{min}}}{\mu^{\theta^R_{max}} + \mu^{\theta^R_{min}}}\sigma^{\theta^R_{min}}(\emptyset) = x(\theta^B) \;\Leftrightarrow\; (11). \end{aligned}$$

Together with Equation (10) (for $p = R$) this then uniquely determines $\sigma^{\theta^R_{max}}(\emptyset)$ and $\sigma^{\theta^R_{min}}(\emptyset)$. The reasoning is the same for $n^B > 1, n^R = 1$.

Part 2: $n^B = n^R = 2$

In this case we have $\Theta^p = \{\theta^p_{max}, \theta^p_{min}\}$ for $p = B, R$. We already know that uniqueness of the empathy levels follows immediately if we have $\rho^p = 0$ for some $p \in \{B, R\}$. So in this regard we only need to consider the case that $\rho^B, \rho^R > 0$. Again, we recall the proof of Theorem 1 and take Case 2 with $p = B$ as a starting point. Consider its three subcases.

Subcase (a): $\rho^B\sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) = x(\hat{\theta}^R)$ and
$\rho^B\sigma^{\bar{\theta}^B}(\tilde{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) = x(\tilde{\theta}^R)$

This subcase is again straightforward as we simply have $\rho^B = x(\hat{\theta}^R) - x(\tilde{\theta}^R)$ and know already that it is $\rho^B \geq x(\theta^R_{max}) - x(\theta^R_{min})$. Therefore it must be $\hat{\theta}^R = \theta^R_{max}$ and $\tilde{\theta}^R = \theta^R_{min}$.

Subcase (b): $\rho^B\sigma^{\bar{\theta}^B}(\hat{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) > x(\hat{\theta}^R)$

Recall Inequality (7). Lemma 1 then implies that

$$\rho^R \sigma^{\tilde{\theta}^R}(\breve{\theta}^B) + (1-\rho^R)\sigma^{\tilde{\theta}^R}(\emptyset) < x(\breve{\theta}^B)$$

as here it is $\{\theta^R \in \Theta^R \mid \theta^R \neq \hat{\theta}^R\} = \{\tilde{\theta}^R\}$. In turn, this induces $\sigma^{\breve{\theta}^B}(\tilde{\theta}^R) = 0$. Moreover, recall that it is $\sigma^{\bar{\theta}^B}(\tilde{\theta}^R) = 0$, $\sigma^{\bar{\theta}^B}(\hat{\theta}^R) = 1$ and $\sigma^{\breve{\theta}^B}(\hat{\theta}^R) = 1$. Further, we know again by Lemma 1 that it must be

$$\rho^B \Big(\sum_{\theta^B} \mu^{\theta^B} \sigma^{\theta^B}(\theta^R_{max}) - \sum_{\theta^B} \mu^{\theta^B} \sigma^{\theta^B}(\theta^R_{min}) \Big) = x(\theta^R_{max}) - x(\theta^R_{min}).$$

If it were $\hat{\theta}^R = \theta^R_{min}, \tilde{\theta}^R = \theta^R_{max}$, then this would imply $\rho^B = x(\theta^R_{min}) - x(\theta^R_{max}) \leq 0$. So it must be $\hat{\theta}^R = \theta^R_{max}, \tilde{\theta}^R = \theta^R_{min}$ which implies $\rho^B = x(\theta^R_{max}) - x(\theta^R_{min})$.

Subcase (c): $\rho^B \sigma^{\bar{\theta}^B}(\tilde{\theta}^R) + (1-\rho^B)\sigma^{\bar{\theta}^B}(\emptyset) < x(\tilde{\theta}^R)$
This subcase proceeds analogously to Subcase (b) and is therefore omitted.

Bibliography

Amir, R. and Lazzati, N. (2014). Endogenous information acquisition in bayesian games with strategic complementarities. mimeo, University of Iowa.

Aumann, R. J. and Maschler, M. (1972). Some thoughts on the minimax principle. *Management Science*, 18(5):54–63.

Bergemann, D., Shi, X., and Välimäki, J. (2009). Information acquisition in interdependent value auctions. *Journal of the European Economic Association*, 7(1):61–89.

Costa-Gomes, M., Crawford, V. P., and Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235.

Costa-Gomes, M. A. and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768.

Crawford, V. P. (2003). Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 93(1):133–149.

Crawford, V. P. and Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113.

Dekel, E., Ely, J. C., and Yilankaya, O. (2007). Evolution of preferences. *The Review of Economic Studies*, 74(3):685–704.

Dimitrova, M. and Schlee, E. E. (2003). Monopoly, competition and information acquisition. *International Journal of Industrial Organization*, 21(10):1623–1642.

Ely, J. C. and Yilankaya, O. (2001). Nash equilibrium and evolution of preferences. *Journal of Economic Theory*, 97(2):255–272.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory*, 24(4):323–344.

Güth, W. and Yaari, M. E. (1992). Explaining reciprocal behavior in a simple strategic game. In *Explaining Process and Change: Approaches to Evolutionary Economics*, pages 23–24. University of Michigan Press.

Hauk, E. and Hurkens, S. (2001). Secret information acquisition in cournot markets. *Economic Theory*, 18(3):661–681.

Heifetz, A., Shannon, C., and Spiegel, Y. (2007a). The dynamic evolution of preferences. *Economic Theory*, 32(2):251–286.

Heifetz, A., Shannon, C., and Spiegel, Y. (2007b). What to maximize if you must. *Journal of Economic Theory*, 133(1):31–57.

Heller, Y. and Mohlin, E. (2015a). Coevolution of deception and preferences: Darwin and Nash meet Machiavelli. Available at SSRN 2490370.

Heller, Y. and Mohlin, E. (2015b). Observations on cooperation. Available at SSRN 2558570.

Hellwig, C. and Veldkamp, L. (2009). Knowing what others know: Coordination motives in information acquisition. *The Review of Economic Studies*, 76(1):223–251.

Herold, F. and Kuzmics, C. (2009). Evolutionary stability of discrimination under observability. *Games and Economic Behavior*, 67(2):542–551.

Ho, T.-H., Camerer, C., and Weigelt, K. (1998). Iterated dominance and iterated best response in experimental "p-beauty contests". *American Economic Review*, 88(4):947–969.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hwang, H.-S. (1993). Optimal information acquisition for heterogenous duopoly firms. *Journal of Economic Theory*, 59(2):385–402.

Jansen, J. (2008). Information acquisition and strategic disclosure in oligopoly. *Journal of Economics & Management Strategy*, 17(1):113–148.

Koçkesen, L., Ok, E. A., and Sethi, R. (2000a). Evolution of interdependent preferences in aggregative games. *Games and Economic Behavior*, 31(2):303–310.

Koçkesen, L., Ok, E. A., and Sethi, R. (2000b). The strategic advantage of negatively interdependent preferences. *Journal of Economic Theory*, 92(2):274–299.

Köhler, W. (1929). *Gestalt Psychology*. Liveright.

Li, L., McKelvey, R. D., and Page, T. (1987). Optimal research for cournot oligopolists. *Journal of Economic Theory*, 42(1):140–166.

Matêjka, F. and McKay, A. (2012). Simple market equilibria with rationally inattentive consumers. *American Economic Review, Papers and Proceedings of the 104th Annual Meeting of the American Economic Association*, 102(3):24–29.

Matêjka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

McKelvey, R. D., McLennan, A. M., and Turocy, T. L. (2014). Gambit: Software tools for game theory. Version 14.1.0., http://www.gambit-project.org.

Mead, G. H. (1934). *Mind, Self, and Society: From the Standpoint of a Social Behaviorist*. University of Chicago Press.

Mengel, F. (2012). Learning across games. *Games and Economic Behavior*, 74(2):601–619.

Myatt, D. P. and Wallace, C. (2012). Endogenous information acquisition in coordination games. *The Review of Economic Studies*, 79(1):340–374.

Nachbar, J. H. (1990). "Evolutionary" selection dynamics in games: Convergence and limit properties. *International Journal of Game Theory*, 19(1):59–89.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326.

Ok, E. and Vega-Redondo, F. (2001). On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, 97:231–54.

Persico, N. (2000). Information acquisition in auctions. *Econometrica*, 68(1):135–148.

Piaget, J. (1932). *Le Jugement Moral chez l'Enfant*. Presses Universitaires de France.

Pruzhansky, V. (2011). Some interesting properties of maximin strategies. *International Journal of Game Theory*, 40(2):351–365.

Robalino, N. and Robson, A. J. (2012). The economic approach to "theory of mind". *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2224–2233.

Robalino, N. and Robson, A. J. (2015). The evolution of strategic sophistication. *American Economic Review*, forthcoming.

Robson, A. J. and Samuelson, L. (2010). The evolutionary foundations of preferences. *Handbook of Social Economics*, 1:221–310.

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., and Convit, A. (2007). Who cares? Revisiting empathy in Asperger syndrome. *Journal of Autism and Developmental Disorders*, 37(4):709–715.

Samuelson, L. (2001). Introduction to the evolution of preferences. *Journal of Economic Theory*, 97(2):225–230.

Shamay-Tsoory, S. G., Aharon-Peretz, J., and Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3):617–627.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163.

Solan, E. and Yariv, L. (2004). Games with espionage. *Games and Economic Behavior*, 47(1):172–199.

Stahl, D. O. and Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327.

Stahl, D. O. and Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.