

### Appendix 3: Cognitive Hierarchy

As a robustness check, we conduct our analysis with the cognitive hierarchy model of Camerer, Ho and Chong (2004). There, the distribution of types is Poisson distributed, i.e., the proportion of  $T_k$  is given by

$$p_k = \frac{e^{-\tau} \tau^k}{k!}.$$

$T_k$  best responds given the belief that the others players are  $T_0$  up to  $T_{k-1}$ .  $T_k$ 's belief about the proportion of  $T_{l < k}$  is

$$g_k(l) = \frac{p_l}{\sum_{h=0}^{k-1} p_h}.$$

The cognitive hierarchy model is developed for normal form games only. In order to adapt the model to games with pre-play communication we must specify how beliefs are updated after messages have been received. We assume Bayesian updating. For the games preceded by one round of communication, let  $q_{ki}(m_i)$  denote the probability that a  $T_k$  player  $i$  sends the message  $m_i$  (and is allowed to send a message).  $T_k$ 's belief that the sender  $i$  is a  $T_{l < k}$  player conditional upon receiving the message  $m_i$  is

$$g_{ki}(l|m_i) = \frac{g_k(l) q_{li}(m_i)}{\sum_{h=0}^{k-1} g_k(h) q_{hi}(m_i)} = \frac{p_l q_{li}(m_i)}{\sum_{h=0}^{k-1} p_h q_{hi}(m_i)},$$

where the latter equality follows from the definition of  $g_k(l)$ .

We retain the assumption that players randomize uniformly when indifferent, but that they prefer to be honest if it does not affect expected payoffs. This implies that the behavior of  $T_0$  and  $T_1$  is the same in the cognitive hierarchy model and in the level- $k$  model.

A feature of the cognitive hierarchy model is that if  $T_k$  plays a strategy that is a best response to  $T_k$  opponent in a two-player game with one round of pre-play communication, then  $T_{m > k}$  will play the same strategy. Since this result will be used repeatedly we state it separately in Lemma 1.

**LEMMA 1:** *Let  $G$  be a symmetric two-player normal form game. If  $T_k$  plays a strategy profile that is a best response to a  $T_k$  opponent in  $G$ ,  $\Gamma_I(G)$  or  $\Gamma_{II}(G)$ , then  $T_{m > k}$  play this strategy too.*

**PROOF:**

Consider the case of one-way communication (the proof for two-way communication and without communication is analogous). Let the strategy played by  $T_k$  be denoted  $s^* = \langle m^*, a^*, f^*(m) \rangle$ . Consider a  $T_k$  player that received the message  $m$ . We know that  $f^*(m)$  is the action that maximizes expected payoff conditional on receiving  $m$  given the belief that the opponent is  $T_{l < k}$  with probability

$$g_k(l|m) = \frac{p_l q_l(m)}{\sum_{h=0}^{k-1} p_h q_h(m)}.$$

Similarly, a  $T_{k+1}$  player that receives the same message  $m$  best responds given the belief that the opponent is a  $T_{l < k+1}$  player with probability

$$g^{(k+1)}(l|m) = \frac{p_l q_l(m)}{\sum_{h=0}^k p_h q_h(m)}.$$

Since  $f^*(m)$  maximizes the expected payoff of  $T_k$  and is a best response to a  $T_k$  sender, by

linearity of expected payoffs it must be a best response also to the mixture of types  $T_{k+1}$  believes to be facing (note that this argument does not extend to more than two players).

Now consider the communication stage of the game. The message  $m^*$  followed by the action  $a^*$  is a best response given the belief that the opponent is  $T_{l < k}$  with probability

$$g_k(l) = p_l / \sum_{h=0}^{k-1} p_h.$$

Similarly a  $T_{k+1}$  player believes that the opponent is  $T_{l < k+1}$  with probability

$$g_{k+1}(l) = p_l / \sum_{h=0}^k p_h.$$

Since  $m^*$  and  $a^*$  maximizes the payoff of  $T_k$  and is a best response against another  $T_k$  player, it must be a best response also to the mixture of types  $T_{k+1}$  believes to be facing.

By induction this reasoning holds for all  $T_{m > k}$  players.

In the cognitive hierarchy model, predicted behavior depends both on the payoff configuration and the average of the type distribution,  $\tau$ . A complete characterization of behavior is therefore intractable, and the remainder of this appendix focuses on  $T_2$  and  $T_3$  in the class of symmetric and generic  $2 \times 2$  games. However, a general characterization of the behavior of  $T_3$  in mixed motive games with two-way communication is also intractable, so in this case we focus on  $T_2$  only. For simplicity, we finally disregard cases in which the combination of  $\tau$  and the payoff structure of the game implies that  $T_{2+}$  is indifferent between strategies as well as games in which neither action is risk dominant.

Two general findings emerge from the analysis. First, when  $\tau$  is close to zero,  $T_2$  and  $T_3$  players are practically certain that the opponent is  $T_0$  and consequently play the same action as  $T_1$ . However,  $T_2$  and  $T_3$  may send another message since they take into account the possibility that the opponent is (a responsive)  $T_1$ . Second, for sufficiently large  $\tau$ ,  $T_{2+}$  play as in the level- $k$  model in all games except possibly  $T_{3+}$  in mixed motive games with two-way communication. In this part of the parameter space, the level- $k$  model is robust to the assumption about lexicographic beliefs. For intermediate levels of  $\tau$ ,  $T_2$  and  $T_3$  best-respond to the mixture of lower-level types they believe they are facing.

An interesting new finding is that the behavior in the Stag Hunt hypothesized by Aumann (1990) emerges endogenously in the model. With the payoffs in Aumann's original example, depicted in Figure 2, a  $T_{3+}$  player sends the message  $h$  and plays  $L$  as sender, and plays  $L$  irrespective of the received message as receiver, whenever  $\tau$  is between 0.547 and 1.646. As a sender,  $T_{3+}$  does so in order to induce  $T_1$  and  $T_2$  to play  $H$ , but believes that there such a high probability of meeting a randomizing  $T_0$  that it is better to play  $L$ .  $T_{3+}$  ignores the received message because of the likelihood of meeting a  $T_2$  opponent, who sends  $h$  messages that are not self-signalling.

In dominance solvable games,  $T_1$  sends and plays the dominant strategy, so by Lemma 1,  $T_{1+}$  does so too (irrespective of whether communication is possible). We now proceed to characterize the behavior in the two remaining classes of games.

### Coordination games

As before, we assume that  $H(igh)$  is the payoff dominant equilibrium, i.e.,  $u_{HH} > u_{LL}$ .

OBSERVATION 8: (No communication)  $T_{1+}$  plays the risk dominant action.

PROOF:

$T_1$  plays the risk dominant action. Consequently, by Lemma 1 all higher level types do the same.

Absent communication, behavior is the same in the level- $k$  and cognitive hierarchy models.

**OBSERVATION 9:** (*One-way communication*) If  $H$  is the risk dominant action,  $T_{1+}$  sends  $h$  and plays  $H$  as sender and responds to messages as receiver. If  $L$  is the risk dominant action,  $T_1$  sends  $l$  and plays  $L$  as sender and respond to messages as receiver, but the behavior of  $T_{2+}$  depends on the payoff structure of the game:

**Case 1** ( $u_{LH} > u_{LL}$ ): Let  $\alpha \equiv (u_{LL} - u_{HL}) / (u_{HH} - u_{LH})$ . If  $\tau < (\alpha - 1) / 2$ , then  $T_2$  plays  $\langle h, L, H, L \rangle$  and  $T_3$  plays as follows:

$T_3$  plays  $\langle h, L, H, L \rangle$  if  $\tau < \sqrt{\alpha} - 1$  and  $\tau < (\sqrt{\alpha + 1} + 1) / \alpha$ ,

$T_{3+}$  plays  $\langle h, L, L, L \rangle$  if  $(\sqrt{\alpha + 1} + 1) / \alpha < \tau < \sqrt{\alpha} - 1$ ,

$T_{3+}$  plays  $\langle h, H, H, L \rangle$  if  $\sqrt{\alpha} - 1 < \tau < (\sqrt{\alpha + 1} + 1) / \alpha$ ,

$T_3$  plays  $\langle h, H, L, L \rangle$  if  $\tau > \sqrt{\alpha} - 1$  and  $\tau > (\sqrt{\alpha + 1} + 1) / \alpha$ .

If  $\tau > (\alpha - 1) / 2$ , then  $T_{2+}$  play  $\langle h, H, H, L \rangle$ .

**Case 2** ( $u_{LH} < u_{LL}$ ): Let  $\beta \equiv (u_{LH} - u_{HL}) / (u_{HH} - u_{LL})$ . If  $\tau < (\beta - 1) / 2$ , then  $T_2$  plays  $\langle l, L, H, L \rangle$  and  $T_3$  plays  $\langle l, L, H, L \rangle$  if  $\tau < \sqrt{\beta} - 1$  and  $\langle h, H, H, L \rangle$  if  $\tau > \sqrt{\beta} - 1$ . If  $\tau > (\beta - 1) / 2$ , then  $T_{2+}$  plays  $\langle h, H, H, L \rangle$ .

PROOF:

First consider the case when  $H$  is risk dominant. As in the level- $k$  model,  $T_1$  plays  $\langle h, H, H, L \rangle$  (facing randomizing  $T_0$  receivers and truthful  $T_0$  senders). Since this strategy is a best-response to itself,  $T_{2+}$  plays the same strategy.

Now consider the case when  $L$  is risk dominant so that  $T_1$  plays  $\langle l, L, H, L \rangle$ . For  $T_2$  senders, the strategy  $\langle l, H \rangle$  is dominated by  $\langle h, H \rangle$ , so we need not consider that strategy. The expected payoff for the remaining three sender strategies are

$$\begin{aligned}\pi(\langle l, L \rangle) &= g_2(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2(1) u_{LL}, \\ \pi(\langle h, L \rangle) &= g_2(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2(1) u_{LH}, \\ \pi(\langle h, H \rangle) &= g_2(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_2(1) u_{HH}.\end{aligned}$$

If  $u_{LH} > u_{LL}$ , then it is clear that  $T_2$  senders play either  $\langle h, L \rangle$  or  $\langle h, H \rangle$ . The payoff from playing  $\langle h, L \rangle$  is higher whenever  $\tau$  is sufficiently low,

$$\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2(u_{HH} - u_{LH})} = (\alpha - 1) / 2.$$

Similarly, if  $u_{LH} < u_{LL}$ , then  $T_2$  senders prefer  $\langle l, L \rangle$  over  $\langle h, H \rangle$  whenever

$$\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2(u_{HH} - u_{LL})} = (\beta - 1) / 2.$$

$T_2$  receivers face truthful  $T_1$  and  $T_2$  senders, so they respond to messages. It is clear that for sufficiently high  $\tau$ ,  $T_{2+}$  plays  $\langle h, H, H, L \rangle$ .

We now go on to consider the behavior of  $T_3$  when  $\tau$  is below the thresholds above. First suppose that  $u_{LH} > u_{LL}$  and  $\tau < (\alpha - 1)/2$ . Then  $T_3$  senders prefer  $\langle h, L \rangle$  over  $\langle h, H \rangle$  whenever

$$g_3(0) \frac{1}{2} (u_{LL} + u_{LH}) + (g_3(1) + g_3(2)) u_{LH} > g_3(0) \frac{1}{2} (u_{HL} + u_{HH}) + (g_3(1) + g_3(2)) u_{HH},$$

which simplifies to  $(1 + \tau/2)\tau < (\alpha - 1)/2$ . Since the left hand side is larger than  $\tau$ , this condition may or may not hold. Both sides of the inequality are positive, so the condition is equivalent to  $\tau < \sqrt{\alpha} - 1$ . Suppose now that  $u_{LH} < u_{LL}$ . Then  $T_3$  senders prefer  $\langle l, L \rangle$  over  $\langle h, H \rangle$  whenever

$$(1 + \tau/2)\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2(u_{HH} - u_{LL})} = (\beta - 1)/2$$

Both sides of the inequality are positive, so this condition is equivalent to  $\tau < \sqrt{\beta} - 1$ .

Finally,  $T_3$ 's behavior as receivers depend on the  $T_2$  senders. It is only when  $T_2$  senders send  $h$ , but play  $L$  that  $T_3$  may not respond to messages. If  $T_3$  receives a  $l$  message, it comes from a  $T_0$  player and  $T_3$  best responds by playing  $L$ . The payoff from each action upon receiving  $h$  is

$$\begin{aligned} \pi(H|h) &= g_3(0|h)u_{HH} + g_3(1|h)u_{HH} + g_3(2|h)u_{HL}, \\ \pi(L|h) &= g_3(0|h)u_{LH} + g_3(1|h)u_{LH} + g_3(2|h)u_{LL}. \end{aligned}$$

Playing  $L$  is preferable whenever

$$\frac{\tau^2}{1 + 2\tau} > \frac{u_{HH} - u_{LH}}{u_{LL} - u_{HL}} = 1/\alpha.$$

To illustrate the first case when  $L$  is risk dominant, Figure A3 displays the behavior of  $T_3$  as a function of  $\tau$  and the payoffs of the game. First note that  $\alpha$  has to be larger than 1 because  $L$  is risk dominant. Above the thick line in Figure A3,  $T_{2+}$  plays  $\langle h, H, H, L \rangle$  and below it  $T_2$  plays  $\langle h, L, H, L \rangle$ . Figure A3 shows the four different cases for the behavior of  $T_3$  in the latter case. For example, for the Stag Hunt depicted in Figure 2,  $\alpha = 7$ , implying that  $T_{3+}$  plays  $\langle h, L, L, L \rangle$  whenever  $0.547 < \tau < 1.646$ .

**OBSERVATION 10:** (*Two-way communication*)  $T_1$  randomizes messages and responds to received messages. Let  $\lambda \equiv (u_{HH} - u_{LH}) / (2u_{LL} - u_{LH} - u_{HH})$ . If  $L$  is the risk dominant action and  $0 < \lambda < (\beta - 1)/2$ , then  $T_{2+}$  plays  $\langle h, H, L \rangle$  if  $\tau < \lambda$ ,  $\langle l, L, L \rangle$  if  $\lambda < \tau < (\beta - 1)/2$ , and  $\langle h, H, H \rangle$  if  $\tau > (\beta - 1)/2$ . If either  $\lambda < 0$  or  $0 < (\beta - 1)/2 < \lambda$ ,  $T_{2+}$  plays  $\langle h, H, L \rangle$  if  $\tau < \alpha$  and  $\langle h, H, H \rangle$  if  $\tau > \alpha$ . If  $H$  is the risk dominant action, the behavior of  $T_2$  depends on the payoff structure of the game:

**Case 1** ( $u_{LH} + u_{HH} > u_{LL} + u_{HL}$ ):  $T_{2+}$  plays  $\langle h, H, L \rangle$  if  $\tau < \alpha$  and  $\langle h, H, H \rangle$  if  $\tau > \alpha$ .

**Case 2** ( $u_{LH} + u_{HH} < u_{LL} + u_{HL}$ ): Let  $\gamma \equiv (u_{LL} - u_{HL}) / (2u_{HH} - u_{LL} - u_{HL})$  and  $\delta \equiv (u_{HH} - u_{LL}) / (u_{LL} - u_{HL})$ . If  $\tau < \gamma$ , then  $T_2$  plays  $\langle l, H, L \rangle$ ;  $T_3$  plays  $\langle l, H, L \rangle$  if in addition  $\tau < (\sqrt{4\delta\gamma^2 + 1} - 1) / 2\gamma\delta$ , but plays  $\langle h, H, H \rangle$  if  $\tau > (\sqrt{4\delta\gamma^2 + 1} - 1) / 2\gamma\delta$ . If  $\tau > \gamma$ , then  $T_{2+}$  plays  $\langle h, H, H \rangle$ .

**PROOF:**

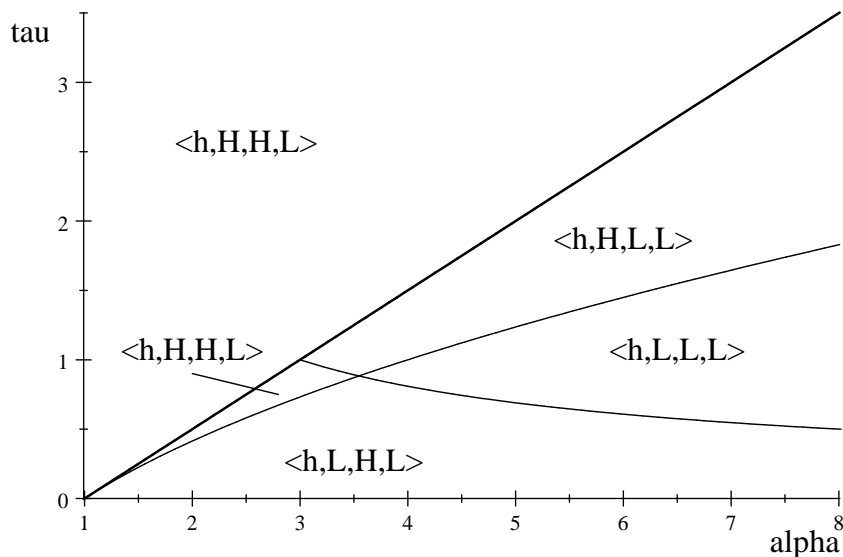


FIGURE A3. LEVEL-3 IN COORDINATION GAMES

$T_1$  believes that the opponent is truthful and therefore best responds to the received message, while sending random messages (not knowing what action will be taken).

$T_2$  faces truthful  $T_0$  and responding  $T_1$ . Since zero-step and one-step thinkers send both messages with equal probabilities,  $g_2(l|m) = g_2(l)$ . The strategy  $\langle l, H, H \rangle$  is clearly dominated by  $\langle h, H, H \rangle$  and  $\langle h, L, L \rangle$  is dominated by  $\langle h, H, L \rangle$ . The remaining strategies gives the following expected payoff:

$$\begin{aligned} \pi(\langle h, H, L \rangle) &= g_2(0) \frac{1}{2} (u_{LL} + u_{HH}) + g_2(1) \frac{1}{2} (u_{LH} + u_{HH}), \\ \pi(\langle h, H, H \rangle) &= g_2(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_2(1) u_{HH}, \\ \pi(\langle l, L, L \rangle) &= g_2(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2(1) u_{LL}, \\ \pi(\langle l, H, L \rangle) &= g_2(0) \frac{1}{2} (u_{LL} + u_{HH}) + g_2(1) \frac{1}{2} (u_{LL} + u_{HL}). \end{aligned}$$

First suppose that  $L$  is risk dominant. This implies that  $u_{LL} + u_{LH} > u_{HL} + u_{HH}$  and consequently  $u_{LH} > u_{HL}$  so that  $\langle h, H, L \rangle$  dominates  $\langle l, H, L \rangle$ .  $T_2$  prefers  $\langle h, H, H \rangle$  over  $\langle h, H, L \rangle$  whenever

$$\tau > (u_{LL} - u_{HL}) / (u_{HH} - u_{LH}) = \alpha,$$

and  $\langle l, L, L \rangle$  over  $\langle h, H, H \rangle$  whenever  $\tau < (\beta - 1)/2$ . Finally,  $T_2$  prefers  $\langle l, L, L \rangle$  over  $\langle h, H, L \rangle$  whenever

$$\tau > \frac{u_{HH} - u_{LH}}{2u_{LL} - u_{LH} - u_{HH}},$$

given that the right hand side is positive. Hence,  $\langle l, L, L \rangle$  is only optimal whenever  $0 < \lambda < \tau < (\beta - 1)/2$  and the payoffs satisfy  $0 < \lambda < (\beta - 1)/2$ .

Second, suppose that  $H$  is risk dominant and  $u_{LH} + u_{HH} > u_{LL} + u_{HL}$  so that  $\langle h, H, L \rangle$  dominates  $\langle l, H, L \rangle$  and  $\langle h, H, H \rangle$  dominates  $\langle l, L, L \rangle$ .  $T_2$  plays  $\langle h, H, H \rangle$  rather than  $\langle h, H, L \rangle$  if  $\tau > \alpha$ .

Finally, suppose that  $H$  is risk dominant and  $u_{LH} + u_{HH} < u_{LL} + u_{HL}$ . Now  $\langle l, H, L \rangle$  dominates  $\langle h, H, L \rangle$  and  $\langle h, H, H \rangle$  dominates  $\langle l, L, L \rangle$ . Therefore,  $T_2$  plays  $\langle h, H, H \rangle$  if  $\tau > (u_{LL} - u_{HL}) / (2u_{HH} - u_{LL} - u_{HL})$ .

Since  $\langle l, L, L \rangle$ ,  $\langle h, H, L \rangle$  and  $\langle h, H, H \rangle$  are best responses if the opponent plays the same strategies, by Lemma 1,  $T_{3+}$  play like  $T_2$  in these cases. Finally, consider  $T_3$  when  $T_2$  plays  $\langle l, H, L \rangle$ . In this case, whenever  $T_3$  receives an  $h$  message, he believes that it comes from a  $T_0$  or  $T_1$  player. Suppose first  $T_3$  receives the message  $h$ . If  $T_3$  sent  $h$ , then it is optimal to play  $H$ . If  $T_3$  sent the message  $l$ , the payoffs from playing  $L$  and  $H$  are

$$\begin{aligned}\pi(\langle l, L, \cdot \rangle | h) &= g_3(0)u_{LH} + g_3(1)u_{LL}, \\ \pi(\langle l, H, \cdot \rangle | h) &= g_3(0)u_{HH} + g_3(1)u_{HL}.\end{aligned}$$

Playing  $H$  is preferred whenever  $\tau < (u_{HH} - u_{LH}) / (u_{LL} - u_{HL}) = 1/\alpha$ . Since  $H$  is risk dominant,  $\alpha < 1$  and since  $\gamma < 1$ , this condition always hold. Now consider the case when  $T_3$  receives the message  $l$ . If  $T_3$  sent  $l$ , then it is optimal to play  $L$  (since  $T_0$  is truthful and  $T_1$  and  $T_2$  best-responds). Suppose that  $T_3$  sent  $h$ . Then expected payoffs are:

$$\begin{aligned}\pi(\langle h, \cdot, L \rangle | l) &= g_3(0|l)u_{LL} + g_3(1|l)u_{LH} + g_3(2|l)u_{LH} \\ \pi(\langle h, \cdot, H \rangle | l) &= g_3(0|l)u_{HL} + g_3(1|l)u_{HH} + g_3(2|l)u_{HH}\end{aligned}$$

Playing  $L$  is preferred whenever  $\tau(1 + \tau) < \alpha$ .

Which message will  $T_3$  send? Suppose first that  $\tau(1 + \tau) < \alpha$  so that  $T_3$  plays  $\langle h, H, L \rangle$  or  $\langle l, H, L \rangle$ . These strategies give the following ex ante payoffs

$$\begin{aligned}\pi(\langle h, H, L \rangle) &= g_3(0) \frac{1}{2}(u_{LL} + u_{HH}) + g_3(1) \frac{1}{2}(u_{LH} + u_{HH}) + g_3(2)u_{LH}, \\ \pi(\langle l, H, L \rangle) &= g_3(0) \frac{1}{2}(u_{LL} + u_{HH}) + g_3(1) \frac{1}{2}(u_{LL} + u_{HL}) + g_3(2)u_{LL}.\end{aligned}$$

It follows from the condition  $u_{LH} + u_{HH} < u_{LL} + u_{HL}$  that  $\langle l, H, L \rangle$  dominates  $\langle h, H, H \rangle$ . Now consider the case when  $\tau(1 + \tau) > \alpha$  so that  $T_3$  play either  $\langle h, H, H \rangle$  or  $\langle l, H, L \rangle$ . The payoff from each strategy is

$$\begin{aligned}\pi(\langle h, H, H \rangle) &= g_3(0) \frac{1}{2}(u_{HL} + u_{HH}) + g_3(1)u_{HH} + g_3(2)u_{HH}, \\ \pi(\langle l, H, L \rangle) &= g_3(0) \frac{1}{2}(u_{LL} + u_{HH}) + g_3(1) \frac{1}{2}(u_{LL} + u_{HL}) + g_3(2)u_{LL}.\end{aligned}$$

Sending  $h$  is preferred whenever  $\tau + \tau^2\delta\gamma > \gamma$ , i.e. when  $\tau > (\sqrt{4\gamma^2\delta + 1} - 1) / 2\gamma\delta$  (since  $\gamma > 0$  and  $\delta > 1$ ).

Note that two-way communication always entails play of the payoff dominant equilibrium in the Stag Hunt game depicted in Figure 2. For that particular game,  $\alpha = 7$  so that  $T_{2+}$  plays  $\langle h, H, L \rangle$  if  $\tau < 7$  and  $\langle h, H, H \rangle$  otherwise.

*Mixed motive games*

As before, we assume without loss of generality that  $u_{HL} > u_{LH}$  so that each player prefers the equilibrium where he is the one to play *H*(igh).

**OBSERVATION 11:** *(No communication) Let  $\theta = (u_{LH} - u_{HH}) / (u_{HL} - u_{LL})$ . If *H* is the risk dominant action,  $T_1$  plays *H*. If  $\tau < (1/\theta - 1)/2$ ,  $T_2$  plays *H*;  $T_3$  plays *H* if in addition  $\tau + \tau^2/2 < (1/\theta - 1)/2$ , but plays *L* if  $\tau + \tau^2/2 > (1/\theta - 1)/2$ . If  $\tau > (1/\theta - 1)/2$ ,  $T_2$  plays *L*;  $T_3$  plays *H* if in addition  $(2 - \tau/\theta)\tau < 1/\theta - 1$ , but plays *L* if  $(2 - \tau/\theta)\tau > 1/\theta - 1$ . If *L* is the risk dominant action,  $T_1$  plays *L*. If  $\tau < (\theta - 1)/2$ ,  $T_2$  plays *L*;  $T_3$  plays *L* if in addition  $\tau + \tau^2/2 < (\theta - 1)/2$ , but plays *H* if  $\tau + \tau^2/2 > (\theta - 1)/2$ . If  $\tau > (\theta - 1)/2$ ,  $T_2$  plays *H*;  $T_3$  plays *L* if in addition  $(2 - \tau\theta)\tau < \theta - 1$ , but plays *H* if  $(2 - \tau\theta)\tau > \theta - 1$ .*

**PROOF:**

First suppose *H* is risk dominant (which implies that  $\theta < 1$ ).  $T_2$  plays *H* rather than *L* if

$$g_2(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_2(1) u_{HH} > g_2(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2(1) u_{LH},$$

which is equivalent to  $1 + 2\tau < 1/\theta$ . Suppose this holds so that  $T_2$  plays *H*. Then  $T_3$  prefers *H* over *L* whenever

$$\begin{aligned} g_3(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_3(1) u_{HH} + g_3(2) u_{HH} \\ > g_3(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_3(1) u_{LH} + g_3(2) u_{LH}, \end{aligned}$$

which simplifies to  $1 + 2\tau + \tau^2 < 1/\theta$ . Suppose instead  $T_2$  plays *L*. Then  $T_3$  prefers *H* over *L* whenever

$$\begin{aligned} g_3(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_3(1) u_{HH} + g_3(2) u_{HL} \\ > g_3(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_3(1) u_{LH} + g_3(2) u_{LL}, \end{aligned}$$

which is equivalent to  $(2 - \tau/\theta)\tau < 1/\theta - 1$ .

Now suppose *L* is risk dominant. Then  $T_2$  plays *H* rather than *L* if

$$g_2(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_2(1) u_{HL} > g_2(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2(1) u_{LL},$$

which is equivalent to  $1 + 2\tau > \theta$ . Suppose that this holds so that  $T_2$  plays *H*. Then  $T_3$  prefers *H* over *L* whenever

$$\begin{aligned} g_3(0) \frac{1}{2} (u_{HL} + u_{HH}) + g_3(1) u_{HL} + g_3(2) u_{HH} \\ > g_3(0) \frac{1}{2} (u_{LL} + u_{LH}) + g_3(1) u_{LL} + g_3(2) u_{LH}, \end{aligned}$$

which simplifies to  $(2 - \tau\theta)\tau > \theta - 1$ . If  $T_2$  instead plays *L*, then  $T_3$  prefers *H* over *L* whenever  $\tau + \tau^2/2 > (\theta - 1)/2$ .

Note that some of the conditions above are quadratic, implying that they may be satisfied both for low and high values of  $\tau$ .

**OBSERVATION 12:** (*One-way communication*) If  $H$  is the risk dominant action, then  $T_{1+}$  sends  $h$  and plays  $H$  as sender and responds to messages as receiver. If  $L$  is the risk dominant action, then  $T_1$  sends  $l$  and plays  $L$  as sender and responds to messages as receiver. The behavior of  $T_{2+}$  depends on the payoff structure of the game:

**Case 1** ( $u_{LH} > u_{LL}$ ): Let  $\eta \equiv (u_{LL} + u_{LH} - u_{HL} - u_{HH}) / 2 (u_{HL} - u_{LH})$ . If  $\tau < \eta$ , then  $T_2$  plays  $\langle l, L, L, H \rangle$  and  $T_3$  plays  $\langle l, L, L, H \rangle$  if  $(1 + \tau/2) \tau < \eta$  whereas  $T_{3+}$  plays  $\langle h, H, L, H \rangle$  if  $(1 + \tau/2) \tau > \eta$ . If  $\tau > \eta$ , then  $T_{2+}$  plays  $\langle h, H, L, H \rangle$ .

**Case 2** ( $u_{LH} < u_{LL}$ ): If  $\tau < (\theta - 1) / 2$ , then  $T_2$  plays  $\langle h, L, L, H \rangle$  and  $T_3$  plays as follows:

$$\begin{aligned} T_3 \text{ plays } \langle h, L, L, H \rangle & \text{ if } (1 + \tau/2) \tau < (\theta - 1) / 2 \text{ and } \tau < \sqrt{\theta}, \\ T_{3+} \text{ plays } \langle h, L, L, L \rangle & \text{ if } (1 + \tau/2) \tau < (\theta - 1) / 2 \text{ and } \tau > \sqrt{\theta}, \\ T_{3+} \text{ plays } \langle h, H, L, H \rangle & \text{ if } (1 + \tau/2) \tau > (\theta - 1) / 2 \text{ and } \tau < \sqrt{\theta}, \\ T_{3+} \text{ plays } \langle h, H, L, L \rangle & \text{ if } (1 + \tau/2) \tau > (\theta - 1) / 2 \text{ and } \tau > \sqrt{\theta}. \end{aligned}$$

If  $\tau > (\theta - 1) / 2$ , then  $T_{2+}$  plays  $\langle h, H, L, H \rangle$ .

**PROOF:**

First let  $H$  be the risk dominant action. A  $T_1$  sender faces a randomizing receiver and therefore plays  $H$  and sends  $h$ . A  $T_1$  receiver, on the other hand, responds to the sent message, believing it comes from a truthful  $T_0$  opponent. By Lemma 1,  $T_{2+}$  plays the same strategy as  $T_1$ .

If instead  $L$  is the risk dominant action, a  $T_1$  sender instead sends and plays  $L$ , but responds to messages as receiver. A  $T_2$  sender faces a tradeoff between playing  $L$  (the best response against  $T_0$ ) and sending  $h$  and playing  $H$  (the best response against  $T_1$ ). The expected payoffs from the three relevant sender strategies are:

$$\begin{aligned} \pi (\langle l, L \rangle) &= g_2 (0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2 (1) u_{LH}, \\ \pi (\langle h, H \rangle) &= g_2 (0) \frac{1}{2} (u_{HL} + u_{HH}) + g_2 (1) u_{HL}, \\ \pi (\langle h, L \rangle) &= g_2 (0) \frac{1}{2} (u_{LL} + u_{LH}) + g_2 (1) u_{LL}. \end{aligned}$$

Suppose  $u_{LH} > u_{LL}$  so that  $\langle l, L \rangle$  dominates  $\langle h, L \rangle$ . Then a  $T_2$  sender plays  $\langle l, L \rangle$  if

$$\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2 (u_{HL} - u_{LH})} = \eta,$$

but plays  $\langle h, H \rangle$  otherwise.  $T_2$  receivers face truthful  $T_0$  and  $T_1$  senders, so they respond to messages. If  $\tau$  is above the threshold above,  $T_{2+}$  play  $\langle h, H, L, H \rangle$ . However, if  $\tau$  is below the threshold,  $T_3$  senders trade off truthfully playing  $L$  or  $H$ . They play  $L$  if  $(1 + \tau/2) \tau < \eta$  and otherwise play  $H$ .

Suppose now that  $u_{LL} > u_{LH}$  so that  $T_2$  senders prefer sending  $h$  when they intend to play  $L$ . They prefer doing so over  $\langle h, H \rangle$  whenever

$$\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2 (u_{HL} - u_{LL})} = (\theta - 1) / 2.$$

$T_2$  receivers face truthful senders, so they respond to messages. If  $\tau > (\theta - 1) / 2$ ,  $T_{2+}$  plays  $\langle h, H, L, H \rangle$ . A  $T_3$  sender plays  $\langle h, L \rangle$  rather than  $\langle h, H \rangle$  if  $(1 + \tau/2) \tau < (\theta - 1) / 2$ .

A  $T_3$  receiver believes that an  $l$  message is truthful, so they play  $H$  in that case. An  $h$  message comes either from a  $T_0$  or  $T_3$ . When receiving a  $h$  message, the payoff from each action is:

$$\begin{aligned}\pi(H|h) &= g_3(0|h)u_{HH} + g_3(2|h)u_{HL}, \\ \pi(L|h) &= g_3(0|h)u_{LH} + g_3(2|h)u_{LL}.\end{aligned}$$

So,  $T_3$  play  $\langle L, H \rangle$  if

$$\tau < \sqrt{\frac{u_{LH} - u_{HH}}{u_{HL} - u_{LL}}} = \sqrt{\theta},$$

and play  $\langle H, H \rangle$  otherwise.

Two-way communication in mixed motive games is particularly cumbersome to characterize generally. The following observation therefore focuses on the behavior of  $T_2$ . (For a particular payoff configuration, however, it is straightforward to derive the behavior of  $T_{3+}$  players.)

**OBSERVATION 13:** (*Two-way communication*)  $T_1$  sends  $h$  and  $l$  with equal probabilities and responds to messages. Let  $v \equiv (u_{HL} - u_{LL}) / (2u_{LH} - u_{LL} - u_{HL})$ . If  $L$  is risk dominant and  $0 < v < \eta$ , then  $T_2$  plays  $\langle h, L, H \rangle$  if  $\tau < v$ ,  $\langle l, L, L \rangle$  if  $v < \tau < \eta$ , and  $\langle h, H, H \rangle$  if  $\tau > \eta$ . If either  $v < 0$  or  $0 < \eta < v$ , then  $T_2$  plays  $\langle h, L, H \rangle$  if  $\tau < \theta$  and  $\langle h, H, H \rangle$  if  $\tau > \theta$ . If  $H$  is risk dominant and  $u_{LH} + u_{HH} > u_{LL} + u_{HL}$ , then  $T_2$  plays  $\langle l, L, H \rangle$  if  $\tau < (u_{LH} - u_{HH}) / (2u_{HL} - u_{LH} - u_{HH})$  and  $\langle h, H, H \rangle$  otherwise. If instead  $u_{LH} + u_{HH} < u_{LL} + u_{HL}$ ,  $T_2$  plays  $\langle h, L, H \rangle$  if  $\tau < \theta$  and  $\langle h, H, H \rangle$  if  $\tau > \theta$ .

**PROOF:**

$T_1$  believes that the opponent is truthful and therefore sends random messages, but responds to the received message. The strategy  $\langle l, H, H \rangle$  is dominated by  $\langle h, H, H \rangle$  and the expected payoff for  $T_2$ 's other strategies are:

$$\begin{aligned}\pi(\langle h, L, L \rangle) &= g_2(0) \frac{1}{2}(u_{LL} + u_{LH}) + g_2(1)u_{LL}, \\ \pi(\langle h, L, H \rangle) &= g_2(0) \frac{1}{2}(u_{HL} + u_{LH}) + g_2(1) \frac{1}{2}(u_{LL} + u_{HL}), \\ \pi(\langle h, H, H \rangle) &= g_2(0) \frac{1}{2}(u_{HL} + u_{HH}) + g_2(1)u_{HL}, \\ \pi(\langle l, L, H \rangle) &= g_2(0) \frac{1}{2}(u_{HL} + u_{LH}) + g_2(1) \frac{1}{2}(u_{LH} + u_{HH}), \\ \pi(\langle l, L, L \rangle) &= g_2(0) \frac{1}{2}(u_{LL} + u_{LH}) + g_2(1)u_{LH}.\end{aligned}$$

Suppose  $H$  is risk dominant. Then  $\langle h, H, H \rangle$  dominates  $\langle l, L, L \rangle$  and  $\langle h, L, L \rangle$ . First suppose that  $u_{LH} + u_{HH} > u_{LL} + u_{HL}$  so that  $\langle l, L, H \rangle$  dominates  $\langle h, L, H \rangle$ .  $T_2$  plays  $\langle h, H, H \rangle$  rather than  $\langle l, L, H \rangle$  if

$$\tau > \frac{u_{LH} - u_{HH}}{2u_{HL} - u_{LH} - u_{HH}}.$$

If instead  $u_{LL} + u_{HL} > u_{LH} + u_{HH}$ , then  $T_2$  plays  $\langle h, H, H \rangle$  rather than  $\langle h, L, H \rangle$  if

$$\tau > \frac{u_{LH} - u_{HH}}{u_{HL} - u_{LL}} = \theta.$$

Now consider the case when  $L$  is risk dominant. This implies that  $u_{LL} > u_{HH}$ , so  $\langle h, L, H \rangle$  dominates  $\langle l, L, H \rangle$  and  $\langle h, L, H \rangle$  dominates  $\langle h, L, L \rangle$ . There are three remaining strategies to consider.  $T_2$  plays  $\langle h, H, H \rangle$  rather than  $\langle h, L, H \rangle$  if  $\tau > \theta$ .  $T_2$  may prefer to play  $\langle l, L, L \rangle$ .  $\langle l, L, L \rangle$  preferred over  $\langle h, L, H \rangle$  whenever

$$\tau > \frac{u_{HL} - u_{LL}}{2u_{LH} - u_{LL} - u_{HL}} = \nu,$$

given that the right hand side is positive (otherwise the condition cannot hold).  $\langle l, L, L \rangle$  is preferred over  $\langle h, H, H \rangle$  whenever

$$\tau < \frac{(u_{LL} + u_{LH}) - (u_{HL} + u_{HH})}{2(u_{HL} - u_{LH})} = \eta.$$

Hence, in order for  $\langle l, L, L \rangle$  to be optimal,  $\tau$  must be between  $\nu$  and  $\eta$  and the payoffs must satisfy  $\eta > \nu > 0$ .