Contents

Chapter 5

Trust Games Generalized

Altruism is for those who can't endure their own desires.
–Stephen Dunn

This chapter provides a more detailed look into the central model of this dissertation: *Trust Games* [Bicchieri, 2006]. As mentioned in our chapter on cooperation (**??**) , we are interested in norms and how they relate to politeness. One such norm is trustworthiness, and we model here how it can be activated by initiating a costly exchange with a trust game. Trust games give us a baseline for the type of incentives seen in requests and other speech acts where the face threat to the speaker and hearer can vary.

We then generalize and extend these results to give us constraints and incentives for a wider range of behavior. Although the normalized values in the original trust game reflected a preference order based on what outcomes would be most favorable to each agent, we would like to expand on these intuitions and even provide cases where the conflict between incentives can provide interesting behavior.

5.1 Costly Signaling vs. Reciprocal Exchange

Before we move into our exposition on trust games being the appropriate model for speech acts like requests, we should clarify why the thus far dominant paradigm in game-theoretic pragmatics, signaling games, is not as critical to our analysis for the moment. Part of this is rooted in understanding the phenomena of costly signaling and reciprocity, as outlined in papers like BliegeBird et al. [2005]. According to this exposition, there are two basic forms of costly signaling: wasteful or conspicuous consumption and exchange of symbolic capital.

In our chapter on cooperation, **??**, we discussed the first notable treatment of politeness from a game-theoretic perspective, van Rooij [2003], which used ideas similar to van Rooij [2004] [1] in that the longer formulations of polite requests may indicate a costly message parallel to the phenomenon of the *Peacock's Tail*, a reproductive strategy outlined via the Zahavian Handicap Principle. [Zahavi, 1975] I.e. just as costly, wasteful biological signals lead to the emergence and

[1] These are the same author.

evolution of honest communication between groups with conflicting preferences on selection, so should costly language like politeness indicate a form of honesty.

While this approach has merit, it does not capture the implicit exchange-based payoff structure underlying a request, the speech act we claim to be the canonical driver of polite speech. Why not? First, for a request to succeed, there must be sufficient incentive for the requester to grant the favor. Assuming the speaker and hearer are not in a dominance relationship, the speaker must acknowledge the potential autonomy/ need for acceptance of the hearer in said request, otherwise we would have an imperative [Clark and Carlson, 1982] . This is typically done through some sort of face payment [Goffman, 1955]. It is not done by using speech as a marker of fitness, but rather as a medium of exchange.

Second, when speakers make requests, the exchange is one of understood reciprocity. This implies that the rewards are levied not through selection, but rather through the payoffs in the game itself and the maintenance of a repeated interaction or reciprocal society. Third, the Handicap Principle can also explain impolite speech used among groups with selection criteria geared toward counterculture movements. For instance, learning slang can itself be a costly endeavor used to promote in-group identity [Wilson et al., 2000]. Its ability to seemingly account for both politeness and impoliteness suggests that costly speech (e.g. acquiring elaborate or specialized jargon and turns of phrase) does play a role in group selection, but that this phenomenon is orthogonal to making a request by paying someone face. As Nowak [2006] details that the mechanisms enabling cooperation fall under selection and reciprocity, we turn from the selection-oriented Handicap Principle to games focused on reciprocal behavior, i.e. those of trust and exchange introduced in **??**.

Vanilla Trust Game

To recall the notion from before, trust games, seen in Figure 5.1 depict a scenario where Player X has an initial option to defer to Player Y for a potentially larger payoff for both. However, similar to the Prisoner's Dilemma, Player Y could defect on Player X and keep more for himself.

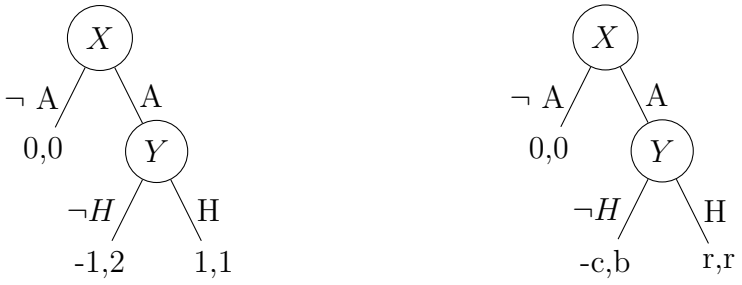


Figure 5.1: Normalized and generalized trust games. Utilities are U_X, U_Y.

Motivation and Utility Structure

Let us address the reasons for adopting this game, including its equilibrium path and payoff profiles. Requests necessarily involves asymmetry in action; our model therefore depicts the person asking for help as surrendering control to the other player to resolve the game. Further, this model incorporates the conflict between asking for help and risking a loss of face or opportunity, in accordance with the theory of social exchange posited by Homans [1958]. Last, the game should be amenable to modifications like repetition or additions to its payoff structure, which we explore in this chapter. To address the payoff structure, we consider each action profile at a time.

- $\neg A_X$: X not asking for help simply means leaving the status quo in place. $(0, 0)$

- A_X, H_Y: If Y helps X, then they arrive at a mutually beneficial outcome. This does cost Y time and energy, and thus his utility is not as high as if he defects. $(1, 1) \Rightarrow (r, r)$

- $A_X, \neg H_Y$: When X asks Y for help via a polite request, Y should experience an increase in face. However, Y does not have to deliver on the request, which saves him time/energy. If Y does not help X, he has lost positive face, the opportunity to ask someone else, and time/ energy on the request. $(-1, 2) \Rightarrow (-c, b)$

To connect these this situation to the tree from Brown and Levinson [1987], observe that just as there are a multitude of linguistic variations for making a request, so are there multiple trust game interpretations. We can depict the face-payment/ face-threats in the requests below in Figure 5.2, seen with escalating politeness markers and their accompanying payoffs.

a. *Get that now you imbecile!*

b. *Can you get that?*

c. *Pardon me, sir. Can you help me?*

d. *Pardon me, sir. Can you please help me?*

In general, we will stick with a slightly simpler payoff scheme. The payoffs are only meant to reflect relative preferences and not actual monetary or material outcomes. Proceeding via backward induction in all of these games , as Y moves last, we see that for a one-shot game, the act of trust will not occur for a rational Player X in equilibrium. We can solve this by making the roles symmetric and repeating the game.

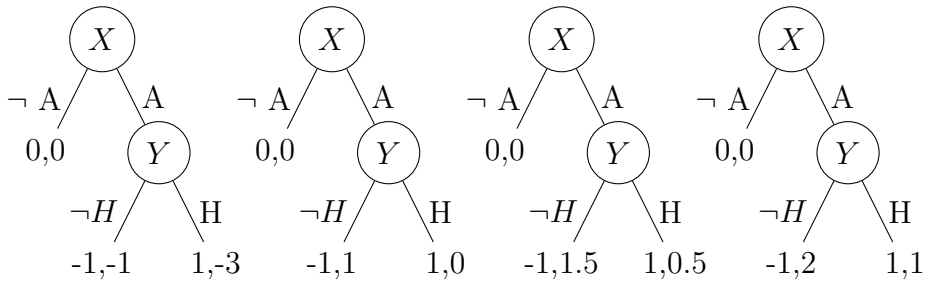


Figure 5.2: Trust Games with various payoffs mirroring the utterances above. As the utterances include more face payment, we see an increase in the utility for Y. This also assumes a cost–neutral set of messages for X.

5.2 Trust Games Generalized and Symmetrized

Trust games share qualities with the Prisoner's Dilemma but with an asymmetric and sequential structure. This fits our notion of requests as requiring cooperative behavior from speakers with different roles in a single instance. In a larger population, however, interactions that occur can be generalized with frequently occurring similar roles.

Extending the work done in political science and economics, much of the work on the Prisoner's Dilemma in evolutionary biology models the interaction as one with a cost to cooperation and a benefit derived from the other's cooperation. As seen in [Nowak, 2006], we can display the Prisoner's Dilemma as seen in Table 5.1.

	C	D
C	3,3	-2,5
D	5,-2	0,0

	C	D
C	b-c,b-c	-c,b
D	b,-c	0,0

Table 5.1: Variants of the Prisoner's Dilemma with sample payoffs and formalized parameters on costs and benefits.

Although we will not delve too deeply into evolutionary dynamics yet, to take advantage of the population-level results from economics and biology, we will symmetrize the trust game in the following section. This is done because:

- Speakers do not act as merely senders of information; they also receive it.
- In parallel, speakers both give respect/admiration (face) and receive it.
- Both linguistic conventions and societal norms require population-wide acceptance to take hold.
- Within scenarios like the request-help model, we have two sorts of speech acts that require different roles. Speakers act in both capacities in repeated interaction.

Figure 5.3: Trust Games G_1 and G_2 with roles reversed.

Symmetrizing

In a large population, we can then think of two players X and Y playing the game with roles potentially reversed. To construct this game table from our basic example in Figure 5.3, we let each player occupy the different roles, constructing four types T_i for each combination i of the asymmetric game roles with the set of types : $T = \{\ AH\ ,\ \neg AH\ ,\ A\neg H\ ,\ \neg A\neg H\ \}$ We then add the utilities received by the player in both roles interacting with another type to arrive at his final utility. If we consider these as games G_1 and G_2 with utilities U_1 and U_2 we have

$$U_X(T_X, T_Y) = U_1(T_X, T_Y) + U_2(T_Y, T_X)$$

This is the utility of Player X playing as after acting as Sender and Receiver. The process for Y is identical, as shown below. This results in Table 5.2. For example, consider the play where X is type AH and Y is type $A\neg H$. For each player we have:

$$U_X(AH_X, A\neg H_Y) = U_1(A_X, \neg H_Y) + U_2(A_Y, H_X) = -1 + 1 = 0$$

$$U_Y(AH_X, A\neg H_Y) = U_1(A_X, \neg H_Y) + U_2(A_Y, H_X) = 2 + 1 = 3$$

In G_1 the score is $(-1, 2)$ and in G_2 the score is $(1, 1)$. This gives us a final score of $(-1 + 1, 2 + 1)$ or $(0, 3)$. We can imagine a pair of hypothetical conversations as such:

Game 1: Agent X is type AH and Y is type $A\neg H$

· **Xavier:** I'd really appreciate it if you could help me out.

· **Yvonne:** I can't.

Game 2: Agent X is type AH and Y is type $A\neg H$

· **Yvonne:** I'd like you to help me with this. Could you?

· **Xavier:** No problem. What do you need?

	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2,2	1,1	0,3	-1, 2
$\neg AH$	1,1	0,0	1,1	0,0
$A\neg H$	3,0	1,1	1,1	-1, 2
$\neg A\neg H$	2,-1	0,0	2,-1	$\boxed{0,0}$

Table 5.2: Symmetric Trust Game. Nash Equilibrium boxed

General Payoffs

To generalize the basic example, we posit a cost c for asking (A) and a benefit b for not helping ($\neg H$). Further, we posit a symmetric reward (r, r) for helping (H) and being helped. [2] . We will assume in general that $b > r$, i.e. the *benefit* of not helping is greater than the mutual *reward* for helping. Just as before, we combine two games, where each player occupies the different roles.

Figure 5.4: Generalized Trust Games G_1 and G_2 with roles reversed.

The process is the same for the trust games in the general form as it was previously. For instance, if we want to find the entry of a player X of type $A\neg H$ against a player Y of type $\neg A\neg H$:

$$U_X(A\neg H_X, \neg A\neg H_Y) = U_1(A_X, \neg H_Y) + U_2(\neg A_Y, \neg H_X) = -c + 0 = -c$$

$$U_Y(A\neg H_X, \neg A\neg H_Y) = U_1(A_X, \neg H_Y) + U_2(\neg A_Y, \neg H_X) = b + 0 = b$$

Another way to construct the final payoff profile is to add the components of the respective payoff profiles. In G_1 we have the outcome of $(-c, b)$ and in G_2 we have the outcome of $(0, 0)$. Adding these outcomes in each component gives us the appropriate payoff profile of $(-c, b)$. For every other strategy combination, we perform a similar computation. Performing these computations allows us to construct the table in Table 5.3.

[2] The rewards could be different: (r_1, r_2) or we could take $r \approx b - c$ where $r < b$

	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r,r	r-c, r+b	-c, b
$\neg AH$	r,r	0,0	r,r	0,0
$A\neg H$	r+b,r-c	r,r	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0	b,-c	$\boxed{0,0}$

Table 5.3: Symmetric Trust Game Generalized. Nash Equilibrium boxed.

Player Types

For a single player, we can think of the various types like AH and $A\neg H$ as prescribed behavior depending on the given context of asking for help or being asked. For a population of speakers, we can think of player types or the behavior itself. These dual notions can be equally useful depending on our goals. We will follow these descriptions for the following player types. Our primary concern will be finding conditions to promote the *Friendly* type.

AH:Friendly $\neg AH$:Altruistic $A\neg H$:Insistent $\neg A\neg H$:Reclusive

One-Shot Equilibrium Analysis

We have a Nash Equilibrium at the strategy profile $(\neg A\neg H, \neg A\neg H)$, shown boxed in the tables. This fits the backward induction found in the asymmetric game, as the second player will have an incentive not to help, and thus the first player will have incentive not to ask. To see this in the symmetric game, we will iteratively eliminate (weakly or strictly) dominated strategies in the general game. We will do this for the general game, as the game with concrete values works out equivalently.

Although there are no strictly dominated strategies, we can perform an iterated elimination of weakly dominated strategies. Note that along every profile, the strategy $\neg A\neg H$ dominates $\neg AH$ since we posited that $b > r$. Eliminating it gives us the game in Table 5.4. Now, we perform the same again, observing that $A\neg H$ weakly dominates AH again under the constraint that $r < b$ and thus both $r + b > 2r$ and $b - c > r - c$. This leaves us with the game in Table 5.5 .

	AH	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r-c, r+b	-c, b
$A\neg H$	r+b,r-c	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	b,-c	0,0

Table 5.4: Altruistic $\neg AH$ eliminated.

This game is exactly our version of the Prisoner's Dilemma from before in Table 5.1, and it surprisingly gives us a model of interaction with trustworthiness ruled out. A further analysis of the game shows that not helping weakly dominates helping in the initial stage, and as we eliminate strategies, not asking strictly

dominates asking, giving the Prisoner's Dilemma in the final stage (Table 5.5). This supports the analysis [Bicchieri, 2006, Bicchieri and Muldoon, 2011] that trustworthiness, i.e. taking the action H of helping, is a norm that must be enforced, whereas trusting itself, i.e. taking the action A is not a norm, and we see that trust varies over the iterated elimination of weakly dominated strategies.

	$A\neg H$	$\neg A\neg H$
$A\neg H$	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0

Table 5.5: Resulting Prisoner's Dilemma when we eliminate two iterations of weakly dominated strategies.Nash Equilibrium boxed.

Mixed Strategy Equilibrium

After calculating the pure strategy Nash Equilibrium another question arises: is there a strictly mixed strategy Nash Equilibrium that yields non-trivial payoffs? We claim there is not. As the game is symmetric, consider the matrix form for Player X given here. We put here the actions for X on the left, and above, we place the probabilities that Y will play each strategy. Note that $a+b+c+d=1$, so in the calculations for $EU_x()$ we will use $d=(1-a-b-c)$

$$M_X = \begin{array}{c} \\ AH \\ \neg AH \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 3 & 1 & 1 & -1 \\ 2 & 0 & 2 & 0 \end{bmatrix} \end{array};$$

$$\begin{aligned} EU_x(AH) =& 2a+b-1(1-a-b-c) \\ EU_x(\neg AH) =& a+c \\ EU_x(A\neg H) =& 3a+b+c-1(1-a-b-c) \\ EU_x(\neg A\neg H) =& 2a+2c \end{aligned}$$

This equilibrium will exist when X is indifferent to his own strategies for variations in what Y does.

$$EU_x(\neg A\neg H) = EU_x(\neg AH) \Rightarrow 2a+2c = a+c \Rightarrow a=-c$$

As these are probabilities, this only occurs when $a=c=0$. Thus we have:

$$
\begin{aligned}
EU_x(AH) =& EU_x(A\neg H) \Rightarrow \\
2a + b - 1(1 - a - b - c) =& 3a + b + c - 1(1 - a - b - c) \Rightarrow \\
b - 1(1 - b) =& b - 1(1 - b)
\end{aligned}
$$

which is true for all values of b. Notice that this reduces the game to the degenerate game of only receiving 0, as the outcomes involving b and d yield payoffs of 0 when paired up. These strategies are also both of the ones that produce no interaction, as they both include not asking. Thus there are no non-trivial strictly mixed equilibria in the basic game with concrete payoffs.

Pareto-Efficient Equilibria and Symmetric Games

As we can see above, the Symmetric Trust Game contains strategies similar to the Prisoner's Dilemma with respect to the dominance of its least efficient, non-cooperative strategies. We would like to derive conditions under which the perverse incentives of this game give way to those favoring mutual reciprocity and coordinated action, much like the Stag Hunt. In our section on repeated games, we saw how a repeated Prisoner's Dilemma can transform into a game resembling the Stag Hunt [Skyrms, 2004, Aumann, 1990]. To recall the differences among the games, let us make the following remarks on Table 5.6:

	C	D
C	P,P	Q,R
D	R,Q	S,S

PD	C	D
C	3,3	-2,5
D	5,-2	0,0

SH	C	D
C	4,4	0,2
D	2,0	1,1

Table 5.6: 2×2 symmetric game vs. Prisoner's Dilemma and Stag Hunt.

We can observe in the tables that the primary obstacle to cooperation is that defection strictly dominates cooperation. The Prisoner's Dilemma has asymmetric strategy profiles contain strategies for one player that are preferential to the mutually beneficial cooperative profile ($R > P$). Further, for the other player they contain the least preferable strategy($Q < S$). We would like the first inequality to be false. On the general level, we have in the Prisoner's Dilemma that

$$Q < S < P < R$$

What we would like is a game more like the Stag Hunt where the cooperative outcome is also the best for both players, despite the fact that other equilibria exist. This will require additional mechanisms. In general, we want outcomes that resemble:

$$Q < S < R < P$$

I.e. we want the most efficient, cooperative outcomes to be equilibria. Our foray into repeated games and deriving their payoff matrices will resemble this goal. After that, we will revisit the one-shot scenario when we eliminate certain strategies from play. Going forward, we wish to examine the formation of *norms* like politeness and trustworthiness as supported by repeated game strategies. That will give us payoff tables like the Stag Hunt with multiple equilibria, the basis for the conventionalization process.

5.3 Remarks on the Folk Theorem(s)

Despite the pessimistic picture given by the Nash Equilibrium in the one-shot scenario, we can take advantage of repeated scenarios to bring about more efficient equilibria in long-term players, due to the multitude of strategies in the repeated Symmetric Trust Game(STG), as seen in the strategy space from Figure 5.5.

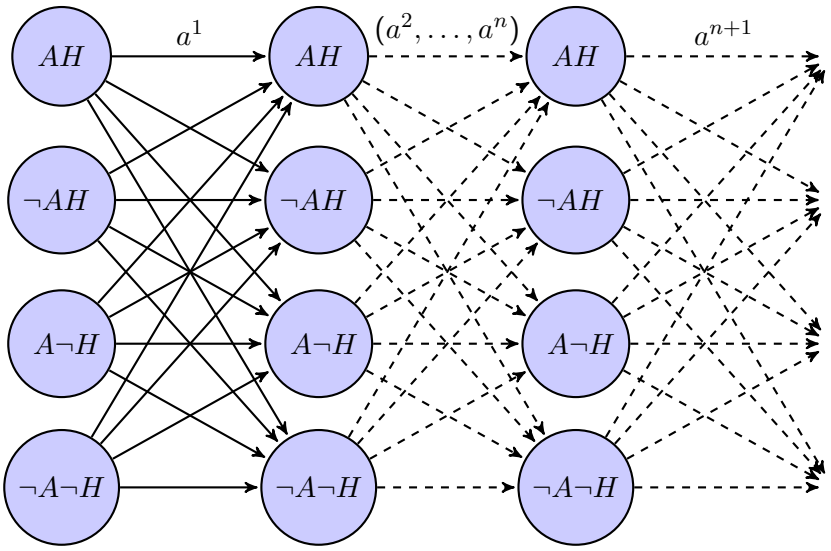


Figure 5.5: Markov Chain depicting infinite repetition of the STG

The previous work done on the repeated Prisoner's Dilemma gives us a model to follow, and in the coming sections, we will apply strategies similar to *Tit-for-Tat* and *Grim Trigger*. A few questions now are

- How do we enforce the cooperative outcomes?
- How do payoff parameters affect the long-term play of the game?
- How do the strategies in the repeated Prisoner's Dilemma gives us a baseline for this game?

When we look back at our symmetric trust game in Table 5.3, we want to investigate how various conditions like repetition could lead to the emergence of profiles predicted by the Folk Theorem, with an example from the normalized game with concrete payoffs given Table 5.7. Two major points will follow every analysis when discussing the folk theorem. The outcomes must be

	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2,2	1,1	0,3	-1, 2
$\neg AH$	1,1	0,0	1,1	0,0
$A\neg H$	3,0	1,1	1,1	-1, 2
$\neg A\neg H$	2,-1	0,0	2,-1	0,0

Table 5.7: Feasible pure strategy profiles amenable to the folk theorem(boxed); profiles up to (but not including) the circled outcomes also feasible.

- *Socially Feasible* in that they fall within the space of the one-shot outcomes;
- *Individually Rational* in that they outperform the min-max profile of the one-shot game for each player; $(0, 0)$ in our case.

The parameters in our stage game could alter the feasible outcomes in the repeated version. This we can see in both the payoff tables and the graphs of the feasible regions to come. We will primarily focus on the table with general values in Table 5.3, as the initial Symmetric Trust Game with concrete values is an example of the case where $c \leq r < b$. When considering this game, we see two pertinent cases. These revolve around whether $r > c$ or $r < c$.

Initially, we only considered values where $r > c$, but we can plausibly entertain scenarios where there is a high cost to the speech act and a low reward for helping. [Blum-Kulka et al., 1989]. These might include scenarios like those documented in Bohns and Flynn [2010]:

- Asking for use of a car,
- Asking to borrow a phone,
- Asking for help with a term paper, etc.

A third case, where $b < c$ is entertained in the tables, but not in the later analysis. Now we proceed on these grounds, having the various cases subject to the folk theorem highlighted in Table 5.8, Table 5.9, and Table 5.10.

From looking at the first two tables, we can see that the action $\neg A\neg H$ is clearly the worst one for the opponent to take. That said, the best way for a player to minimize his maximum loss in this case is to choose $\neg A\neg H$ as well. We thus have a minmax profile at the point where the game also has a Nash Equilibrium, and we see therefore why payoffs less than or equal to zero should be excluded from applications of the folk theorem towards finding long-term equilibria.

Case I: $c < r < b$

When conditioning on $r > c$, as seen in Table 5.8, we have that $r - c > 0$, and thus we do have an equilbrium path to outcomes that would be generated by profiles like $(A\neg H, AH)$ in the one-shot game. We therefore arrive at feasibility

$c < r < b$	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r,r	r-c, r+b	-c, b
$\neg AH$	r,r	0,0	r,r	0,0
$A\neg H$	r+b,r-c	r,r	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0	b,-c	0,0

Table 5.8: Case I: $c < r < b$.

region where the pure strategy profiles of $(A\neg H, \neg AH)$ and $(\neg AH, A\neg H)$ are now accessible. This also means that $b - c > 0$ as $b > r$. Thus the symmetric strategy profile $(A\neg H, A\neg H)$ is feasible. In the figure for the feasibility regions in Figure 5.7, we have an example of this as $c = 1$, $b = 3$, and $r = 2$

Case II: $r < c < b$

This feasibility regions strongly resembles that for the Prisoner's Dilemma, but we should remark that we still will only consider the cases where $U_X > 0$ and $U_Y > 0$. We see that $b - c > 0$ as $b > c$. Thus the symmetric strategy profile $(A\neg H, A\neg H)$ is feasible. As $r < c$ we have that $r - c < 0$ and thus we cannot include the pure strategy profiles of $(A\neg H, \neg AH)$ and $(\neg AH, A\neg H)$. In the figure for the feasibility regions in Figure 5.7, we have an example of this as $c = 2$, $b = 3$, and $r = 1$.

$r < c < b$	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r,r	r-c, r+b	-c, b
$\neg AH$	r,r	0,0	r,r	0,0
$A\neg H$	r+b,r-c	r,r	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0	b,-c	0,0

Table 5.9: Case II: $r < c < b$

Case III: $r < b < c$

This case is where there is a high cost to the message or signal of trust. Notice in this case the only surviving feasible outcome that is also a symmetric profile is the cooperative profile of (AH, AH). As before in Case II, we have that $r < c$ and therefore $r - c < 0$. Thus we cannot include the pure strategy profiles of $(A\neg H, \neg AH)$ and $(\neg AH, A\neg H)$. In the figure for the feasibility regions in Figure 5.7, we have an example of this as $c = 3$, $b = 2$, and $r = 1$.

Feasibile Regions

We now visualize the space of feasible outcomes amenable to the folk theorem. Although this may seem like a subtle point, the regions we will depict in these

$r < b < c$	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r,r	r-c, r+b	-c, b
$\neg AH$	r,r	0,0	r,r	0,0
$A\neg H$	r+b,r-c	r,r	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0	b,-c	0,0

Table 5.10: Case III: $r < b < c$.

diagrams will truly only exist for high discount values, as some outcomes require extremely patient players depending on the payoffs and strategies in the stage game [Mailath and Samuelson, 2006]. This is why we will feature calculations later in this chapter that will determine which patience levels sustain the various strategies. As mentioned, there are two primary conditions on the *feasibility* of payoffs amenable to the **Folk Theorem.** First, they must be within the convex hull of the stage game's payoffs. Second, each component must be strictly better than the payoffs in the joint minmax profile. In our case, this will be $(0, 0)$ unless otherwise noted.

If we look at the region for the variant Prisoner's Dilemma in Figure 5.6, we see that the region looks strikingly similar to the regions we see in Figure 5.7. This should not surprise us, as it reminds of the link between these two games and their similar strategy spaces

	C	D
C	5,5	-1,7
D	7,-1	1,1

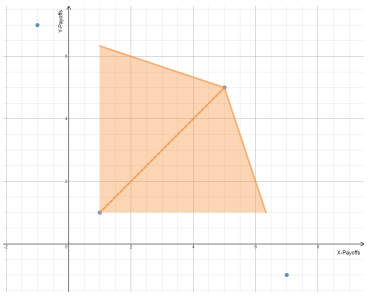
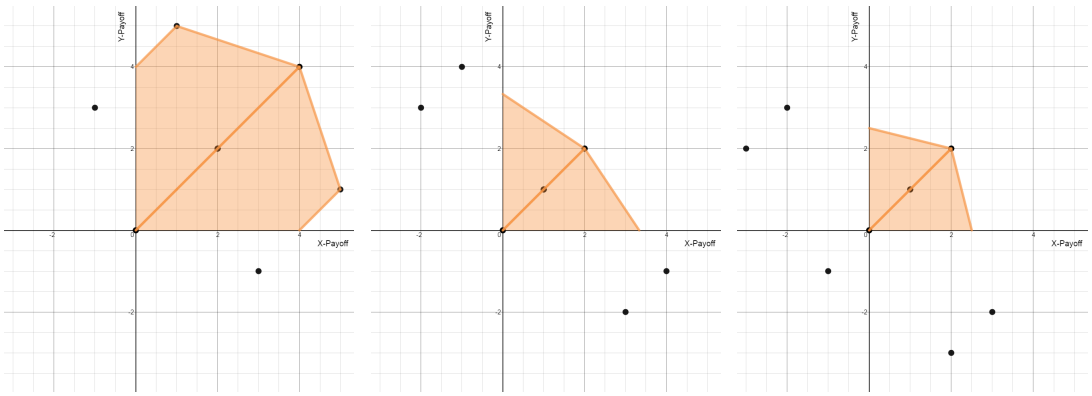


Figure 5.6: Feasible equilibria for the Prisoner's Dilemma. The payoffs for agent X are along the $x - axis$; likewise with Y. We do **NOT** include the boundaries $x = 1$ and $y = 1$ from the minmax profile $(1, 1)$.

In addition the tables, we can also look at the feasible regions as we change the parameters, seen in Figure 5.7. With some manipulation of the parameter c, we get that long-run outcomes giving $(A\neg H, AH)$ disappear, as $c > r$, and thus the outcome $(r + b, r - c)$ will not be *individually rational*, as $r - c < 0$. If $c > b$, as in Table 5.10, we will also eliminate outcomes with $b - c$ in the payoffs, as $b - c < 0$. As the cost of asking c increases, we see a smaller region forming, as the relative benefits of various outcomes decreases. This could make it easier to enforce the cooperative outcomes, as the long-run penalty for being unhelpful increases in relative scale. Since the action $\neg A\neg H$ produces zero or negative payoffs for each possible strategy against it and since for each parameter $r, b, c > 0$, the minmax profile will always be $(\neg A\neg H, \neg A\neg H)$ for our game. This means any feasible set



(a) Case I:$c < r < b$ **(b)** Case II:$r < c < b$ **(c)** Case III:$r < b < c$

Figure 5.7: Three cases for c on the same scale. As the value of c increases, less efficient outcomes are available as the payoffs approach $(0, 0)$. The payoffs for agent X are along the $x - axis$; likewise with Y. We do **NOT** include the boundaries $x = 0$ and $y = 0$.

of normalized payoffs must exist in the first quadrant of our graph where $U_X > 0$ and $U_Y > 0$.

As we look through these cases, we see a progression of the values of the message cost c on one hand and a contraction of the space of feasible payoffs. This contraction invariably leads to the cooperative outcome of (AH, AH) being the only surviving symmetric profile in the set of feasible equilibria in the repeated game. This is encouraging in the sense that we have a link now between message cost and provision of the cooperative outcome. In some ways this is disconcerting however, as it might be hard to imagine the cost of a speech act being greater than the reward for getting help.

Strategies in the Repeated Symmetric Trust Game

One of our primary aims is to account for the stability of politeness in a rational population, despite its apparent inadequacy in one-shot scenarios. We therefore want conditions by which various norms of behavior are symmetric equilibria of a repeated interaction. We will focus now on the profiles along the diagonal of our matrix, those where both conversants would be following the same norms.

We will ignore two types of outcomes however: asymmetric profiles and cases of high reward. Although we can derive from the folk theorem that there are asymmetric equilibria, we will not explore them at this time as we are interested in groups following norms. Recall that our primary model posits a situation where the *benefit* of not helping outweighs the potential reward of *helping*. We also notice that when $r > b$, the backward induction solution to the asymmetric trust game is trivial, namely always trusting. That may not be the case for the symmetric trust game however. The figure in Figure 5.8 depicts an example of this context, but we will ignore this case for now. An interesting note is that the payoffs and feasible region resemble that of the Stag Hunt in some ways.

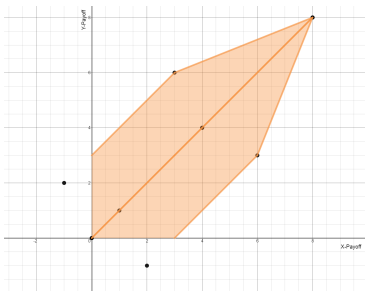


Figure 5.8: Example of feasible region for omitted case with $c < b < r$; here $r = 3, b = 2, c = 1$.

Within the scope of a repeated game, there are many complexities in strategies that can take place. We will focus our attention on the classes of imitation and trigger strategies and the naïve strategies that play against them. Within the class, we will innovate new strategies designed to fit certain qualities we deem desirable in the Symmetric Trust Game, much in the sense of mechanism design [Myerson, 1997].

To be able to compare payoffs in the stage game to those in the repeated game, recall that a rational player i in a repeated game against j should seek to maximize the *normalized sum S of discounted payoffs:*

$$S_i = (1 - \delta) \sum_{t=0}^{\infty} \delta^t v_i(a_i^t, a_j^t) \tag{5.1}$$

Also recall that for our discount parameter $0 < \delta < 1$,

$$\sum_{j=0}^{\infty} \delta^j = \frac{1}{1 - \delta}$$

One-Shot Deviation and Subgame-Perfect Nash Equilibrium

Strategies under the folk theorem must satisfy a condition of producing subgame-perfect Nash Equilibria; i.e. they must gives us equilibrium states from which no agent would deviate in any subset of the game's successive history if we truncated play at a given point. Since we are considering naïve, repetitive strategies against trigger strategies that will change play on deviation from the preferred state, in a repeated game with sufficiently patient players, we can consider that any subgame will play out equivalently to an infinitely repeated game with a potential deviation in the first round.

Deriving symmetric strategy profiles in equilibrium amounts to constructing automata that produce a path of subgame-perfect Nash Equilbria. These automata give us a matrix of discounted payoffs. Thus a sufficient criterion is to find the discount value under which the strategy is a best response to itself, and thus a Nash Equilibrium .

The primary way to encourage maintenance of a repeated strategy is to punish deviation. Punishment is a way of minimizing the score of the other player.

Considering the general Symmetric Trust Game we see that the punishing strategy in general is $\neg A \neg H$, as playing against it gives the worst outcome. This means that when we invoke the Folk Theorem we should have trigger strategies that potentially use $\neg A \neg H$ in the case of a one-shot deviation. As the *tit-for-tat* has also shown successful in the repeated Prisoner's Dilemma, we will investigate whether the Symmetric Trust Game can feature a similar strategy that imitates the opponent's last move for the remainder of the game. Note that this is not necessarily as harsh as a punishing strategy.

5.4 Friendly and Insistent Strategies

This section compares strategies that will produce the cooperative (friendly) (AH, AH) and insistent $(A\neg H, A\neg H)$ outcomes in the repeated game. We explore imitation and punishment, giving an example of each for the friendly case, as the insistent case will work in a similar fashion.

Friendly Imitator

We proceed as before with a strategy that we will denote as *Friendly Imitator(FI)*, taking the perspective of the row player X. It begins with the Friendly action of AH. Upon encountering another action from Y, it switches to imitating that action, as seen in Figure 5.9. As for the other strategies, we will have monotonically repeated versions of the stage games strategies. Instead of writing $All\neg AH$ fpr these naïve strategies , we will simply write $\neg AH$.

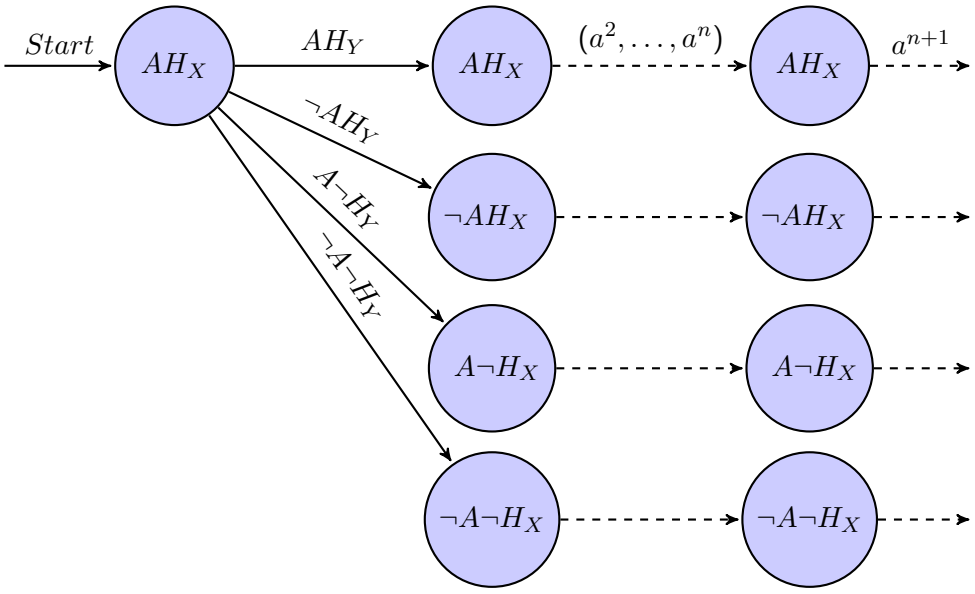


Figure 5.9: Friendly Imitator for X vs. Y

We will use the row player payoff matrix here to calculate the normalized payoffs in the repeated game for the general case $0 < \delta < 1$. For this first set, the respective payoffs per round will follow the first row in the first round, with

the imitated strategies following the diagonal. In the remaining examples, we will omit this stage game matrix and list the payoffs in the discounted, repeated game.

$$M_X = \begin{array}{c} \\ AH \\ \neg AH \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} AH & \neg AH & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r & r-c & -c \\ r & 0 & r & 0 \\ r+b & r & b-c & -c \\ b & 0 & b & 0 \end{bmatrix} \end{array}$$

$$\begin{aligned}
U(FI|FI) =& (1-\delta)(2+\delta\sum_{t=0}^{\infty}\delta^t(2)) = 2r \\
U(FI|\neg AH) =& (1-\delta)(r+\delta\sum_{t=0}^{\infty}\delta^t(0)) = r(1-\delta) \\
U(FI|A\neg H) =& (1-\delta)(r-c+\delta\sum_{t=0}^{\infty}\delta^t(b-c)) \\
=& (1-\delta)\frac{(r-c)(1+\delta)+\delta(b-c)}{1-\delta} = r-c+\delta(b-r) \\
U(FI|\neg A\neg H) =& (1-\delta)(-c+\delta\sum_{t=0}^{\infty}\delta^t(0)) = -c(1-\delta)
\end{aligned}$$

For this second set of calculations, the respective payoffs per round will follow the first column in the first round, with the imitated strategies following the diagonal.

$$\begin{aligned}
U(\neg AH|FI) =& (1-\delta)(r+\delta\sum_{t=0}^{\infty}\delta^t(0)) = r(1-\delta) \\
U(A\neg H|FI) =& (1-\delta)(r+b+\delta\sum_{t=0}^{\infty}\delta^t(b-c)) \\
=& (1-\delta)\frac{(r+b)(1-\delta)+\delta(b-c)}{1-\delta} = r+b-\delta(r+c) \\
U(\neg A\neg H|FI) =& (1-\delta)(b+\delta\sum_{t=0}^{\infty}\delta^t(0)) = b(1-\delta)
\end{aligned}$$

$$M_X = \begin{array}{c} \\ FI \\ \neg AH \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} FI & \neg AH & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r(1-\delta) & r-c+\delta(b-r) & -c(1-\delta) \\ r(1-\delta) & 0 & r & 0 \\ r+b-\delta(r+c) & r & b-c & -c \\ b(1-\delta) & 0 & b & 0 \end{bmatrix} \end{array}$$

From these inequalities, we can build the repeated game payoff matrix. There are two ways to consider here the discounted payoffs:

- One is that we now need to determine the parameter δ that will give us a case where the strategy FI is a best response to itself. This will occur where $2r > r + b - \delta(r + c)$ or $\delta > \frac{b-r}{r+c}$.

- The other is that for a given discount factor δ, we have constraints on the parameters r, b, c.

Looking back at the payoff matrix in the stage game, we are essentially comparing the difference between the tempting payoff and the symmetric payoff to the sum of the symmetric payoff and cost of trust. We can observe that this discount factor will vary directly with that difference. I.e. the greater benefit gained by not helping, the greater the investment in the long-term relationship must exist. As we increase the cost of interaction or the symmetric reward, the less patient players have to be with each other to favor the cooperative outcome.

Using the second interpretation, given a relationship and its accompanying discount factor δ , we can construct a set of payoffs that could be sustained under that given discount. For instance, in our previous analysis, we found that $\delta > \frac{b-r}{r+c}$ is a discount constraint under which we expect cooperation. Given a discount of $\frac{1}{2}$ for instance, we would then be able to conclude that if $\frac{b-r}{r+c} < \frac{1}{2}$, then we can expect cooperation. A case where there is a high cost of interaction or a low difference between the reward from helping and the benefit of not helping might satisfy this.

Friendly Punisher

We now proceed with the strategy that denoted as *Friendly Punisher*. It begins with the Friendly strategy of AH. Upon encountering another strategy, it switches to the trigger strategy $\neg A\neg H$ as a way of punishing the other player. To see why this is a punishing strategy, notice that the strategy produces the worst possible payoff for each corresponding play from the other, provided that $r, b, c > 0$. As for the other strategies, we will have monotonically repeated versions of the stage games strategies. Instead of writing $All\neg AH$, we will simply write $\neg AH$.

We will use the row player payoff matrix seen before to calculate the normalized payoffs in the repeated game for the general case $0 < \delta < 1$. For this first set, the respective payoffs per round will follow the first row in the first round, with the punishing strategies following the last row's final three entries.

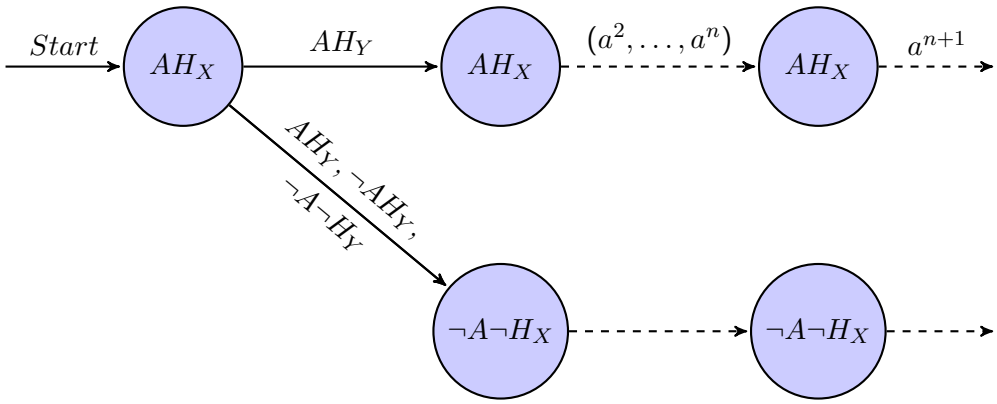


Figure 5.10: Friendly Punisher Strategy for X vs. Y

$$
\begin{aligned}
U(FP|FP) =& (1-\delta)(2r + \delta \sum_{t=0}^{\infty} \delta^t(2r)) = 2r \\
U(FP|\neg AH) =& (1-\delta)(1 + \delta \sum_{t=0}^{\infty} \delta^t(0)) = r(1-\delta) \\
U(FP|A\neg H) =& (1-\delta)(r - c + \delta \sum_{t=0}^{\infty} \delta^t(b)) \\
=& (1-\delta)\frac{(r-c)(1-\delta) + \delta b}{1-\delta} = r - c + \delta(b + c - r) \\
U(FP|\neg A\neg H) =& (1-\delta)(-c + \delta \sum_{t=0}^{\infty} \delta^t(0)) = c(\delta - 1)
\end{aligned}
$$

For this second set of calculations, the respective payoffs per round will follow the first column in the first round, with the punished strategies following the last column's final three entries.

$$
\begin{aligned}
U(\neg AH|FP) =& (1-\delta)(r + \delta \sum_{t=0}^{\infty} \delta^t(0)) = r(1-\delta) \\
U(A\neg H|FP) =& (1-\delta)(r + b + \delta \sum_{t=0}^{\infty} \delta^t(-c)) \\
=& (1-\delta)\frac{r + b - \delta(r + b + c)}{1-\delta} = r + b - \delta(r + b + c) \\
U(\neg A\neg H|FP) =& (1-\delta)(b + \delta \sum_{t=0}^{\infty} \delta^t(0)) = b(1-\delta)
\end{aligned}
$$

$$M_X = \begin{array}{c} \\ FP \\ \neg AH \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} FP & \neg AH & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r(1-\delta) & r-c+\delta(b+c-r) & c(\delta-1) \\ r(1-\delta) & 0 & r & 0 \\ r+b-\delta(r+b+c) & r & b-c & -c \\ b(1-\delta) & 0 & b & 0 \end{bmatrix} \end{array}$$

The repeated game payoff matrix of discounted payoffs can help up find the parameter δ that will give us where FP is a best response to itself. This will occur where $2r > r + b - \delta(r + b + c)$ or $\delta > \frac{b-r}{r+b+c}$.

Friendly Norm Enforcer

According to [Bicchieri, 2002], trustworthiness *is* a norm although trust itself is not. We take this in our game to denote that the action H should be enforced more stringently than A, as H ensures the Pareto-optimal outcome (r, r). One way to do this is through an *Enforcer Strategy* that only triggers among instances of $\neg H$ in the game. Otherwise, it may play consistently or imitate the previously seen strategy.

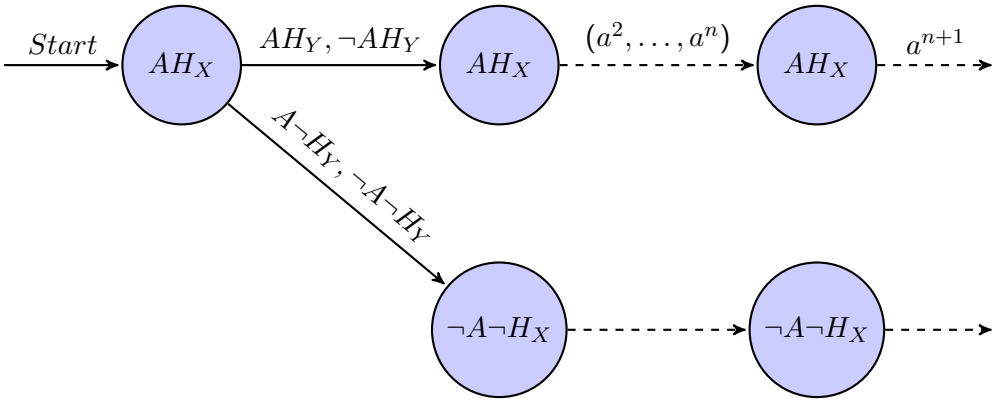


Figure 5.11: Friendly Enforcer Strategy

$$M_X = \begin{array}{c} \\ FE \\ \neg AH \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} FE & \neg AH & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r & r-c+\delta(b+c-r) & -c(1-\delta) \\ r & 0 & r & 0 \\ r+b-\delta(r+b+c) & r & b-c & -c \\ b(1-\delta) & 0 & b & 0 \end{bmatrix} \end{array}$$

From the automaton, we can build the repeated game payoff matrix. Just as before in our encounter with Friendly Punishment strategies, we see that FE is a best response to itself where $2r > r + b - \delta(r + b + c)$ or $\delta > \frac{b-r}{r+b+c}$. The salient difference here that we have a higher utility for altruistic players.

Insistent Strategies

Under some permutations of the game parameters, we saw that the insistent strategy of $A\neg H$ could yield feasible symmetric strategy profiles in the repeated game according to the folk theorem. In particular, these are the cases where $c < b$. This could occur in case I, where $c < r < b$ or in case II, where $r < c < b$. These are cases where the message cost is low compared to the benefit of not helping. As this strategy is the major obstacle to the friendly outcomes, we would like to see how its discount values compare to the previous ones.

In preliminary tournament simulations similar to Axelrod [1984], the insistent strategy has emerged as the most successful in pairwise competition for its exploitative nature. [3] As the calculations follow the same paradigm as we saw with the friendly strategies, we will omit them until reaching the results at the end of the section. We will however give the automata, as these are more revealing of the strategy itself.

Insistent Imitator

Insistent Imitator(II). begins with the insistent action of $A\neg H$. Upon encountering another strategy, it switches to repeating that strategy, seen in Figure 5.12. As for the other strategies, we will have monotonically repeated versions of the stage games strategies.

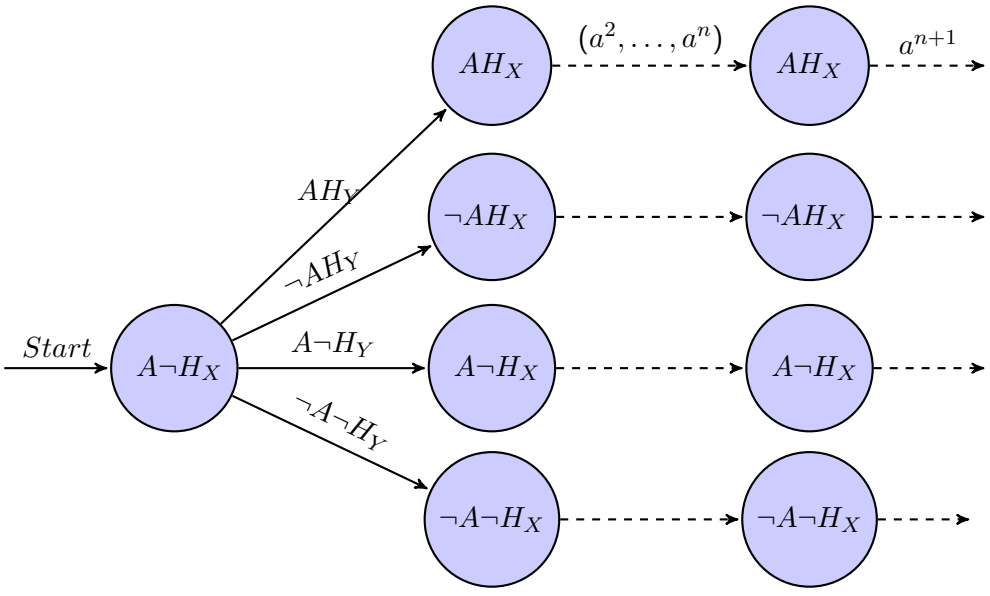


Figure 5.12: Insistent Imitator for X

From the automaton, we can build the repeated game payoff matrix. II is a best response to itself where $b - c > b(1 - \delta)$ or $\delta > \frac{c}{b}$.

[3] Mühlenbernd and Quinley (forthcoming)

$$M_X = \begin{array}{c} \\ AH \\ \neg AH \\ II \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} AH & \neg AH & II & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r & r-c+\delta(r+c) & c \\ r & 0 & r(1-\delta) & 0 \\ r+b+\delta(r-b) & r(1-\delta) & b-c & -c \\ b & 0 & b(1-\delta) & 0 \end{bmatrix} \end{array}$$

Insistent Punisher

Insistent Punisher begins with the insistent action of $A\neg H$. Upon encountering another strategy, it switches to the trigger strategy $\neg A\neg H$ as a way of punishing the other player. IP is a best response to itself where $b - c > b(1 - \delta)$ or $\delta > \frac{c}{b}$.

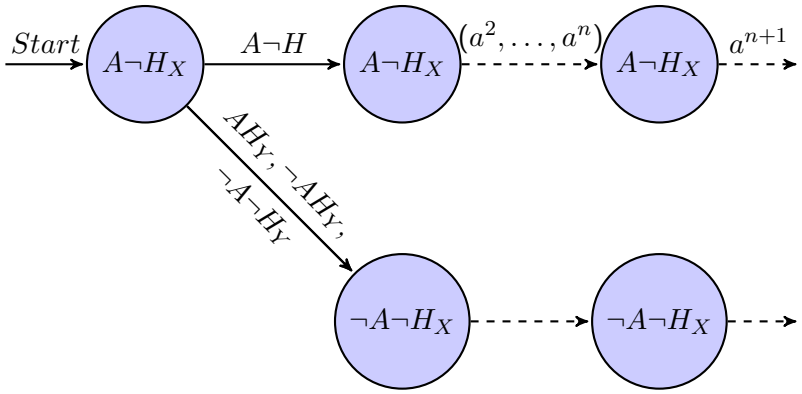


Figure 5.13: Insistent Punisher Strategy for X vs. Y

$$M_X = \begin{array}{c} \\ AH \\ \neg AH \\ IP \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{cccc} AH & \neg AH & IP & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r & r-c-\delta r & c \\ r & 0 & r(1-\delta) & 0 \\ r+b+\delta r & r(1-\delta) & b-c & -c \\ b & 0 & b(1-\delta) & 0 \end{bmatrix} \end{array}$$

Commentary and Results

With the friendly strategies, we have that the more credible the threat of punishment, the less patient players need to be to enforce the Nash Equilibrium [Myerson, 1997]. This means that the opt-out condition signified by the strategy FP as compared to the imitation strategy signified by FI can hold cooperation in place for a less stable or future-oriented pair of players, as the punishing strategy opts out of future interaction. With the threat of punishment, we see that if any of the game's parameters increase, we have, for a fixed discount value, a further expectation of cooperative behavior. This result, however, does not hold for the insistent strategies.

The results on the insistent strategies are a contrast to those seen in the friendly strategies. In both cases, we saw that a discount value $\delta > \frac{c}{b}$ can sustain

Type	Strategy	Sym Nash Equilibrium	Discount
Friendly	Imitator	$(AH,\ AH)$	$\delta > \frac{b-r}{r+c}$
Friendly	Punisher	$(AH,\ AH)$	$\delta > \frac{b-r}{r+b+c}$
Insistent	Imitator	$(A\neg H,\ A\neg H)$	$\delta > \frac{c}{b}$
Insistent	Punisher	$(A\neg H,\ A\neg H)$	$\delta > \frac{c}{b}$

Table 5.11: Here we see the results from using the imitation and punishment dynamics on the symmetric trust game. Only the symmetric equilibria are highlighted.

the mutually performed insistent strategy $A\neg H$. As the insistent strategies choose the Ask option in the asymmetric game, this would favor a cheaper message cost. Punishment does not have the same effect as a threat here. This may be because the norm of not helping is already violated, and this strategy is already the most exploitative. As to whether the insistent strategies favor less patient players, this depends on the game parameters. In the normalized example with concrete payoffs, both of these patience thresholds give $\delta > \frac{1}{3}$.

Another thing to notice here is that we can explore what happens when $b-c > 2r$. This would not only make the insistent strategy profile $(A\neg H,\ A\neg H)$ a Nash Equilibrium but also an evolutionarily stable strategy. In contrast, when $b - c < 2r$, then we have that the friendly profile $(AH,\ AH)$ is not only a Nash Equilibrium but also an evolutionary stable strategy. If we look back to our original formulation of the trust game, we see that this is a comparison between the sum of the payoffs on one branch to the sum of the payoffs on the other.

5.5 Strategic Restrictions and One-Shot Games

Much of our discussion revolves around the similarity between the Prisoner's Dilemma and the Symmetric Trust Game. We would like to strengthen that connection by highlighting the similarity between the two games as a case of what we call *strategic restriction*. If we look at their various payoff matrices, we see a striking resemblance between strategies in both.

By *strategic restriction*, we mean keeping the payoffs and outcomes of the game the same but eliminating certain strategies from the field of play. We differentiate this from a subgame, which as we saw before is a truncation of a repeated game along with that game's history. We mention this topic here and not earlier, as we need a game with more than two strategies to have a meaningful discussion. Not surprisingly, *iterated elimination of dominated strategies* is a special case of strategic restriction. There are many other rationales for strategic restriction besides finding a Nash Equilibrium in a game with a larger strategy space, and we now see some of the impacts here.

We can define a general concept of strategic restrictions and expansions that function much like the subset relation. Consider games G_1 and G_2 with respective action spaces A_1 and A_2 and utility functions on those spaces U_1 and U_2.

Definition 1. *We say* $G_1 \preceq_{STRAT} G_2$ *iff* $\forall a \in A_1$, $a \in A_2$ *and* $U_1(a) = U_2(a)$.

Looking at this definition and the generalized Prisoner's Dilemma we can see that we have more than a casual relationship between it and our Symmetric Trust Game Ṫo make this clearer, we define the following terms:

Definition 2. *We say* G_1 *is a* ***strategic restriction*** *of* G_2 *iff* $G_1 \preceq_{STRAT} G_2$. *Conversely,* G_2 *is a* ***strategic expansion*** *of* G_1.

Categorical Strategic Restrictions

What happens when we restrict the strategies of the symmetric game to a few *types*? In one case, we can eliminate actions from the original asymmetric game wholesale, and as the new game is symmetric, that eliminates two rows or columns. We call that *Categorical Strategic Restriction*. E.g. eliminating $\neg A$ removes both $\neg AH$ and $\neg A\neg H$. Motivations for this might be external factors like societal norms or internal factors like awareness of available actions. If we restrict strategies categorically, we get payoff matrices like the following seen in Table 5.12 and Table 5.13.

	$\neg AH$	$\neg A\neg H$
$\neg AH$	0,0	0,0
$\neg A\neg H$	0,0	0,0

	AH	$A\neg H$
AH	2r,2r	r-c,r+b
$A\neg H$	r+b,r-c	b-c,b-c

Table 5.12: Variations of Trust: The game on the right is a Prisoner's Dilemma as $b > r$ and $b - c < 2r$.

Just as we see with the general payoff parameters, the same patterns would occur in the basic game with its concrete payoffs. This is not the only way to produce the Prisoner's Dilemma as the next table shows.

	$A\neg H$	$\neg A\neg H$
$A\neg H$	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0

	AH	$\neg AH$
AH	2r,2r	r,r
$\neg AH$	r,r	0,0

Table 5.13: Variations of Trustworthiness: The game on the left is the Prisoner's Dilemma; the right is a trivial game.

Independent Strategic Restrictions

We can also perform strategic restrictions one at a time by treating each action in the Symmetric Trust Game as independent from the others. As such, we can restrict our game around those types. We will now give the game tables created by eliminating those strategies for both the general payoffs and our specific payoffs. Going forward, we will focus more attention to the general payoffs generated by eliminating one of the following four actions.

AH:Friendly $\neg AH$:Altruistic $A\neg H$:Insistent $\neg A\neg H$:Reclusive

	$\neg AH$	$A\neg H$	$\neg A\neg H$
$\neg AH$	0,0	r,r	0,0
$A\neg H$	r,r	b-c,b-c	-c, b
$\neg A\neg H$	0,0	b,-c	$\boxed{0,0}$

	$\neg AH$	$A\neg H$	$\neg A\neg H$
$\neg AH$	0,0	1,1	0,0
$A\neg H$	1,1	1,1	-1, 2
$\neg A\neg H$	0,0	2,-1	$\boxed{0,0}$

Table 5.14: Eliminate Friendly Types: Equilibrium of ($\neg A\neg H$,$\neg A\neg H$).

We can eliminate Friendly players as done above or eliminate Altruistic players. In each of these cases, the Nash Equilibrium of both players playing $\neg A\neg H$ remains.

	AH	$A\neg H$	$\neg A\neg H$	
AH	2r,2r		r-c,r+b	-c, b
$A\neg H$	r+b,r-c		b-c,b-c	-c, b
$\neg A\neg H$	b,-c		b,-c	$\boxed{0,0}$

	AH	$A\neg H$	$\neg A\neg H$
AH	2,2	0,3	-1, 2
$A\neg H$	3,0	1,1	-1, 2
$\neg A\neg H$	2,-1	2,-1	$\boxed{0,0}$

Table 5.15: Eliminate Altruistic Types: Equilibrium of ($\neg A\neg H$,$\neg A\neg H$).

Further, we can eliminate Insistent players or Reclusive Types. For each elimination of these types, we can obtain an equilibrium potentially different from that of the larger game. An interesting point to note is what the natural Nash Equilibrium is when we eliminate other strategies, seen boxed in the various tables.

	AH	$\neg AH$	$\neg A\neg H$
AH	$\boxed{2r,2r}$	1r,1r	-c, b
$\neg AH$	1r,1r	0,0	0,0
$\neg A\neg H$	b,-c	0,0	$\boxed{0,0}$

	AH	$\neg AH$	$\neg A\neg H$
AH	$\boxed{2,2}$	1,1	-1, 2
$\neg AH$	1,1	0,0	0,0
$\neg A\neg H$	2,-1	0,0	$\boxed{0,0}$

Table 5.16: Eliminate Insistent Types: This game has TWO symmetric pure strategy Nash Equilibria. In addition to ($\neg A\neg H$, $\neg A\neg H$) we have (AH,AH).

5.6 Strategic Restriction and Repetition

We now consider strategic restriction of the symmetric trust games under repeated strategies like imitation and punishment. With eliminated strategies, we can apply the folk theorem again to derive strategies and discount values under which we see new equilibria in the repeated game emerge. Part of the motivation here is that in certain interactions, agents may not always have every choice from the larger space of actions available or apparent to them. This can be from a

	AH	$\neg AH$	$A\neg H$
AH	2r,2r	r,r	r-c,r+b
$\neg AH$	r,r	0,0	r,r
$A\neg H$	r+b,r-c	r,r	b-c,b-c

	AH	$\neg AH$	$A\neg H$
AH	2,2	1,1	0,3
$\neg AH$	1,1	0,0	1,1
$A\neg H$	3,0	1,1	1,1

Table 5.17: Eliminate Reclusive Types: This game has a new symmetric pure strategy Nash Equilibria at ($A\neg H$, $A\neg H$). The general game also potentially has asymmetric equilibria at ($A\neg H$, $\neg AH$) and ($\neg AH$, $A\neg H$) if $r \geq b - c$.

lack of awareness or norm against the particular behavior. We will first perform independent restriction, eliminating player types from the larger Symmetric Trust Game.

Independent Restriction: Imitation vs. Punishment

For each independent restriction, we will examine both the imitation and punishment strategies. We will call the repeated matrix under *Imitation* I_X and the repeated game matrix under *Punishment* P_X. A central question will be whether the credible threat of punishment can lead to a reduced discount value compared to imitation.

Friendly Types Restricted

	$\neg AH$	$A\neg H$	$\neg A\neg H$
$\neg AH$	0,0	r,r	0,0
$A\neg H$	r,r	b-c,b-c	-c, b
$\neg A\neg H$	0,0	b,-c	0,0

Table 5.18: Eliminate Friendly Types: NE of ($\neg A\neg H$,$\neg A\neg H$).

With the action AH eliminated, we would like to find the conditions under which insistence ($A\neg H$) becomes stable. Note that for $r > b - c$ we might also be interested in the asymmetric equilibria. This would not conflict with $r < b$ for a sufficiently high message cost c.

$$I_X = \begin{array}{c} \\ \neg AH \\ II \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} \neg AH & II & \neg A\neg H \end{array} \\ \begin{bmatrix} 0 & r(1-\delta) & 0 \\ r(1-\delta) & b-c & -c(1-\delta) \\ 0 & b(1-\delta) & 0 \end{bmatrix} \end{array};$$

$$P_X = \begin{array}{c} \\ \neg AH \\ IP \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} \neg AH & IP & \neg A\neg H \end{array} \\ \begin{bmatrix} 0 & r(1-\delta) & 0 \\ r(1-\delta) & b-c & -c(1-\delta) \\ 0 & b(1-\delta) & 0 \end{bmatrix} \end{array}$$

Under the imitation and punishment strategies, we have the repeated game payoff matrices I_X and P_X, respectively, for the row player X. Both II and IP are best responses to themselves where $b - c > b(1 - \delta)$ or $\delta > \frac{c}{b}$. Thus in the case where discount values are sufficiently great, the insistent type is stable under either trigger strategy. This occurs for a low cost of interaction and a high benefit to not helping.

Altruistic Types Restricted

	AH	$A\neg H$	$\neg A\neg H$
AH	2r,2r	r-c,r+b	-c, b
$A\neg H$	r+b,r-c	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	b,-c	0,0

Table 5.19: Eliminate Altruistic Types:Nash Equilibrium of ($\neg A\neg H$,$\neg A\neg H$)

With the Altruistic Types eliminated, we will consider the conditions for (AH, AH) to be a Nash Equilibrium in the repeated game following the Friendly Imitator and Friendly Punisher strategies. Under the imitation and punishment strategies, we have the repeated game payoff matrices I_X and P_X, respectively, for the row player X.

$$I_X = \begin{array}{c} \\ FI \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FI & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r-c-\delta(r-b) & -c(1-\delta) \\ r+b-\delta(r+c) & b-c & -c \\ b(1-\delta) & b & 0 \end{bmatrix} \end{array};$$

$$P_X = \begin{array}{c} \\ FP \\ A\neg H \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FP & A\neg H & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r-c-\delta(r-b-c) & -c(1-\delta) \\ r+b-\delta(r+b+c) & b-c & -c \\ b(1-\delta) & b & 0 \end{bmatrix} \end{array}$$

FI is a best response to itself where $2r > r + b - \delta(r + c)$ or $\delta > \frac{b-r}{r+c}$ and FP is a best response to itself where $2r > r + b - \delta(r + b + c)$ or $\delta > \frac{b-r}{r+b+c}$. Thus punishment requires less patient players.

$$I_X = \begin{array}{c} \\ AH \\ II \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} AH & II & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r-c+\delta(r+c) & -c \\ r+b-\delta(b-r) & b-c & -c(1-\delta) \\ b & b(1-\delta) & 0 \end{bmatrix} \end{array};$$

$$P_X = \begin{array}{c} \\ AH \\ IP \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} AH & IP & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r-c-\delta r & -c \\ r+b-\delta r & b-c & -c(1-\delta) \\ b & b(1-\delta) & 0 \end{bmatrix} \end{array}$$

It also could be the case that $b - c > 2r$, in which case Insistent Imitator/Punisher would be the most efficient in the one-shot case. As we typically assume $b > c$ it is already a feasible equilibrium in the repeated game. We thus have the above matrices for Insistent Imitator and Insistent Punisher. In both cases, II and IP are best responses to themselves when $b-c > b(1-\delta)$ or $\delta > \frac{c}{b}$.

Insistent Types Restricted

	AH	$\neg AH$	$\neg A\neg H$
AH	2r,2r	r,r	-c, b
$\neg AH$	r,r	0,0	0,0
$\neg A\neg H$	b,-c	0,0	0,0

Table 5.20: Eliminate Insistent Types: NE: ($\neg A\neg H$, $\neg A\neg H$) and possibly (AH,AH).

In this case the cooperative outcome (AH,AH) is one of the two Nash Equilibria in addition to ($\neg A\neg H$, $\neg A\neg H$), provided that $2r > b$. This should occur in cases except those with a very high message cost and low rewards to being helped.

$$I_X = \begin{array}{c} \\ FI \\ \neg AH \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FI & \neg AH & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r(1-\delta) & -c(1-\delta) \\ r(1-\delta) & 0 & 0 \\ b(1-\delta) & 0 & 0 \end{bmatrix} \end{array};$$

$$P_X = \begin{array}{c} \\ FP \\ \neg AH \\ \neg A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FP & \neg AH & \neg A\neg H \end{array} \\ \begin{bmatrix} 2r & r(1-\delta) & -c(1-\delta) \\ r(1-\delta) & 0 & 0 \\ b(1-\delta) & 0 & 0 \end{bmatrix} \end{array}$$

Both FI and FP are best responses to themselves where $2r > b(1-\delta)$ or $\delta > \frac{b-2r}{b}$. In the case where $b < 2r$, the cooperative outcome is a Nash Equilibrium from the beginning and will remain so for any positive discount value.

Reclusive Types Restricted

No matter the magnitude of r, we will still have that $2r < r + b$ as $r < b$, and thus the cooperative outcome needs the mechanism of repetition to succeed. We therefore want to find the conditions under which (AH, AH) is a Nash Equilibrium.

	AH	$\neg AH$	$A\neg H$
AH	2r,2r	r,r	r-c,r+b
$\neg AH$	r,r	0,0	r,r
$A\neg H$	r+b,r-c	r,r	b-c,b-c

Table 5.21: Eliminate Reclusive Types: This game has a new symmetric pure strategy Nash Equilibrium at ($A\neg H$, $A\neg H$) if $b-c > r$. It also has asymmetric equilibria at ($A\neg H$, $\neg AH$) and ($\neg AH$, $A\neg H$)

$$I_X = \begin{array}{c} \\ FI \\ \neg AH \\ A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FI & \neg AH & A\neg H \end{array} \\ \begin{bmatrix} 2r & r(1-\delta) & r-c+\delta(b-r) \\ r(1-\delta) & 0 & r \\ r+b-\delta(r+c) & r & b-c \end{bmatrix} \end{array};$$

$$P_X = \begin{array}{c} \\ FI \\ \neg AH \\ A\neg H \end{array} \begin{array}{c} \begin{array}{ccc} FI & \neg AH & A\neg H \end{array} \\ \begin{bmatrix} 2r & r & r-c+\delta(b-r) \\ r & 0 & r \\ r+b-\delta(r+c) & r & b-c \end{bmatrix} \end{array}$$

FI and FP are best responses to themselves where $2r > r+b-\delta(r+c)$ or $\delta > \frac{b-r}{r+c}$. For this t set of calculations, the respective payoffs per round will follow the first row in the first round, with punished strategies following the last row. We should remark before proceeding that this strategic restriction has removed the default punishment from the game, so we adapt to the strategy that minimizes the score of the other in equiprobable cases. If we take it to be that higher message costs give us $b < 2c$, then the minimizing strategy is the insistent one $A\neg H$, as a randomized player against $A\neg H$ would get $\frac{2r+b-2c}{3}$ as opposed to $\frac{2r}{3}$.

Categorical Strategic Restriction

Categorically restricting a certain action can lead to reduced game matrix resembling the Prisoner's Dilemma. We now return to an analysis of those restrictions, borrowing in cases from our chapter on repeated games.

	$A\neg H$	$\neg A\neg H$
$A\neg H$	b-c,b-c	-c, b
$\neg A\neg H$	b,-c	0,0

	AH	$\neg AH$
AH	2r,2r	r,r
$\neg AH$	r,r	0,0

Table 5.22: Variations of Trustworthiness:The game on the right is trivial, as the NE is the most efficient; the one on the left is the classical Prisoner's Dilemma.

$$M_X = \begin{matrix} \\ IP \\ \neg A\neg H \end{matrix} \begin{matrix} \begin{matrix} IP & \neg A\neg H \end{matrix} \\ \begin{bmatrix} b-c & -c(1-\delta) \\ b(1-\delta) & 0 \end{bmatrix} \end{matrix}$$

Grim Trigger (or Insistent Punisher) gives us the redefined matrix, where cooperation has an advantage and an advantage when $b-c > b(1-\delta)$ or $\delta > c/b$. I.e. we need the discount parameter to be greater than the ratio of the cost of cooperation compared to the benefit of the other cooperating.

	$\neg AH$	$\neg A\neg H$
$\neg AH$	0,0	0,0
$\neg A\neg H$	0,0	0,0

	AH	$A\neg H$
AH	2r,2r	r-c,r+b
$A\neg H$	r+b,r-c	b-c,b-c

Table 5.23: Variations of Trust: The game on the right is potentially a version of the Prisoner's Dilemma for the appropriate payoffs.

Restricting a player's ability to ask leaves us with a degenerate game, but if player are required to ask, we have a variant of the Prisoner's Dilemma, since $r > b$ implies that $A\neg H$ strictly dominates AH. This time, the *Grim Trigger* strategy is equivalent to the *Friendly Punisher* strategy.

$$M_X = \begin{matrix} \\ FP \\ A\neg H \end{matrix} \begin{matrix} \begin{matrix} FP & A\neg H \end{matrix} \\ \begin{bmatrix} 2r & r-c+\delta(b-r) \\ r+b-\delta(r+c) & b-c \end{bmatrix} \end{matrix}$$

Friendly Punisher gives us the advantage in cooperating when $2r > r+b-\delta(r+c)$ or $\delta > \frac{b-r}{r+c}$. I.e. we need the discount parameter to be greater than the ratio of the cost of helping compared to the combined reward from helping and message cost.

Results on Strategic Restriction

Here we have the various results over the permutations of strategic restriction in Table 5.24.

From the last sections on imitation and punishment in section 5.4, we had the results seen in Table 5.25.

We can compare these tables and see that the discount values are identical for the promoted symmetric Nash Equilibrium given by discounted repetition each game. Notice that as there are only two action in the categorically restricted games, the trigger strategies of *Imitator* and *Punisher* are indistinguishable from each other and *Grim Trigger*.

Restriction	One-Shot NE	Rep NE	Discount(Imi)	Discount(Pun)
F:AH	$(\neg A\neg H,\ \neg A\neg H)$	$(A\neg H,\ A\neg H)$	$\delta_{IMI} > \frac{c}{b}$	$\delta_{PUN} > \frac{c}{b}$
A: $\neg AH$	$(\neg A\neg H,\ \neg A\neg H)$	$(AH,\ AH)$ $(A\neg H,\ A\neg H)$	$\delta_{IMI} > \frac{b-r}{r+c}$ $\delta_{IMI} > \frac{c}{b}$	$\delta_{PUN} > \frac{b-r}{r+b+c}$ $\delta_{PUN} > \frac{c}{b}$
I: $A\neg H$	(AH,AH) $(\neg A\neg H,\ \neg A\neg H)$	(AH,AH)	$\delta_{IMI} > \frac{b-2r}{b}$	$\delta_{PUN} > \frac{b-2r}{b}$
R:$\neg A\neg H$	$(A\neg H,\ A\neg H)$	(AH,AH)	$\delta_{IMI} > \frac{b-r}{r+c}$	$\delta_{PUN} > \frac{b-r}{r+c}$

Table 5.24: Strategic Restrictions and Discounts for sustaining Nash Equilibria in the Repeated Game

Type	Strategy	Rep NE	Discount
Friendly	Imitator	$(AH,\ AH)$	$\delta > \frac{b-r}{r+c}$
Friendly	Punisher	$(AH,\ AH)$	$\delta > \frac{b-r}{r+b+c}$
Insistent	Imitator	$(A\neg H,\ A\neg H)$	$\delta > \frac{c}{b}$
Insistent	Punisher	$(A\neg H,\ A\neg H)$	$\delta > \frac{c}{b}$

Table 5.25: Results from using the imitation and punishment strategies on the symmetric trust game. Only the symmetric equilibria from the repeated game are highlighted.

5.7 Conclusion and Discussion

This chapter gave us several results:

- Punishment Strategies presented a more credible threat to the other players, and thus they gave us a lower threshold by which we can sustain the cooperative outcomes of the games.

- Insistent Strategies and Friendly Strategies could be both Nash Equilibria and Evolutionarily Stable Strategies in the repeated games.

- The strength of insistent strategies makes the conditions for maintaining the cooperative, friendly strategies more stringent.

- Strategic restriction produced results leading to new equilibria under one-shot games, but the discounts required to promote the various strategies remained constant across the restricted and unrestricted Symmetric Trust Game.

- Equilibria in the symmetric game's one-shot version correspond to the asymmetric game

- If we making asking or helping mandatory, we land at games strategically equivalent to the Prisoner's Dilemma.

Restriction	Rep NE	Discount
A	N/A	N/A
$\neg A$	$(AH,\ AH)$	$\delta > \frac{b-r}{r+c}$
H	$(A\neg H,\ A\neg H)$	$\delta > \frac{c}{b}$
$\neg H$	(AH,AH)	$\delta > 0$

Table 5.26: Categorical restrictions with *Grim Trigger*.

· We notice that as we increase the message cost c in the symmetric trust game that we obtain conditions that lead to more cooperative outcomes; .e.g less patient players are required to sustain the Pareto-optimal profiles. Note that this is the area of speech acts controlled by the speaker, and thus this will lead into adopting the work of [Brown and Levinson, 1987].

Bibliography

Robert J. Aumann. Nash Equilibria Are Not Self-Enforcing, chapter 34, pages 667–677. Elsevier, 1990.

Robert Axelrod. The Evolution of Cooperation. Perseus Books Group, Cambridge, 1984.

C. Bicchieri. The grammar of society: The nature and dynamics of social norms. Cambridge University Press, 2006.

Cristina Bicchieri. Covenants without swords group identity, norms, and communication in social dilemmas. Rationality and Society, 14(2):192–228, 2002.

Cristina Bicchieri and Ryan Muldoon. Social norms. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Spring 2011 edition, 2011.

Rebecca BliegeBird, EricAlden Smith, Michael Alvard, Michael Chibnik, Lee Cronk, Lourdes Giordani, EdwardH Hagen, Peter Hammerstein, FraserD Neiman, RebeccaBliege Bird, et al. Signaling theory, strategic interaction, and symbolic capital 1. Current Anthropology, 46(2):221–248, 2005.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. Cross-cultural pragmatics: Requests and apologies, volume 31. Ablex Pub, 1989.

Vanessa K Bohns and Francis J Flynn. "why didn't you just ask?" underestimating the discomfort of help-seeking. Journal of Experimental Social Psychology, 46(2):402–409, 2010.

Penelope Brown and Stephen C. Levinson. Politeness: Some universals in language use. Cambridge University Press, Cambridge, 2nd edition, 1987.

Herbert H Clark and Thomas B Carlson. Hearers and speech acts. Language, pages 332–373, 1982.

Erving Goffman. On face-work: an analysis of ritual elements in social interaction. Psychiatry: Journal for the Study of Interpersonal Processes, 1955.

George C Homans. Social behavior as exchange. American journal of sociology, pages 597–606, 1958.

George J. Mailath and Larry Samuelson. Repeated games and reputations: long-run relationships. Oxford University Press, Oxford, 2006.

Roger B Myerson. Game theory: analysis of conflict. Harvard University Press, 1997.

Martin A Nowak. Five rules for the evolution of cooperation. science, 314(5805): 1560–1563, 2006.

Brian Skyrms. The Stag Hunt and the Evolution of Social Structure. Cambridge University Press, Cambridge, 2004.

Robert van Rooij. Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge, pages 45–58. ACM, 2003.

Robert van Rooij. Signaling games select horn strategies. Linguistics and Philosophy, 27:493–527, 2004.

D.S. Wilson, C. Wilczynski, A. Wells, and L. Weiser. Gossip and other aspects of language as group-level adaptations. The evolution of cognition, pages 347–365, 2000.

A. Zahavi. Mate selection–a selection for a handicap. Journal of theoretical Biology, 53(1):205–214, 1975.