Contents

Chapter 4

Agent Preferences and Interaction Modeling

In carrying out this programme we lay ourselves open to the attack that we are inappropriately reviving the economic homunculus.
[Brown and Levinson, 1987]

Our previous chapter covered external mechanisms of repetition that can alter the incentives of a game. There are other mechanisms that can alter a game, and these are internal to the agents: preferences. We use these mechanisms to account for the presence of cooperative outcomes in the absence of repeated games.

In much of the foundational work done in behavioral economics, Richard Thaler [Thaler, 2000] distinguishes between the notions of *real* humans and the models of *homo economicus* often found in standard game-theoretic literature. We find this distinction a useful default for modeling agent behavior. Just as we began the book with thoughts on our goals, we should also elicit what we would like to avoid.

- Preferences equivalent to individual payoffs.
- Unlimited rationality and introspection.
- Cold-blooded individualism.
- A single game model for all interactions.

These items fit in more naturally with the nature of politeness; that is, it gives heuristics for multiple situations and relationships. We are not alone in making this point, as works like Camerer [2003] also lend strength to the need for incorporating psychological insights into economic and game-theoretic modeling. We will strengthen our claim to the validity of modifying agent preferences looking at the literature from behavioral economics. This not only gives us a more general model of rational behavior, it also sets us up for experimental directions in the final chapters **??**. As to how the model can become more general, for instance, we can differentiate payoffs from preferences by providing a more complex utility function that incorporates payoffs like face or sympathy. As we saw in the last chapter, a preference feature like discounting might also differentiate valuations of current situations from long-term benefits.

4.1 Augmenting Game Models with Psychology

Here we outline the case for importing evidence and notions from psychology into economics as found in Camerer [2003]. The principal problem as discussed is that there is a general inconsistency between economic models and findings in psychology. As such, psychology should inform economics as paralleled in other sciences. This line of thought traces itself historically to Adam Smith's *Theory of Moral Sentiments* and *The Wealth of Nations*, where the latter focused on the derivation of group benefit from selfish motives. As the former posited ideas on decision making and care for a group, incorporating these notions into economic modeling could give better predictions. Relevant areas include the evaluation of utility, the incorporation of social awareness into reasoning, models of equilibria, and procedures for discounting future decisions.

Case from Behavioral Economics

We outline Camerer's case as follows:

- Behavioral economists have found subjects in experiments do not evaluate their choices based on the actual probabilities but rather on a set of heuristics designed to "overweigh low probabilities ... and insure against low-probability disasters" [Camerer, 2003].
- Humans also have a pronounced preference for instant reward, and this preference skews what previous models of discounting predicted. Incorporating this preference further predicts behavior like procrastination.
- The various notions of equilibrium are in need of an overhaul. Using heuristics from psychology like *experience-related attraction,* we can study the process of how a population might arrive at a given equilibrium state. This also gives us a way to see how an equilibrium might give way to another based on changes in the system.
- Although traditional ideas of preferences have equated them with payoffs, behavioral economists have found that people often make decisions with the group in mind. This includes having a notion of fairness and sympathy.
- *Camerer's Recommendation:* Incorporating these ideas should improve modeling. These models predict more accurately many of the phenomena found in actual human behavior.

As mentioned in some of the later chapters on experiments, one of the directions being investigated is the role of the hormonal correlates of emotion-based rewards as drivers of preferences. This includes oxytocin, which reveals several surprising results, as seen in Zak [2008] Some of these results resemble sympathetic decision heuristics when certain social structures are in play (prior friendship) or when certain speech acts are engaged. This gives a further reason to integrate preferences found in cognitive science with economic modeling, as developments in game-theoretic models of communication have the potential to provide further

grounds for testing. One such alteration to preferences that resembles behavior modulated under increased levels of oxytocin is the incorporation of a sympathy preference. Such a preference can bring about a shift toward Pareto-optimal outcomes by encouraging cooperation when a traditionally rational player might defect otherwise.

Preferences and Payoffs

Traditional game-theoretic modeling starts with the premise that an agent *prefers* an action A to B *iff* under a utility function with material payoff U we have $U(A) > U(B)$. Should the utilities be equal, i.e. $U(A) = U(B)$, then we would say that the agent is indifferent to the choices. This notion obscures a subtle fact: our material payoffs within a certain scenario may be different from our preferences. For example, we may also value today more than tomorrow(discounting) or place partial value on the welfare of another(sympathy).

On a purely economic level, we can consider the evidence from ultimatum games where one party rejects the offer it is deemed too low [Fehr and Schmidt, 1998] or cooperative dilemmas like the Prisoner's Dilemma where despite incentive to defect, those experimented on will cooperate [Fehr and Gachter, 2000]. Our motivation for investigating and deriving such preferences is that we want explanatory power for linguistic interchanges that occur across a multitude of relationship types, be they partitioned across quality (dominance, reciprocity, and communality) or quantity (one-shot, short-term, long-term). We will begin by a general intuition of preferences before proceeding to more formal notions and examples.

By preferences, we mean mechanisms that change how an agent evaluates utility. As a folk example, observe that if I have a sympathetic preference for my friends, I will not mind giving them a ride to work despite the fact that it costs me personally. Notice that this differentiates from the concept of a norm, where the norm might be that friends *should* help each other out, lest there be a social cost imposed on them. The preference is that I *want* to help out; this originates in the agent's calculations for even a one-shot game. [1] This should also play into our considerations of dominance relationships vs. communality, where it might be the case that a norm promotes being cooperative in the dominance relationship despite a preference against it. The case of communality gives us one where there is a preference towards cooperation instead.

Let us consider two examples, motives of sympathy and spite. If I feel sympathy for someone else, not only will I value my own payoff, but theirs as well. If I feel spite for someone else, I will feel disfavorable towards their success. These two emotions can be thought to elicit *values*, which we will henceforth represent by the iconic function V. I.e. any representation of a material preference on a game's utility U will be labeled as V when expressing a valuation altered by some preference.

Here we consider sympathy under the limiting case of a single interlocutor. Given a game and a set of players, we can define a *sympathy distribution* based on

[1] Deontic Modality?

the extent to which each player values the success of the others. For each agent i, there is a distribution, $\sigma_i \in \Sigma(U)$, such that $\sum_j \sigma_i(U_j) = 1$ for each other agent j. Based on the sympathy distribution and the utility function U of the original game, we define a new utility function V. In the case of sympathy, we have [2]

$$V_i = \sigma_i(U_i) \cdot U_i + (1 - \sigma_i(U_i)) \cdot U_j$$

Keeping in mind that we will focus on a single interlocutor, we can now arrive at a simplified utility function for two agents, A and B. Let $\sigma_A(U_A) = r$ and $\sigma_A(U_B) = s$. Observe that $r + s = 1$. Thus we have :

$$V_A(s) = rU_A + sU_B$$

Ff

It could also be the case that an agent has a disutility in seeing his interlocutor succeed. This is a model of spite, or hatred, as seen in Mialon and Klump([Klumpp and Mialon, 2012]). Given a game and a set of players, we can define a *spite distribution* based on the extent to which each player disprefers the success of the others. For each agent i, there is a distribution, $\delta_i \in \Delta(U)$, such that $\sum_j \delta_i(U_j) = 1$ for each other agent j. Based on the spite distribution and the utility function U of the original game, we define a new utility function V. In the case of spite, we have[3]

$$V_i = \delta_i(U_i) \cdot U_i - (1 - \delta_i(U_i)) \cdot U_j$$

Extending our previous notation, we can also arrive at a simplified utility function for two agents , A and B, under conditions of spite. Let $\delta_A(U_A) = r$ and $\delta_A(U_B) = s$. Observe that $r + s = 1$. Thus we have :

$$V_A(s) = rU_A - sU_B$$

In both of these cases, we can observe that agents with zero sympathy or zero spite calculate their utilities exactly the same as the classical agents from game theory with straightforward utility calculations. These two mechanisms for altering utilities will show up later when we consider examples of face-saving, deception, dominance, and impoliteness. These mechanisms are not the only ones that can alter an agent's decision-process however. The rest of this chapter concerns itself with a deeper look at some of these processes, beginning with the sympathy-spite continuum and moving on to equilibria and long-term relationships.

[2]Is δ a bad choice of notation? It is used by Sally but conflicts with discounting with discounting? Use σ instead.

[3]Should we use another letter? η for hate?

Preferences vs. Norms

Besides preferences, norms are another mechanism that can alter a game's dynamics. Observe that we are interested in both the norms that drive linguistic behavior, and thus the conventionalization process, and the conventions themselves. In our discussion of norms in **??** , we saw that norms are one such mechanism that promotes the arrival of players at Pareto-optimal outcomes. Preferences differ in the following respects from norms:

- Preferences affect how agents calculate their utilities
- Norms are often subsets of the strategy space
- Preferences are internal to the agents.
- Norms are specific to a game.
- An agent cannot be punished for a preference
- An agent *can* be punished for breaking a norm/ playing a different strategy.

An open question is whether there is a causal link between preferences and norms. [4] For instance, does a norm of justice drive a preference for fairness? Or can a preference of spite counter a norm of cooperation? Does a preference for future returns, i.e. for maximized long-term utility, drives the strong reciprocity norm in the Prisoner's Dilemma? By this we mean the repeated game equilibrium of mutual *tit-for-tat* that preserves the Pareto-Optimal state of cooperation.

To remark on the preference of sympathy, we can contrast it with the norm of reciprocity. Consider the case of two friends going out for drinks. Under a sympathy preference, one friend might feel happy buying drinks for another. This in no way obligates the second towards reciprocity. Contrast this with two colleagues. If one buys drinks for the other, the norm of reciprocity between them might predict that the favor would be returned at some point. Although this is a folk example, the intuitions of sanctioning to preserve norms when sympathy does not exist will feature later when we construct restrictions on repeated games to model long-term relationships.

These are some of the few considerations that we wish to bring to the table, and each of them comport with the layered model of agency and game dynamics brought to light in works like [Cárdenas and Ostrom, 2001] and seen in Figure 4.1. This three-layer model of agency begins with the game dynamics, proceeds to the agent's group affiliation, and then finishes with the agent's individual properties. The idea is that we want to separate the various dynamics to give explanations for certain behaviors in the absence of typical circumstances. We will not investigate all of these, but this itemization gives us a baseline for understanding the various ways in which a game can change.

[4] Am I allowed to stipulate open questions?

Identity	Group Dynamics	Material Payoffs
· Wealth, Occupation, Experience	· Shared norms	· Net payoffs
· Other-regarding	· Heterogeneity	· Feasible Strategies
· Values	· Inequality	· Expected Cost
· Gender, Age, Skills, Education	· Group Identity	· Reputation
· Group Membership	· Cooperative vs. Competitive	· Reciprocity
		· Pr(Next Round)

Figure 4.1: Emendation of Model from [Cárdenas and Ostrom, 2001]: Preferences and mechanisms altering game play. Each category interacts with the next one to produce payoffs and preferences that steer probabilities of cooperation.

Further Remarks on Strategic Equivalence

We mentioned earlier that these mechanisms can transform games into other strategically equivalent forms. We now give a small discussion of why this works for 2×2 games, with reference to lecture notes by Maxwell Stinchcombe.[5]

Consider the game matrices seen here:

$$U_A = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}$$

Now consider the function $f(U_A(Left)) = U_A - R$ and $f(U_A((Right)) = U_A - Q$. We now have the following matrix, with preference orderings preserved, based on the relative scales of elements of P, Q, R, S that can be scaled and still maintain the preference orderings.

$$U_A = \begin{bmatrix} P-R & 0 \\ 0 & S-Q \end{bmatrix}; kU_A = \begin{bmatrix} k(P-R) & 0 \\ 0 & k(S-Q) \end{bmatrix}$$

For instance, in the Prisoner's Dilemma, we would still have that $P - R < 0$ and $S - Q > 0$ and thus the strict dominance of defection. Since these orderings are preserved, we could also scale the matrix as need be without disturbing the presence of the various equilibria. We can make this remark as easily for the column player, and thus we now have a constructive method for determining when preferences or external mechanisms give rise to strategically equivalent games. In general, we can construct linear functions that map each payoff for a given outcome in one game to that outcome in the other.

[5]Found at `http://www.laits.utexas.edu/~mbs31415/gtnotesF08.pdf`.

How do preferences for fairness, face, sympathy, etc. emerge?

A primary claim found in [Mailath and Samuelson, 2006] is that links between current and future behavior create incentives that would otherwise not appear in a one-shot game. One application of this stems from Sally [2002], where we see that he modulates sympathy according to the optimality of outcomes. A further application is found in Rabin [1993], where he considers fairness that can govern equilibria based on the size of the stake. In addition, we might have the case that variation in repeated interaction might make individuals update their own future discounting, another alteration to preferences.

Our perspective is often more based on the mechanism design side, where we want conditions that provide for thresholds of factors like sympathy that would push through cooperative behavior. Thus we will not address in great detail the evolution or emergence of preferences through update mechanisms. We transition now to a deeper look at sympathy and spite preferences.

4.2 Sympathy and Spite

We have emphasized the point that norms and repetition can push a cooperative dilemma with one low-paying equilibrium towards a coordination problem with multiple, and potentially more rewarding, equilibria. The same is true for preferences. We have seen this earlier with discounting in repeated games, and we will see this again in the symmetric trust game. One preference that is particularly relevant is that of sympathy. It can be triggered when two individuals know each other well or when one individual notices the comparable misfortune of another. The second kind becomes apparent when we see examples like a poor beggar on the street or an old woman carrying a heavy box. On a simple level, norms are strategy profiles for a group, whereas preferences are a way to determine an individual's strategy. We now turn to sympathy, as seen in Sally [2000, 2001]

Background on Sympathy

David Sally expounded upon the preference for sympathy when considering cooperation, coordination games, and pragmatics, and we reiterate some of his claims here. [6] We will also comment on areas in which these preferences apply.

- *Sympathy is not an emotion, but the interpersonal process of identification with another*
- *This process has a number of phases ranging from motor-mimicry to cognitive perspective-taking*
- *This process may happen non-consciously and involuntarily*
- *One outcome of the process is an enlarged self-interest*

[6] Claims on sympathy from [Sally, 2000] (verbatim, p.575)

Open Questions on Preferences and Sympathy

Given these claims, we could investigate several questions. Why are sympathy, fear(risk-aversion), altruism, benevolence not norms but preferences? [Bicchieri, 2006, Sally, 2000]. In contrast however, is fairness a preference that gives us justice? By this we mean that a preference for outcomes that are close to each other might drive strategic profiles that divide utility evenly into equilibrium. How does fairness interact with sympathy? Can such a preference in utility calculations trigger the emergence of a norm? Sympathy (and other preferences, e.g. hatred) form more general utility functions, but how do we incorporate knowledge of these? Are these subject to exploitation? How should we integrate Social Distance & Sympathy [Goffman, 1967, Brown and Levinson, 1987] to square these with Sally's notions? Although we cannot answer all of these questions here, we begin with variants of the Prisoner's Dilemma and sympathy.

Self-Regard and Sympathy: An Example

If we take the Prisoner's Dilemma and a variant on self-regard vs. sympathy, we can construct the following example. First, we follow notation used in works like Skyrms [2004] reduce a game table in normal form to a payoff matrix for each player A and B.

$$U_A = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}; \qquad U_B = \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}$$

Within an interaction, an agent will have some amount of self-regard r and sympathy s for his opponent. Note two things: 1) the case where $r = 1$ is the classical notion of utility, and 2) it will be the case that within a dyad $r + s = 1$. Under a single condition of self-regard r_A and sympathy s_A, we can now see the following new calculation for an agent A's utility.

For the moment, we should remark that we need not always operate under the assumption that the sympathetic utility distribution is symmetric; i.e. we could assume that A and B do not have identical sympathy preferences for each other. A classic motivation for such asymmetry might be considering the dynamic between parents and children, where parents might have a higher degree of sympathy towards their children than vice versa.

Within this two-person example, we will begin with identical sympathy conditions for both players, namely saying that $s_A = s_B$. With that in mind, we will use r and s to reflect these values of self-regard and sympathy. Note that this gives us a new utility function V. [7]

$$V_A(s) = rU_A + sU_B = r \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix} + s \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3r + 3s & 0r + 5s \\ 5r + 0s & 1r + 1s \end{bmatrix}$$

[7] Consider V as a stand-in for *value*.

Note further that we are operating under the assumption that $r + s = 1$ and that this game is symmetric. Thus, we have the following payoff matrices for A and B.

$$V_A(s) = \begin{bmatrix} 3 & 5s \\ 5r & 1 \end{bmatrix}; V_B(s) = \begin{bmatrix} 3 & 5r \\ 5s & 1 \end{bmatrix}$$

As we have that a sympathy value $s = 0$ gives us the classical utility model, let us now take the extreme value where $s = 1$. Thus we have,

$$V_A(1) = \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}; V_B(1) = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}$$

Notice now that this generates the normal form game seen in Table 4.1 :

	C	D
C	3;3	5;0
D	0;5	1;1

Table 4.1: Prisoner's Dilemma with 100% sympathy. Pareto-optimal Nash equilibrium of (C, C)

As a result of being tied to the other's outcome, the strategy of cooperation strictly dominates that of defection. We have now made the Prisoner's Dilemma a sort of non-dilemma, where the Nash Equilibrium is also Pareto-optimal. More interesting however, is when we investigate what happens between total self-interest and total sympathy. We want to compare the values from each matrix, noting when we have incentives to switch strategy. I.e. we want to know for what values of s do we have that $V(C) > V(D)$ or $V(D) > V(C)$.

$$V_A(0) = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}; V_A(s) = \begin{bmatrix} 3 & 5s \\ 5(1-s) & 1 \end{bmatrix}; V_A(1) = \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}$$

Figure 4.2: Case I: $V(C) > V(D)$; **Case II:** $V(D) > V(C)$

Consider the case seen in Figure 4.2. The inequalities $3 > 5(1-s)$ and $5s > 1$ give us the solutions of $s > .4$ and $s > .2$. The inequalities $5(1-s) > 3$ and $1 > 5s$ give us the solutions of $s < .4$ and $s < .2$. With that said, however, there is an interior region where $.2 < s < .4$. This is a regions where neither cooperation nor defection strictly dominates the other. If for example, we plug in a value of $s = .3$, we obtain the following matrices:

$$V_A(.3) = \begin{bmatrix} 3 & 1.5 \\ 3.5 & 1 \end{bmatrix}; V_B(.3) = \begin{bmatrix} 3 & 3.5 \\ 1.5 & 1 \end{bmatrix}$$

This results in the following game that we claim to be strategically equivalent to Hawks & Doves. For the particular value of $s = .3$, we have a mixed strategy equilibrium of (.5,.5), i.e. a percentage of cooperation 50% of the time. For the more general instance of an unknown sympathy value s, we have the mixed strategy equilibrium where

$$\begin{aligned} EU(C) &= EU(D) \\ 3p + 5s(1-p) &= 5(1-s)p + 1(1-p) \\ p &= 5s - 1 \end{aligned}$$

Note that we claimed earlier that $.2 < s < .4$ in this case. This gives us that $0 < p < 1$ as $5(.2) - 1 = 0$ and $5(.4) - 1 = 1$.

$s = .3$	C	D
C	3;3	1.5;3.5
D	3.5;1.5	1;1

	D	H
D	1;1	2;7
H	7;2	0;0

Table 4.2: Strategic equivalence between Hawks & Doves and modified Prisoner's Dilemma with sympathy value $s = .3$. Here we see the two anti-coordination equilibria.

As to one motivation for why we are interested in the transformation of incentives under sympathy, we are interested in a predictive theory as to why we might hear overly apologetic or demanding language in scenarios that are *prima facie* cooperative dilemmas. This could be a result of an inbalance/ abnormality in sympathy on the part of one speaker or both speakers, as situation of reciprocity might appear to become one of dominance. As we also connect this to the hormonal drivers of preferences [Zak et al., 2005, Zak, 2011], one such prediction might include these cases appearing when higher levels of oxytocin are present (e.g. close friends, partners).

Sympathy and the General Prisoner's Dilemma

In later chapters, we will see that the general Prisoner's Dilemma will appear as a subset of the strategies in the variants of the trust game. Here we examine sympathy preferences and equilibria in the generalized version of the Prisoner's Dilemma. We begin with comparing the payoff matrices under the standard utility $U = V(0)$ function and the sympathetic one V:

$$V_A(0) = \begin{bmatrix} b-c & -c \\ b & 0 \end{bmatrix}; V_A(s) = \begin{bmatrix} b-c & -c(1-s)+bs \\ b(1-s)-cs & 0 \end{bmatrix}$$

We could also write this as:

$$V_A(s) = \begin{bmatrix} b-c & -c+s(b+c) \\ b-s(b+c) & 0 \end{bmatrix}$$

What keeps the column payoffs separated? We need

$$b - s(b+c) > -c + s(b+c)$$
$$\Rightarrow s < \frac{1}{2}$$

This should not surprise us, as caring more about the other's payoffs switches one agent's preferences to that of his partner. This means that we can preserve the general order of cooperation being dominated by defection as seen in the Prisoner's Dilemma. Any more sympathy leads to a shift.

Stag Hunts and Strategic Equivalence

If we want to know the constraint of sympathy by which the actions C and D are bests responses to themselves, we have that $V_A(CC, s) > V_A(CD, s)$ or $b-c > b(1-s)-cs$. This might give us a second Pareto-optimal outcome as in the Stag Hunt. This means we also want $V_A(DD, s) > V_A(DD, s)$ or $0 > -c+s(b+c)$. These constraints give us

$$s > \frac{c}{b+c}; s < \frac{c}{b+c}$$

Is this incongruous? Does this mean that we should set $s = \frac{c}{b+c}$? In the case below we have that $b = 3, c = 1 \Rightarrow s = \frac{1}{4} = .25$ as seen in Table 4.3.

$$V_A(0) = \begin{bmatrix} 2 & -1 \\ 3 & 0 \end{bmatrix}; V_A(.25) = \begin{bmatrix} 2 & 0 \\ 2 & 0 \end{bmatrix};$$

	C	D
C	2;2	0;2
D	2;0	0;0

Table 4.3: Weak Stag Hunt with four pure strategy Nash equilibria; $s = .25$.

In general, this should give us a weak variant of the Stag Hunt seen in the game Table 4.4. By *weak* we mean that the agents will not strictly improve their scores by deviating, although it should be clear that CC is the Pareto-optimal state.

	C	D
C	b-c;b-c	0;b-c
D	b-c;0	0;0

Table 4.4: Weak Stag Hunt with four pure strategy Nash equilibria; $s = \frac{c}{b+c}$.

If our goal is to keep the off-diagonal payoffs under the outcomes CD, DC separated, we might want to ensure that $\frac{c}{b+c} < s < min(\frac{b}{b+c}, \frac{1}{2}$. As $b > c$, we have that $\frac{b}{b+c} > \frac{1}{2}$, so we should further refine this as

$$\frac{c}{b+c} < s < \frac{1}{2}$$

Notice that as we have a very tight connection between the payoffs of $b, c, b-c$ that we will not necessarily have an emergent Stag Hunt in this normalized version. With a more generalized setting, as seen later in Table 4.8, this is possible.

Hawk-Dove and Strategic Equivalence

On the other hand, we might want to know when C becomes a best response to D as in Hawks & Doves. This occurs when $V_A(DC, s) > V_A(CC, s)$ and $V_A(CD, s) > V_A(DD, s)$. This gives us two inequalities:

$$b(1-s) - cs > b - c$$
$$\Rightarrow -c(1-s) + bs > 0$$

This once again would give us the seemingly contradictory constraints of $s > \frac{c}{b+c}$ and $s < \frac{c}{b+c}$, surprisingly a repeat of the previous conditions found for generating a Stag Hunt in the previous section. The problem again here is the too-tightly connected payoffs of the players. In other words, with the right amount of sympathy, we could get something like a weak coordination problem, and with too much we could get anti-coordination. These simple parameters on b and c suggest we should go to the more generalized version of the Prisoner's Dilemma seen in the literature on symmetric games found in section 4.3.

Hate and the Prisoner's Dilemma: An Example of Spite

We have seen the effect of sympathy on the Prisoner's Dilemma as our cardinal example. We now turn to see how a spite motive can affect the preferences and outcomes of this famous problem of cooperation. We adopt notation as before, assuming that an agent has individual self-regard r and spite s. Thus for two players A and B we have

$$V_A(s) = rU_A - sU_B$$

Observe how this tranforms the Prisoner's Dilemma across the gradient of spite:

$$V_A(s) = rU_A - sU_B = r\begin{bmatrix}3 & 0\\ 5 & 1\end{bmatrix} - s\begin{bmatrix}3 & 5\\ 0 & 1\end{bmatrix} = \begin{bmatrix}3r-3s & 0r-5s\\ 5r+0s & 1r-1s\end{bmatrix}$$

We can substitute $r = 1 - s$ and obtain

$$V_A(s) = \begin{bmatrix}3r-3s & 0r-5s\\ 5r+0s & 1r-1s\end{bmatrix} = \begin{bmatrix}3(1-s)-3s & -5s\\ 5(1-s) & 1(1-s)-1s\end{bmatrix} = \begin{bmatrix}3(1-2s) & -5s\\ 5(1-s) & 1(1-2s)\end{bmatrix}$$

Just as before, we can interest ourselves in several cases. The first is the trivial case of zero spite, equivalent to the classical utility function. The second is total hatred, the case where $s = 1$. The last case involves the boundary points of intermediate spite, between which spite provides interesting incentives. In the case where $s = 0$ and $s = 1$, we have:

$$V_A(1) = \begin{bmatrix}-3 & -5\\ 0 & -1\end{bmatrix}; V_B(1) = \begin{bmatrix}-3 & 0\\ -5 & -1\end{bmatrix}$$

Notice now that this generates the normal form game:

	C	D
C	-3;-3	-5;0
D	0;-5	-1;-1

Table 4.5: Prisoner's Dilemma with 100% spite value. Nash equilibrium and Pareto-optimal state of (D, D).

Notice that this removes the dilemma from the Prisoner's Dilemma. Moreover, it magnifies the disincentive to cooperate. As there is no change to the game's equilibrium and the spite's effect on the change in utility goes along a monotonically decreasing path, the intermediate value theorem predicts no such in-between point of interest where the game should be transformed. This result comports with the more general results found in Klumpp and Mialon [2012] on arms-race style games with spite motives. Their result showed that spite motives in games of strategic substitutes exacerbate the perverse incentives towards defection. This motive should provide interesting results when we begin with games like the Stag Hunt and proceed towards other strategically equivalent scenarios. We would like to use this mechanism in later chapters for discussing the presence of impoliteness when the nominal exchange is cooperative.

Stag Hunt and Spite Motives

Converse to the sympathy motive, a spite motive can transform a situation with a Pareto-optimal coordinative outcome into one with potentially dominated cooperation. Consider the Stag Hunt below with a spite-generated payoff.

$$V_A(s) = \begin{bmatrix} 4r - 4s & 0r - 2s \\ 2r + 0s & 1r - 1s \end{bmatrix} = \begin{bmatrix} 4(1-s) - 4s & -2s \\ 2(1-s) & 1(1-s) - 1s \end{bmatrix}$$
$$= \begin{bmatrix} 4(1-2s) & -2s \\ 2(1-s) & 1(1-2s) \end{bmatrix} = \begin{bmatrix} 4 - 8s & -2s \\ 2 - 2s & 1 - 2s \end{bmatrix}$$

Notice that defection is already a best response to itself in the second column, as $1 - 2s > -2s$ for all values. We thus consider where we could have defection outcompeting cooperation in the first column:

$$2(1-s) > 4(1-2s)$$
$$\Rightarrow s > \frac{1}{3}$$

Thus a sufficient spite value can reduce the Stag Hunt to a Prisoner's Dilemma. As an example, consider below where $s = .4$. This could factor in where a group norm could produce a Stag Hunt from a Prisoner's Dilemma but an individual preference could erase that incentive towards cooperation.

	C	D
C	.8;.8	-.8;1.2
D	1.2;-.8	.2;.2

Table 4.6: Stag Hunt with 40% spite value now equivalent to a Prisoner's Dilemma. Nash equilibrium of (D, D) and a strictly dominant strategy of D.

4.3 Sympathy and Symmetry

In the examples above, we saw the instances of sympathy on a classic symmetric game, the Prisoner's Dilemma. We now return to our formulation of sympathy and symmetric games as seen in our discussion on basic game theory as found in **??**. To recap:

We get the Prisoner's Dilemma when we have $R > P > S > Q$, the Stag Hunt if we have $P > R > S > Q$, and Hawks & Doves if it is the case that we have $R > P > Q > S$. While some of these inequalities need not be strict, for the moment we will keep them as such. Let us now compare these constraints on the utility tables:

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

Table 4.7: Schema for a symmetric game

$$PD : R > P > S > Q \tag{4.1}$$
$$SH : P > R \geq S > Q \tag{4.2}$$
$$HD : R > P > Q > S \tag{4.3}$$

Now we will tackle the application of sympathy to explore how it could alter the incentives in these general formulations of the games. First we consider that it is always the case that the responses against the first column dominate the second. We will now compare the original payoff matrix for a player X against one with sympathy value S and associated utility function $V_X(s) = (1-s)U_x + sU_Y$. If we begin with the original payoff matrix under U_X, we can then construct the payoff matrix for V_X.

$$U_X = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \end{matrix}; U_Y = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} P & R \\ Q & S \end{bmatrix} \end{matrix}$$

Note that in the event that we have $s = .5$, we have an identical payoff in the off-diagonal entries. At this point, the Nash Equilibrium simply becomes determined by which is greater among $P, \frac{R+Q}{2}, S$. This is often the Pareto-optimal value of cooperation in the case of the Prisoner's Dilemma. For values greater than that, the payoff matrix reverses its middle entries in the off-diagonal.

$$V_X = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} P & \frac{R+Q}{2} \\ \frac{R+Q}{2} & S \end{bmatrix} \end{matrix}$$

Stag Hunt from Prisoner's Dilemma

We now attempt to form the Stag Hunt where $V_X(CC) > V_X(DC)$ beginning from the assumption that $R > P > S > Q$ as in the Prisoner's Dilemma. For the payoff matrix affected by a sympathy parameter s, we denote the altered payoff matrix as $M_X(s)$

$$M_X(s) = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} P & Q(1-s) + Rs \\ R(1-s) + Qs & S \end{bmatrix} \end{matrix} = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{bmatrix} P & Q + s(R-Q) \\ R + s(Q-R) & S \end{bmatrix} \end{matrix}$$

We want to consider cases where we input a given outcome-generated payoff like Q into the sympathy conditions and output a new payoff $f(Q)$. If we think of $f(Q) = Q + s(R - Q)$ and likewise $f(R) = R + s(Q - R)$,this means we compute where

$$\begin{aligned} &P > R(1-s) + Qs; \\ &R(1-s) + Qs > S; \\ &S > Q(1-s) + Rs \end{aligned}$$

As we note that $Q - R < 0$, this gives us

$$\begin{aligned} s &> \frac{R-P}{R-Q} \\ s &< \frac{S-R}{Q-R} \\ s &< \frac{Q-S}{Q-R} \end{aligned}$$

What does this mean? We could rearrange these conditions to

$$\begin{aligned} s &> \frac{R-P}{R-Q} \\ s &< \frac{R-S}{R-Q} \\ s &> \frac{Q-S}{R-Q} \end{aligned}$$

This first inequality seems to be the most salient, as we are comparing the difference between $U_X(DC)$ and $U_X(CD)$ and the difference between $U_X(DC)$ $U_X(DC)$. In the denominator are the two normally farthest apart. As they get further apart, the advantage of defection grows, and the sympathy required to promote cooperation goes down interestingly. As $U_X(DC)$ and $U_X(CD)$ get closer together, the sympathy required also goes down.

Now we consider the last inequality. Considering here we have a bound on sympathy that will keep the last column in line $s < \frac{R-S}{R-Q}$. This compares the ratio of the payoffs of playing D to the difference in the asymmetric profiles. Just as before, for a general Prisoner's Dilemma with $R > P > S > Q$, a degree of sympathy as outlined below will force the game into a Stag Hunt:

$$\frac{R-P}{R-Q} < s < \frac{R-S}{R-Q}$$

Further Conditions on the Sympathetic Stag Hunt

If we consider the off-diagonal payoffs Q, R in the standard symmetric games above, we can consider cases where $Q > R$ or $Q < R$. Notice that in the three canonical games of the Prisoner's Dilemma, Stag Hunt, and Hawks & Doves, we have that $Q < R$ in every case, motivated by the implicit risk of cooperation or temptation of defection.

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

Table 4.8: Schema for a symmetric game

If we begin with our prior assumptions, we can extend the case to find sympathy constraints that will turn any game into a Stag Hunt by satisfying

$$P > (1-s)R + sQ;$$
$$S > (1-s)Q + sR$$

Depending on whether $Q < R$ or $Q > R$, we can get different results. Given that $Q < R$, we obtain

$$\frac{P-R}{Q-R} < s < \frac{S-Q}{R-Q}$$

Given that $Q > R$, we obtain

$$\frac{S-Q}{R-Q} < s < \frac{P-R}{Q-R}$$

These constraints are meaningless however on the lower bound, as $0 < s < 1$ in meaningful cases. Thus we have that

$$Q < R \Rightarrow 0 < s < \frac{S-Q}{R-Q}$$

and

$$Q > R \Rightarrow 0 < s < \frac{P-R}{Q-R}$$

Spite and Generating the Prisoner's Dilemma

It could also be the case that we want conditions for generating a strictly dominant strategy that nonetheless results in sub-optimal payoffs. This could be done under a spite motive where

$$V_A(s) = (1-s)U_A - sU_B$$

Once again consider the symmetric game scenarios, let us compare the standard symmetric game with a spiteful counterpart:

	C	D
C	P(1-2s) ,P (1-2s)	Q-s(R+Q),R Q-s(R+Q)
D	R -s(R+Q),Q-s(R+Q)	S(1-2s),S(1-2s)

Table 4.9: Schema for a symmetric game with spite motivations

Thus we want the case for the action D to strictly dominate C, regardlesss of the initial parameters. This means we should solve the inequalities where $R - s(R+Q) > P(1-2s)$ and $S(1-2s) > Q - s(R+Q)$. This gives us two conditions on spite:

$$s > \frac{P-R}{2P-(R+Q)}; s < \frac{Q-S}{2S-(R+Q)};$$

In the case of the Stag Hunt, we have that $P > R \geq S > Q$, thus re-working the negatives gives us $s > max(\frac{S-Q}{2S-(R+Q)}, \frac{P-R}{2P-(R+Q)})$ will be sufficient.

These are some of the ways in which sympathy and spite can transform a symmetric game. This is in the one-shot scenario, so naturally to connect this to previous work on repeated games, we use sympathy as a baseline for understanding the relationship of communality.

4.4 Links Between Sympathy and Repeated Games

What we would like to achieve is a variant of strategic equivalence governed by sympathy preferences and repeated game mechanisms. The goal would be to unite the internal other-regarding preferences with external sources of cooperation like mechanisms of repetition. We would also like to understand how repeated interaction between friends, acquaintances, colleagues, or adversaries might generate decision patterns not predicted by their material payoffs.

Sympathy can be viewed in two ways: a weighting between the interests of both players or a shift away from my own preferences governed by inequality. Note that as before we have an equivalent version of the convex combination:

$$(1-s)\begin{bmatrix} P & Q \\ R & S \end{bmatrix} + s\begin{bmatrix} P & R \\ Q & S \end{bmatrix} = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} + s\begin{bmatrix} 0 & R-Q \\ Q-R & 0 \end{bmatrix}$$

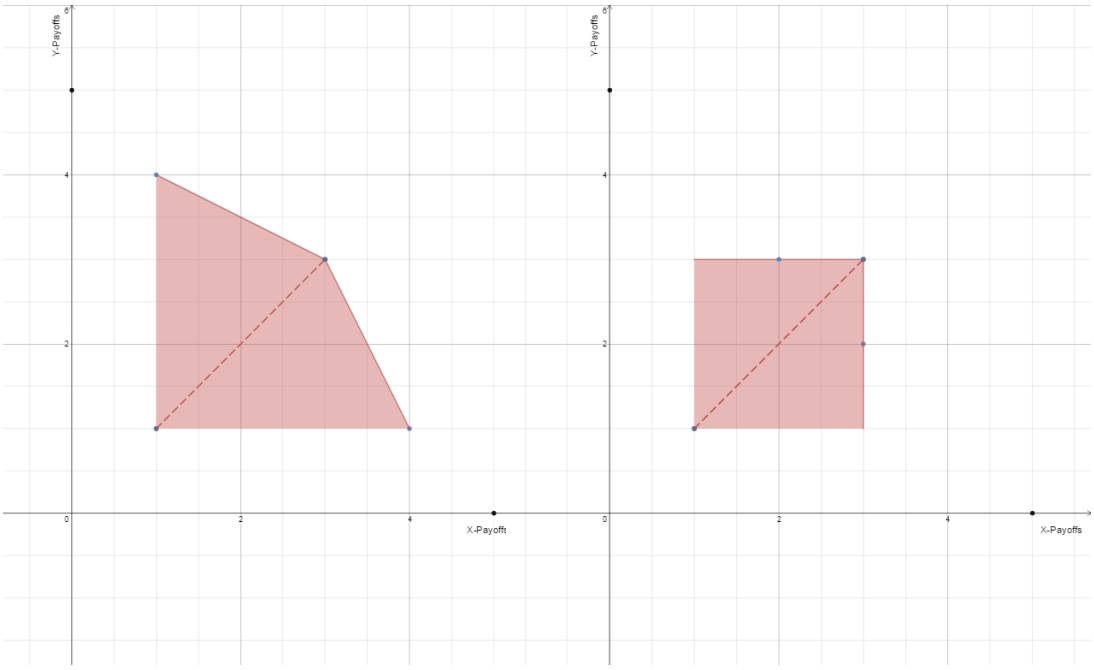

Figure 4.3: Prisoner's Dilemma with sympathy of $s = .2$ and $s = .4$. The black dots represent the original off-diagonal payoffs from CD and DC. Notice how this region of feasible payoffs is slightly smaller than the region predicted by the folk theorem, which would be formed by connecting the black dots to the symmetric outcomes.

The convex combination of self- and other-regard gives possible payoffs in a repeated game forming a region similar to that of the repeated game under the folk theorem. For instance, we can see the figures given in Figure 4.3 to note that the convex hulls of payoffs generated by the points in between the off-diagonal outcomes resemble those seen in the Folk Theorem. We can further see the connection between the off-diagonal outcomes generated by by the line segment between the two points in Figure 4.4. We can contrast this with the convex hull of payoffs for a repeated Stag Hunt seen here in Figure 4.5.

Convex Combinations and Convex Hulls

For background on the previous paragraphs, a convex combination is a way of generating a new vector from a weighted sum of others. In particular, where $\forall \alpha_i \in \alpha_1, \ldots, \alpha_n, 0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^{n} \alpha_i = 1$, we can term the following a *convex combination* of the vectors in $\vec{v_1}, \ldots, \vec{v_n}$:

$$\vec{v} = \sum_{i=1}^{n} \alpha_i \vec{v_i}$$

Notice now that in the case of sympathy, we have just such a property, as both s and $r = 1 - s$ satisfy $r + s = 1$ and $0 \leq r, s \leq 1$. Geometrically, this gives us a way to visualize the payoffs in-between the off-diagonal outcomes. This boundary

could be useful for modifying results from the folk theorem on repeated games. So we now want conditions that gives us boundaries governing the transitions between categories of strategically equivalent scenarios.

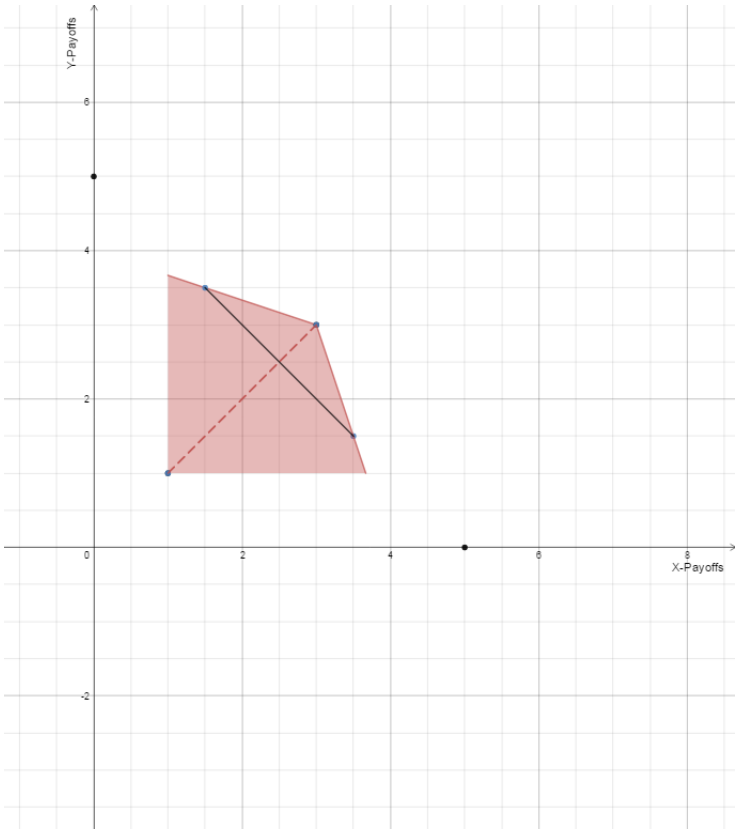

Figure 4.4: A Prisoner's Dilemma with sympathy of $s = .3$. Note the line connecting payoffs in the uncoordinated outcomes that would, if extended, reach the payoffs from original unsympathetic outcomes.

The geometric nature of the weighted sympathy parameter gives us a tie-in to the outcomes of the game and their perceptions in the minds of the players. As sympathetic payoffs are convex combinations of the originals, this means we can achieve every payoff combination between the two off-diagonal outcomes. Consider the table in Table 4.10. Although combinations of sympathy keep the symmetric outcomes the same, we can reach any outcome on the line between (R, Q) and (Q, R) with a sympathy parameter s where $0 \leq s \leq \frac{1}{2}$. This means that we can achieve any outcome where $V_X + V_Y = R + Q$.

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

Table 4.10: Schema for a symmetric game.

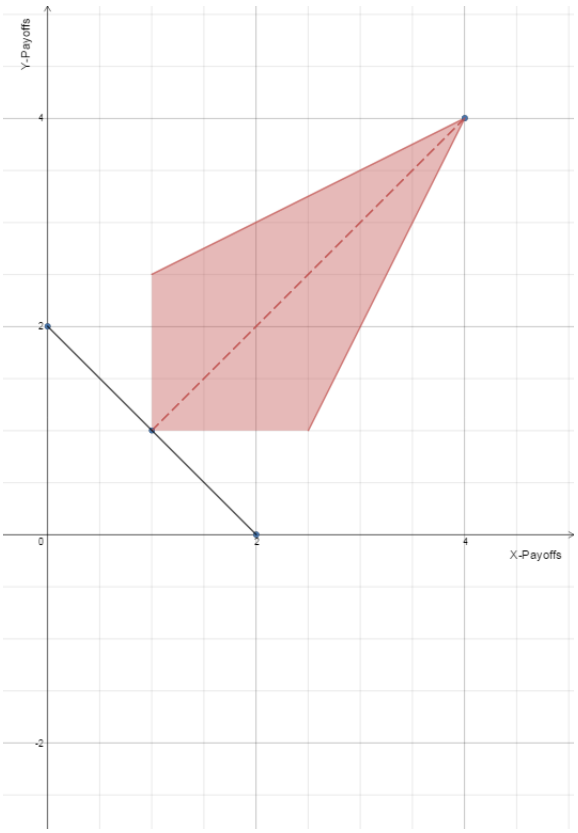

Figure 4.5: A Stag Hunt with sympathy of $s = 0$; i.e. the standard feasible region for the repeated model. The feasible region of repeated game payoffs is the darker shaded region. Note the line connecting payoffs in the uncoordinated outcomes.

Sympathy and Discounts

A crucial idea in the coming chapters is that a sympathetic preference can alter both the one-shot and repeated versions of a game. One way it alters the repeated game is that the discount δ becomes a function of the sympathy value. This allows us to

- Determine the effect sympathy has on promoting cooperative outcomes
- Examine the influence of the sympathy parameter on the minimum discount required to have a cooperative outcome
- Reverse-engineer the sympathy required to have a feasible discount of $\delta < 1$
- Explore the effect of sympathy on the perceived feasible region of outcomes in a repeated game as contrasted to the actual payoffs.

As an example, we consider the repeated Prisoner's Dilemma with sympathy. We want to see whether the patience threshold δ is more forgiving with sympathy and what level of sympathy would promote a feasible discount. Here we see the row player's utility table for the original game and the game with sympathy.

$$V_A(0) = \begin{bmatrix} b-c & -c \\ b & 0 \end{bmatrix}; V_A(s) = \begin{bmatrix} b-c & -c(1-s)+bs \\ b(1-s)-cs & 0 \end{bmatrix}$$

We proceed according to a grim trigger strategy. In the original game, we want that consistent cooperation outperforms defecting before being punished. I.e.

$$\begin{aligned} b-c &> (1-\delta)(b+\delta\sum_{j=0}^{\infty}0\delta^j) \\ \Rightarrow \delta &> \frac{c}{b} \end{aligned}$$

In the case of sympathy, we have:

$$\begin{aligned} b-c &> (1-\delta)(b+\delta\sum_{j=0}^{\infty}0\delta^j) \\ \Rightarrow \delta(s) &> \frac{c-s(b+c)}{b-s(b+c)} \end{aligned}$$

What we now have is that $\delta(s) < \delta \Leftrightarrow c < b$, which is true for all values of s and any game conforming to the Prisoner's Dilemma and variants of it. This means that we require less patience of sympathetic players. For example, given the parameters in the game from Figure 4.4, we have a decrease in the discount value from $\delta = .4$ to $\delta(.1) \approx .302$. The next question is what sympathy value makes a future discount like δ possible.

$$1 > \delta(s) \Rightarrow 1 > \frac{c-s(b+c)}{b-s(b+c)} \Rightarrow c < b$$

This is always true, and this is in fact equivalent to the same condition for $s = 0$ in the original Prisoner's Dilemma repeated. Later, we will see cases where the sympathy value can affect the discount.

4.5 Relationships and Decision Rules

Relationships give us boundaries as to what is acceptable behavior. Agents in communality relationships aim for group benefit, agents in dominance relationships favor one partner over the other, and agents in reciprocity relationships aim for fairness. We can now integrate this into a predictive decision procedure for how agents should account for individuals with whom they have a given relationship. Grinberg et al. [2012] Fiske [1992] Pinker et al. [2008].

As each relationship type specifies a further partition of the feasible region, we can divide the region into acceptable and unacceptable outcomes. Communality

accepts outcomes that favor the group utility $U_X + U_Y$. Dominance favors one agent over another $U_X > U_Y$ or $U_X < U_Y$. Last, reciprocity accepts outcomes within a certain distance from $U_Y = U_X$. The decision rules for accepting or rejecting an outcome according to the relationship can be boiled down to:

- If (U_X, U_Y) is **NOT** in the specified region of the relationship, reject this outcome in the one-shot game.
- If (U_X, U_Y) is **NOT** in the specified region of the relationship, accept this this outcome in the repeated game provided
  - there exists an enforceable equlibrium path towards an acceptable normalized outcome in the feasible region
  - sufficently patient players.
- If (U_X, U_Y) is in the specified subset of the feasible region, accept it in the one-shot or repeated game.

Communality and Amicable Equilibria

Here we define a new notion for subsets of equilibria: *amicable equilibrium*. This notion appears in the intersection of two contexts: repeated games with sympathetic payoffs. The general idea is that agents in a communality relationship like close friends should consider their long-term relationships and how the other agent is doing. This means geometrically the convex hull of their discounted, repeated game payoffs against the convex combination of points between their respective off-diagonal outcomes.

Definition 1. *We say a strategy profile is a* ***k-amicable outcome*** *if the sum of the payoffs $U_X + U_Y \geq k$. A strategy profile is a* ***k-amicable equilbrium*** *if it falls within the feasible region of equilibrium payoffs subject to the folk theorem.*

We see in Figure 4.6 an example of an amicable equilibrium. This is a refinement of the folk theorem, and one that should suffice in cases of sympathetic payoffs. This happens when we remove payoffs worse than those along the line connecting the off-diagonal payoffs, something we could call *sympathetic truncation*. Notice that the only pure strategy Nash Equilibrium surviving the sympathetic truncation would be CC. The motivation here is that the set of outcomes more favorable than the minmax outcome may still leave too much to chance in the case of close associates. Here we have a restriction for those in close contact who would want to improve the outcome of the other.

To the contrary, observe that in the Stag Hunt both Nash equilibria survive the sympathetic truncation, as seen previously in Figure 4.5. That this game has two coordinative equilibria could be the reason. Its model of an already implicit social contract also gives us an instance of why amicability is not required to push towards optimal outcomes and that all feasible outcomes according to the folk theorem are amicable equilibria. The notion should feature in later chapters where we derive cases of cooperation in symmetric trust games that preserve cooperative outcomes in a more complex setting. It is also a constraint in the cases

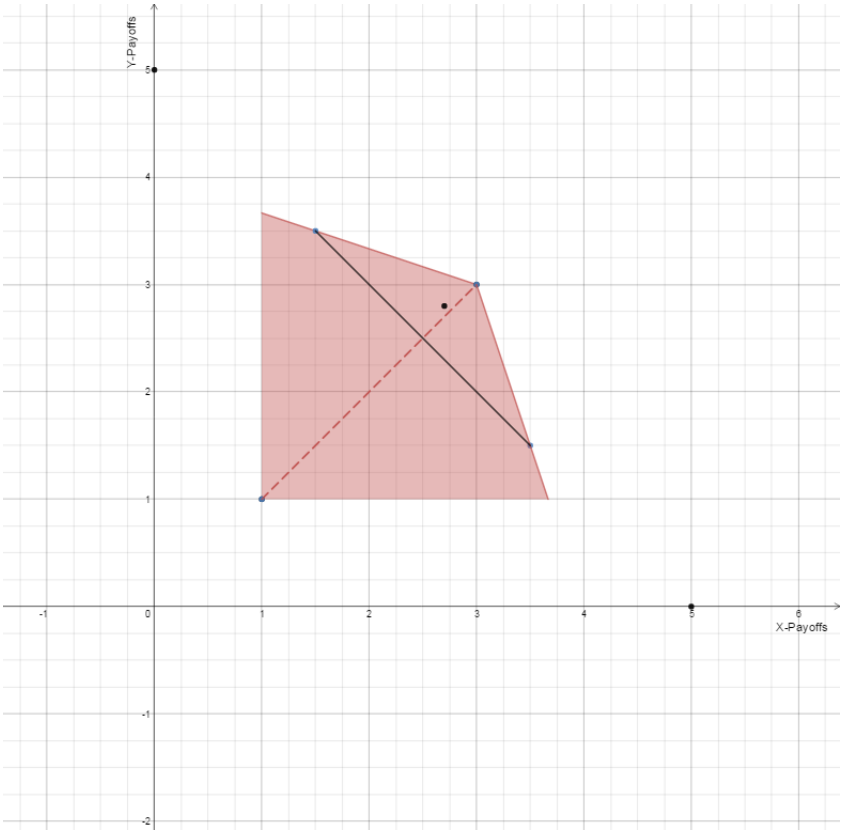

Figure 4.6: A Prisoner's Dilemma with sympathy of $s = .3$. Note the line connecting payoffs in the uncoordinated outcomes that would, if extended, reach the payoffs from original unsympathetic outcomes. The point highlighted of $(V_X, V_Y) = (2.7, 2.8)$

of long-lived sympathy that rules out exploitative outcomes where one partner might defect on another.

To remark on Pareto-efficiency, we mean that for a point P_1 on the line segment showing sympathetic payoff profiles , an amicable equilibrium P_2, and the Pareto-frontier l_P, it is always the case that $d(P_2, l_P) < d(P_1, l_P)$, where we measure distance from a point to a line in the usual Euclidean sense. I.e. the set of intermediate payoffs predicted by sympathetic outcomes give a boundary for amicable equilibria. Amicable equilbria should be further from the origin than these outcomes.

4.6 Power, Face, and Status

In addition to the weight of an FTA on its own (something determined culturally), there are two further concerns given in Brown and Levinson [1987] which we wish to address: power and face. By this we mean the ability of an agent to acquire what he wants in conflict and the preference of an agent to be respected and liked.

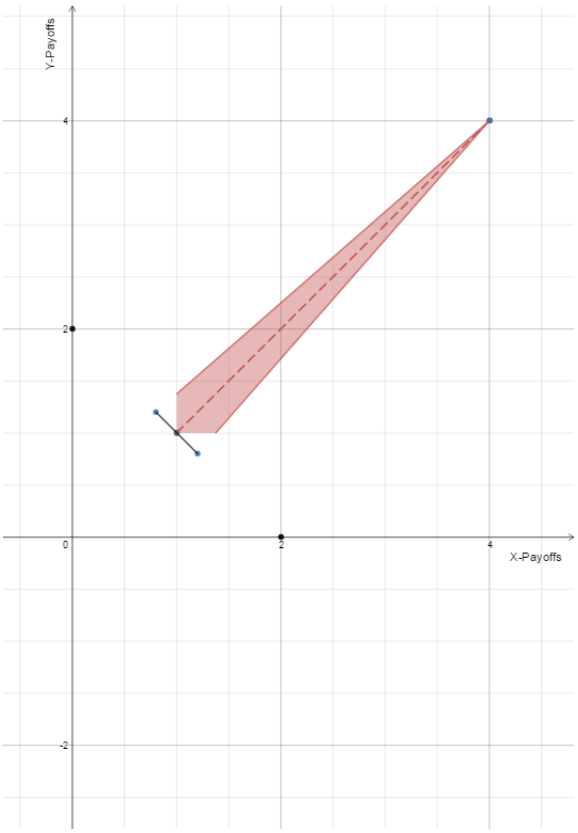

Figure 4.7: A Stag Hunt with sympathy of $s = .40$. The feasible region of payoffs is in fact the darker shaded region, noticeably smaller in this case than in the pure case. Note the sympathetic boundary line connecting payoffs in the uncoordinated outcomes is outside the feasible region. This means that any outcome in the feasible region is by default an amicable equilibrium.

Power and Bargaining

Power is a quantity that only acquires salience when there is an imbalance. We consider this to often be the case, in the sense that differences in circumstances and abilities can place one agent at an advantage to another. This features in linguistic and economic cooperation, as works like Fehr and Schmidt [1998] and Morand [2000] attest. It is the imbalance of power that leads to alliances and the need for shows of cooperation like politeness so as not to disturb it. If, as Goffman [1967] and Brown and Levinson [1987] suggest, face is negotiated, then we should understand power power through the language of asymmetric bargaining.

Definition 2. *We say agents A and B have different* ***power*** α *and* β *if the expected outcomes in a Nash bargaining game are unequal, where we have* $\frac{U_A}{U_B} = \frac{\alpha}{\beta} \neq 1$.

In a simple example, we can imagine two people negotiating a contract splitting $\$100,00$. We would expect a partner with a $3:2$ advantage to take $\$60,00$ and leave his opponent $\$40,00$. In the general case of the asymmetric Nash bargaining game, as seen in [Harsanyi and Selten, 1972], two agents with varying

bargaining powers α and β bargain for a surplus S where $\alpha + \beta = 1$. They then split between them portions t and $S - t$. Finding the equilibrium amounts to maximizing the respective Nash product given below:

$$f(t) = t^{\alpha}(S-t)^{\beta}$$

This function has its maximum where $f'(t) = 0$, which occurs when

$$\begin{aligned} 0 = \alpha t^{\alpha-1}(S-t)^{\beta} + &- \beta(S-t)^{\beta-1}t^{\alpha} \\ \frac{\beta}{S-t}t^{\alpha}(S-t)^{\beta} = &\frac{\alpha}{t}t^{\alpha}(S-t)^{\beta} \\ t = &\frac{\alpha}{\alpha+\beta}S \end{aligned}$$

In other words, we have a Nash Equilibrium where neither partner would choose to deviate when each participant has his respective share of the overall bargain. In the case of a *surplus* S, we can consider the sum of the joint utilities. For instance, in the Prisoner's Dilemma we would add the payoffs for joint cooperation $3 + 3 = 6$ or one agent cooperating $0 + 5 = 5$ to arrive at a surplus. We should expect a more powerful individual to reap a higher share in a given interaction, regardless of whether the interaction is *prima facie* reciprocal. This also comports with psychological evidence and modeling seen in works like Grinberg et al. [2012].

What this gives us is the following partitioning of a common resource S among A and B:

$$U_A = \frac{\alpha}{\alpha+\beta}S; \qquad\qquad U_B = \frac{\beta}{\alpha+\beta}S$$

We expect that agents encountering another and vying for a common resource should bargain according to these terms. Should they have equal bargaining power, we expect them to split a surplus evenly.

In the case of a cooperative dilemma like the Prisoner's Dilemma or Stag Hunt, we might see that agents split the sum of their outcomes in such proportions. For example, consider two agents in a symmetric game of respective power $\alpha = .6$ and $\beta = .4$ contesting a surplus of $S = 6$. It should be the case that we have in equilibrium:

$$U_A = \frac{\alpha}{\alpha+\beta}S = 3.6; \qquad\qquad U_B = \frac{\beta}{\alpha+\beta}S = 2.4$$

This now gives us a way to incorporate dominance into our game models, much as sympathy allowed us to incorporate communality. For instance, in the case of unevenly matched partners in Hawks & Doves, while there are two pure strategy Nash equilibria, only one would satisfy an asymmetry of payoffs according to the

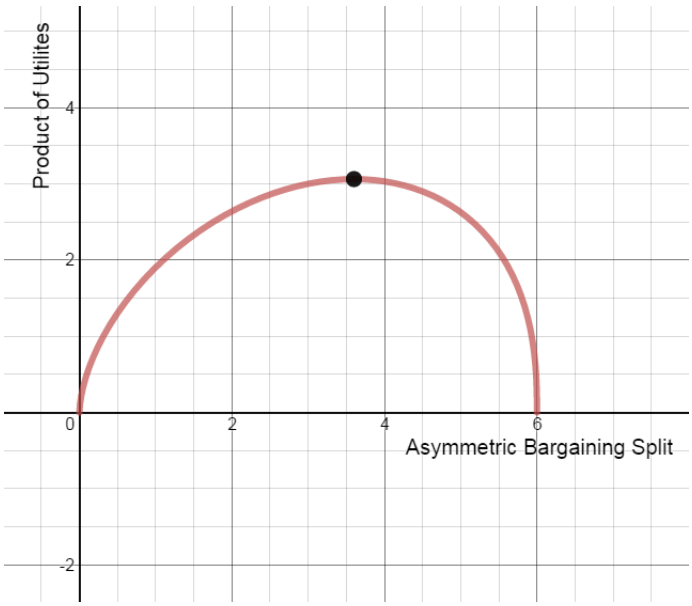

Figure 4.8: A Nash bargaining problem for a surplus of 6 with $\alpha = .6$ and $\beta = .4$. The maximum of the product of the utilities is given by $\alpha S = 3.6$ and $\beta S = 2.4$. In equilibrium, neither would deviate from this bargaining agreement, as their joint utility function would decrease.

Nash bargaining principle outlined above.

We can then adopt a similar constraint on repeated equilibria similar to the *amicable equilibria* from before. In this case, we consider the ratio of payoffs: $\frac{U_X}{U_Y} = \frac{\alpha}{\beta}$. This gives us a line predicting outcomes in the feasible region that satisfy the prescribed asymmetric ratio. Outcomes satisfying these constraints give us a new equilbrium concept: despotic equilibria.

Definition 3. *For a ratio of power R, we say an outcome in a repeated game is an* ***R-despotic outcome*** *if the respective payoff profile (U_X, U_Y) satisfies $U_Y \leq \frac{\beta}{\alpha} U_X$ for agents X and Y with respective bargaining power α and β where $\alpha > \beta$ and $\frac{\alpha}{\beta} = R$. A* ***despotic equilbrium*** *is a despotic outcome in the feasible region of repeated game payoffs subject to the folk theorem.*

If we refer to the line in Figure 4.9, we can see that outcomes in the feasible region subject to the folk theorem, outcomes both in the feasible region and under the line $y = \frac{\beta}{\alpha} x$ would be despotic equilibria. We should also mention that this concept could be up for revision in the sense that we might have a tolerance for outcomes within a certain distance of the line.

Face and Risk-Aversion to Face Threats

We next address properties of *face* and FTAs (face-threatening acts). The natural analog is that agents should be risk-neutral in most cases but risk-averse when it comes to face-threats, both on their part and on the part of others. This risk-aversion might grow proportional to the relative power imbalance of the two individuals. There are several notions of modeling this preference aversion, but

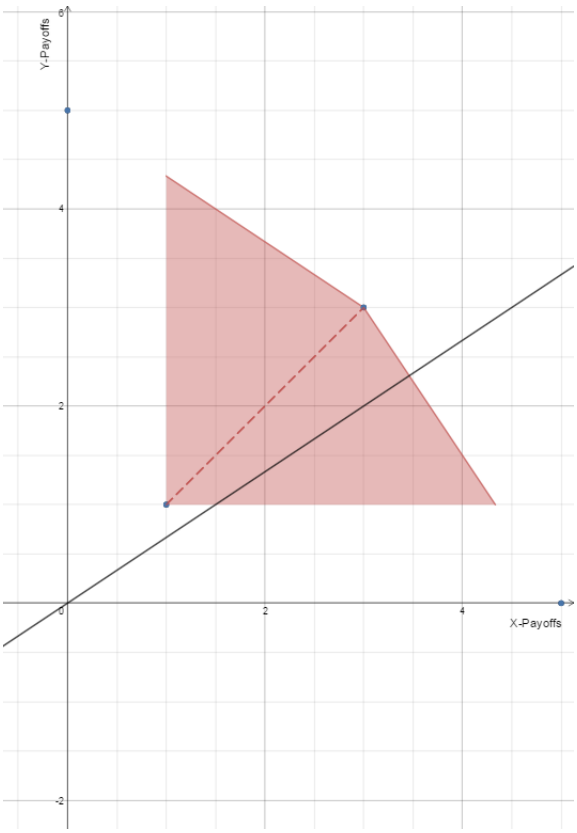

Figure 4.9: In this case we see the feasible region of the repeated Prisoner's Dilemma cut by the line $y = \frac{\alpha}{\beta}x$, where $\alpha = .6$ and $\beta = .4$. Equilibria on this line satisfy the conditions of the Nash bargaining problem.

in general, we can model a measure $A(c)$ of risk-aversion for a given outcome c in a game as

$$A(c) = -\frac{u''(c)}{u'(c)}$$

The canonical example of risk-aversion is to consider an all-or-nothing bet for \$100 constructed by flipping a coin. A risk-averse agent offered a competing choice of a \$50 surplus would always take the surplus. As risk-aversion increases, we see that a more risk-averse agent would take a sure bet c with greater probability over a lottery with the same expected payoff. A risk-neutral agent would take either, and a risk-loving agent would take the \$100 bet. As there are a multitude of functions that capture this, we will demure on further commentary at the moment.

In later chapters, we consider that face is an added parameter that emerges when agents are observed or have the potential to transmit reputational information about encounters, similar to notions of monitoring seen in [Mailath and Samuelson, 2006]. Thus we can incorporate functions that give us outcomes coherent with the preference to avoid harming the face of another in potentially

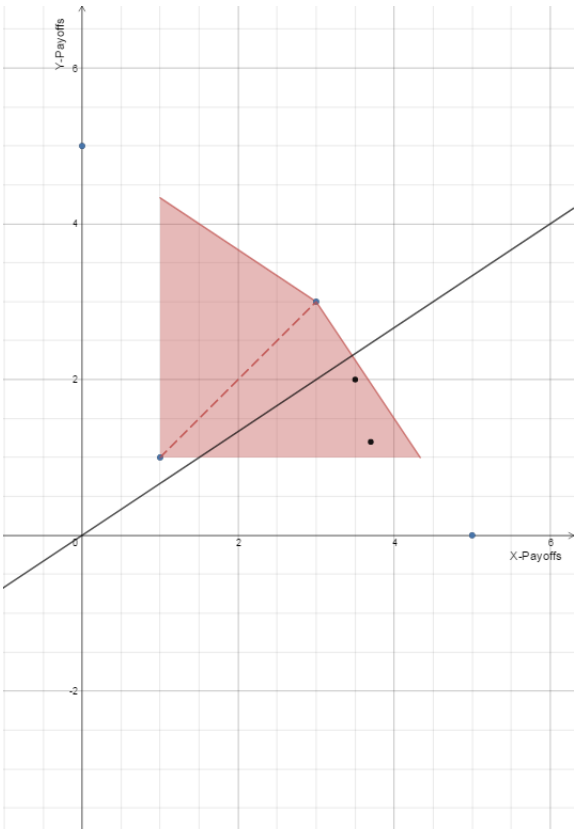

Figure 4.10: In this case we see the feasible region of the repeated Prisoner's Dilemma cut by the line $y = \frac{\alpha}{\beta}x$, where $\alpha = .6$ and $\beta = .4$. We further see a pair of despotic equlibria at the points $(3.5, 2), (3.7, 1.2)$.

repeated interactions.

Face and Utility

We take face as an example of a non-transferrable utility. It may be the case that we gain or lose face via observed and reported actions, but it is rare that an individual may gain at the loss of another, and vice versa. These conditions depend on the interaction structure and relationship. For instance, individuals in a cooperation problem like Prisoner's Dilemma have win-win, win-lose, and lose-lose outcomes.

How does reputation in repeated games square with the notion of face?

The work on *face* originated in Goffman [1967] posits a two-pronged view of social wants: the desire for autonomy (negative face) and the desire for acceptance(positive face). The importance placed on these desires varies culturally, and as such evidence of its importance crops up in the language(s) of that culture. Within a dyad, we further have social and discursive contexts that can alter the kind of strategies played, as seen in Asher and Quinley [2012].

Barring exceptional instances, individuals seek first to protect their own face.

We also claim, based on literature from philosophy and economics like Bicchieri [2002] or Bowles and Gintis [2011], that individuals further act as if they might meet another again or that someone else might be watching. That preference for cooperation in future interactions is the first driver of face maintenance and features prominently in Quinley and Ahern [2012] and Quinley [2011]. The second level of preferences exists primarily over what an agent has partial control, i.e. the indirect transmission of reputation information. As the most reliable way of a bad reputation propagating is when the violation of a norm on an agent's part is common knowledge, agents will seek to minimize making common knowledge any propositions that might damage their reputation and hence their positive face. This gives us a motivation for avoiding threatening the face of a more powerful person or a reason to act kindly towards someone for whom we have sympathy.

4.7 Fairness and Kindness

The (last) preference we will discuss is that of *fairness*. Relationships of reciprocity and strategies like *tit-for-tat* come to mind in this case. Fairness can be understood in several ways:

- A preference for utility distributed according to competence, ability, talent, work, etc.
- A preference that outcomes are not "unjustly" distributed
- A preference for a pairwise equivalence in outcomes

One question is how sympathy and fairness intersect. Do unjust histories of outcomes lead to a reduction in sympathy as seen in Sally [2002]. Does prior unfairness/ resource allocation reduce sympathy on the part of the victim or elevate sympathy on the part of the beneficiary? How does this play into asymmetric games like the trust game?

There are two prominent accounts of fairness in [Rabin, 1993] and Fehr and Schmidt [1998]. We mention them only to cover the notions that modeling fairness can be intention-oriented or behaviorally-based. We also mention them in the sense of reciprocity-based relationships, where a preference for fair treatment might sustain cooperative outcomes. Whereas Fehr's modified utility function predicts utillity loss from both advantageous and disadvantageous inequity, Rabin's function is based on the intention of the other player governed by a probabilistic belief structure.

A modeling question we can ask is whether fairness should be a preference or a norm. I.e. is it something agents should want or is it something leading to optimal outcomes that should be enforced? We will not answer this question here, but it does portend to bear future fruit for further research.

Perhaps another notion of fairness is comparing the difference in utilities versus their absolute sum. For instance, participants may not care so much about their differences if they are both in cases of high utility, as seen in evidence from experimental economics [8]. We could consider in this case the quantity:

[8] Need ref for relative hedonistic adaptation.

$$\frac{|U_A - U_B|}{U_A + U_B}$$

As this quantity decreases, so does the perception of unfairness. As this quantity levels out at zero, so would participants claim absolute fairness. Notice that as the fortunes of A and B increase, so would their interest in fairness decrease. In other words, we should expect a larger tolerance for unbalanced outcomes as the payoff profiles grow farther from the origin. For a natural analog, notice that two CEOs might not care as much about making \$16 million or \$17 million, but two workers might see a clear distinction between \$16 thousand or \$17 thousand for the same job.

Regardless of how fair a restriction of a game might be, we would like to preserve the symmetric game strategies and the ability of both players to punish each other in the event of prior defection. Notice this is in contrast to the case of the amicable equilibria, where mutual friends should be expected not to defect on each other or punish the other one often.

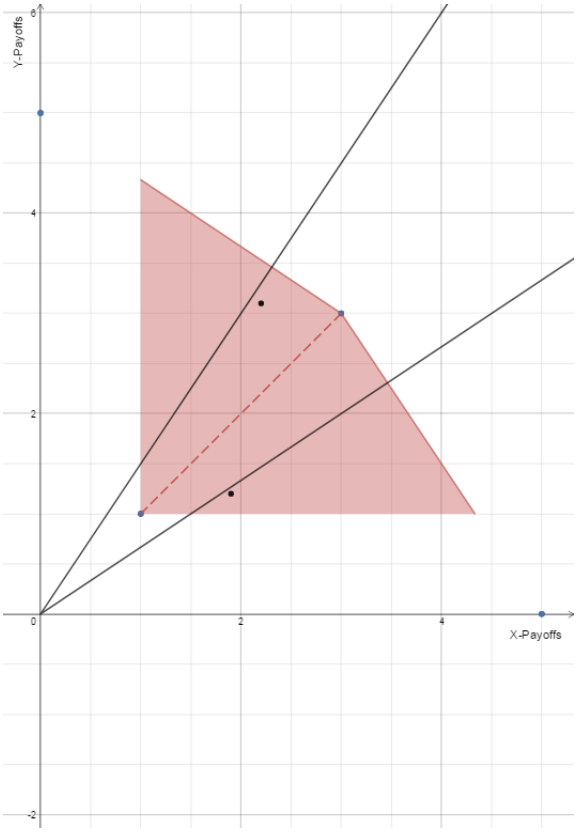

Figure 4.11: In this case we see the feasible region of the repeated Prisoner's Dilemma cut by two lines modeling an inequality tolerance of .25. I.e. $.25 = \frac{|U_X - U_Y|}{U_X + U_Y}$. Equilibria between these two lines could be said to approximate fair ouctomes for a given tolerance. Notice that as the payoffs increase, so does the tolerance for absolute inequality.

We can now introduce a third equilibrium concept intended to model reci-

procal relationships: egalitarian equilibria. We base this on the previously seen expression modeling tolerance for inequality $T = \frac{|U_A - U_B|}{U_A + U_B}$. A higher tolerance can allow for more unequal outcomes. In the case of the standard games, the tolerance is 1. If we only wanted to allow symmetric strategy profiles, we would have a tolerance of zero. This gives us a way to set a limit on how tolerant agents would be of divergence from equality.

Definition 4. *For a given inequality tolerance T, we say an outcome in a repeated game is a* ***egalitarian outcome*** *if the respective payoff profile (U_X, U_Y) satisfies $\frac{|U_X - U_Y|}{U_X + U_Y} \leq T$. An* ***egalitarian equilbrium*** *is an egalitarian outcome in the feasible region of repeated game payoffs subject to the folk theorem.*

This third equilibrium concept completes the correspondence between feasible game solutions and the three canonical relationship types: communality, dominance, and reciprocity. It gives a natural analogue for fairness where low-stakes competition leads agents to prioritize fairness as seen in Rabin [1993]. Moreover, it preserves the desired outcomes seen in *tit-for-tat* strategies on the basis of a simpler mechanism than those of Rabin [1993] or Fehr and Schmidt [1998]. Last, it follows the scheme of the claims in Mailath and Samuelson [2006] that argue that relationships are classes of equilibrium paths in the feasible region of a game.

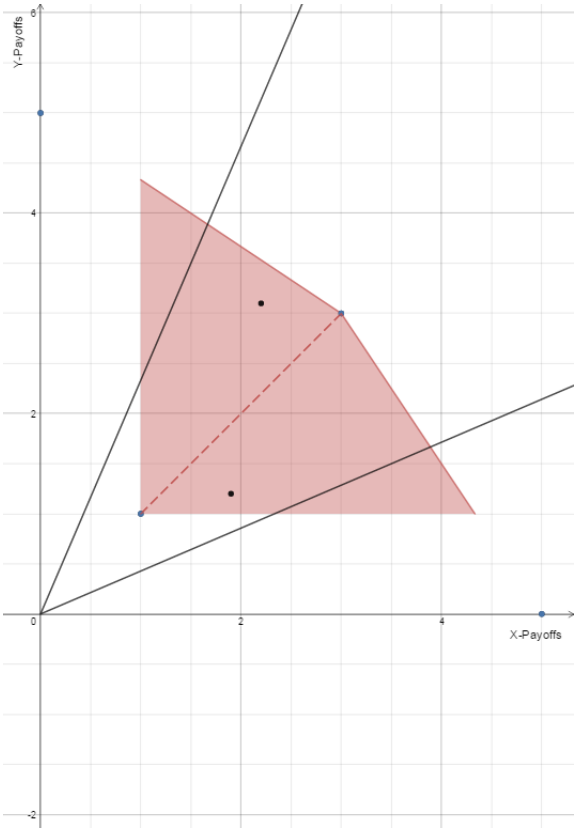

Figure 4.12: In this case we see the feasible region of the repeated Prisoner's Dilemma cut by two lines modeling an inequality tolerance of .40. I.e. $.40 = \frac{|U_X - U_Y|}{U_X + U_Y}$. The black dots are two egalitarian equilibria at $(2.2, 3.1), (1.9, 1.2)$.

It thus follows, in parallel to remarks in Pinker et al. [2008] that relationships are maintained or nullified according to common knowledge, that we have strict geometric boundaries on payoff profiles that can give us more suitable heuristics for seeing why reciprocal relationships adhere to both partners behaving similarly, even without sympathy. A question we hope to pursue is the exclusivity of despotic and egalitarian equilibria, as these regions appear to be geometrically distinct.

4.8 Conclusion

We have brought forth several considerations in mechanism design and preference alterations that make our modeling approach more realistic in cases reflecting psychological intuitions [Kahneman, 2003] [9]

We should highlight again the prevalence of sympathetic and spite motives in contrast to norms of cooperation and fairness. We should also highlight our introduction of amicable, despotic, and egalitarian equilibria, concepts that we hope will refine equilibria in the more complex trust games in the coming chapters according to the canonical relationship types of communality, dominance, and reciprocity seen in Fiske [1992]

[9] Need more references? Thaler? Kahnemann?

Bibliography

Nicholas Asher and Jason Quinley. Begging questions, their answers and basic cooperativity. In *New Frontiers in Artificial Intelligence*, pages 3–12. Springer, 2012.

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.

Samuel Bowles and Herbert Gintis. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press, Princeton, 2011.

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

Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, 2003.

Juan-Camilo Cárdenas and Elinor Ostrom. What do people bring into the game? how norms help overcome the tragedy of the commons. In *4th Toulouse Conference on Environment and Resource Economics 'Property Rights, Institutions and Management of Environmental and Natural Resources', Toulouse (France)*, 2001.

Ernst Fehr and Simon Gachter. Cooperation and punishment. *American Economic Review*, 90(4):980–994, 2000.

Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition and cooperation. CEPR Discussion Papers 1812, C.E.P.R. Discussion Papers, March 1998. URL `http://ideas.repec.org/p/cpr/ceprdp/1812.html`.

A.P. Fiske. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689, 1992.

Erving Goffman. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books, New York, 1967.

Maurice Grinberg, Evgenia Hristova, and Milena Borisova. Cooperation in prisoner's dilemma game: Influence of social relations. In *Proceedings of CogSci*, 2012.

John C Harsanyi and Reinhard Selten. A generalized nash solution for two-person bargaining games with incomplete information. *Management Science*, 18(5-part-2):80–106, 1972.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, pages 1449–1475, 2003.

Tilman Klumpp and Hugo Mialon. On hatred. *Emory Law and Economics Research Paper*, 15(11-120):11–183, 2012.

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

David A Morand. Language and power : an empirical analysis of linguistic strategies used in superior-subordinate communication. *Journal of Organizational Behavior*, 248(December 1997):235–248, 2000.

Steven Pinker, Martin A Nowak, and James J Lee. The logic of indirect speech. *Proceedings of the National Academy of Sciences*, 105(3):833–838, 2008.

Jason Quinley. Politeness and trust games. *Student Papers Session, Proceedings of ESSLLI*, 2011.

Jason Quinley and Christopher Ahern. Questions of trust. *ESSLLI 2012 Student Session*, page 132, 2012.

Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302, December 1993. URL `http://ideas.repec.org/a/aea/aecrev/v83y1993i5p1281-1302.html`.

D. Sally. On sympathy and games. *Journal of Economic Behavior & Organization*, 44(1):1–30, 2001.

D. Sally. What an ugly baby! *Rationality and society*, 14(1):78–108, 2002.

David Sally. A general theory of sympathy, mind-reading, and social inter- action, with an application to the prisoners' dilemma. *Social Science Information*, 39 (4):567–634, 2000.

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

Richard H Thaler. From homo economicus to homo sapiens. *The Journal of Economic Perspectives*, pages 133–141, 2000.

Paul J Zak. The neurobiology of trust. *Scientific American*, 298(6):88–95, 2008.

Paul J Zak. The physiology of moral sentiments. *Journal of Economic Behavior & Organization*, 77(1):53–65, 2011.

Paul J Zak, Karla Borja, William T Matzner, and Robert Kurzban. The neuroeconomics of distrust: sex differences in behavior and physiology. *American Economic Review*, pages 360–363, 2005.