Contents

Chapter 9

Requests and Variants of Trust

> *"The central fact of economics is scarcity. There is never enough of anything to satisfy all those who want it. The first lesson of politics is to disregard the first lesson of economics." T̃homas Sowell [Sowell, 2001]*

Human interaction is built on the problem of scarcity. To resolve this problem, we have a familiar strategy: asking for help. This means engaging in the costly enterprise of trusting others for the chance at a better outcome. This led us to trust games. In the last two chapters we went through general strategies for understanding the repeated versions of trust games in their asymmetric and symmetric variants. The motivation was manifold. First, we constructed the trust game as a coarse model of the circumstances surrounding a speech act like a request. This involved crucially:

- Differing roles between speaker(sender) and hearer(receiver);
- A scarcity in the speaker's ability to perform a task that would benefit him;
- A perceived benefit to the hearer for helping in a one-shot scenario;
- An opportunity cost and opt-out condition for the speaker ;
- A perception of mutual reward should the receiver help the sender.

Second, we constructed conditions that we might see in a population. These were the symmetric and asymmetric versions. The symmetric version is a two-population model where each agent has the four options of AH, $\neg AH$, $A\neg H$, and $\neg A\neg H$. The asymmetric version has senders with the options of A or $\neg A$ and receivers with the option of H or $\neg H$. The rationale for this divide is that in some scenarios members of a population can occupy several roles, whereas in others the roles may be more static.

The results we found gave us rationales for constructing symmetric equilibria promoting the *friendly AH* action as message cost increased. The threshold for maintaining this cooperative outcome decreased under strategies that involved punishment, something consistent with the idea of face loss/ face threat in the politeness literature [Brown and Levinson, 1978, Spencer-Oatey, 2008] and credible threats in adversarial games [Myerson, 1997]. This is because long-run players

would want to maintain their reputation and access to its subsequent benefits. We also found symmetric equilibria promoting the *insistent* $A\neg H$ action as message cost decreased. This is important as a countervailing force creating incentives for not helping is what gives politeness its intrigue. In other words, we would not want it the case that so-called polite speech acts and cooperative responses to them would always be in equilbrium, as this would conflict with the speech acts detailed in Figure 9.2.

We now proceed further into the modeling of requests, proposals, and other forms of discourse. We want cases where we can account for the desired outcomes with and without repetition and where initial circumstances model more closely the kind of expected benefits speakers might encounter. We also want a deeper look at relationships through external circumstances like repetition and internal drivers of behavior like preferences.

This model in Figure 9.1 will serve as a baseline for the further investigations in this chapter and the next. We recognize its limitations, and we hope to address that in the coming sections. We also hope to use it in distinguishing between models of cooperative dilemmas that can give rise to norms and the subsequent coordination problems that can give rise to conventions.

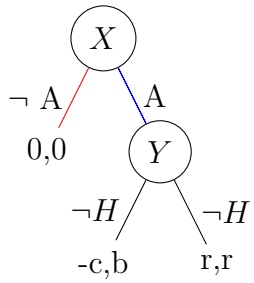

Figure 9.1: Generalized Trust Game

In this chapter we will extend our treatment of trust games to speech acts like requests and varying modifications to their payoffs. In general, our aim is to model various underlying payoffs surrounding requests and investigate how mechanisms like internal preferences and external circumstances affect the stability of the norms of helpfulness and gratitude. In the coming chapters, we will consider a signaling model portraying the various speaker and hearer payoffs under variations of directness, cost, and face-work. We will further give a treatment of proposals, contrasting with requests in the underlying incentive structures and resulting variation in the usage of modal verbs. I.e. we derive conditions for why using a modal verb in a proposal is less felicitous than in a request. We will explore more deeply the dynamics of requests and implicatures under other conditions of asymmetry. When examining the role of the receiver, we will see choice structures revealing a cooperative, or altruistic, method of deception. Here we begin with more details of modeling requests with trust games before moving on to extensions of the model, equilibrium refinements, and comparisons to proposals.

9.1 Modeling Request with Trust Games

Returning to our basic model of the trust game, we have a scenario where a speaker, or sender, has an intention to perform a task with which he would like help. On a very basic and coarse level, he has a choice about whether to ask another person and if so, how much effort to put into the request. This single-agent decision problem we see in the diagram of speech act choice given in [Brown and Levinson, 1987]. Under the branches highlighted in blue, we see that we have a wide variety of speech acts possible to perform. This we see in condensed form in the trust game simply along the dimension of *Ask*. With those varieties in mind, we then have a further action that the hearer could take upon assessing the urgency of the request, his own preferences, and the relationship between himself and the speaker.

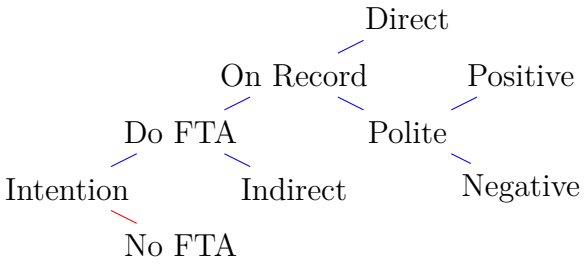

Figure 9.2: Strategic space of speech acts from [Brown and Levinson, 1987]

Notice that this decision tree captures only the choices facing the speaker when making the utterance. To resolve this, works like [Quinley, 2011, Quinley and Ahern, 2012] give a coarser view of the overall dynamics in requests by incorporating trust games, extensions of which we saw in the last chapters. We will adapt these approaches in the coming sections. In the chapters on trust games, we saw that more costly utterances could lead to the stability of the so-called *friendly* outcomes. But just how do these costs play out? We hope to address that with a basic signaling model.

Brown and Levinson's typology of speech acts involved in politeness (Figure 9.2)fits in nicely with the trust game seen in Figure 9.1. What we need to do now is

- determine the speech acts captured by the dynamics of the game
- expand on the strategies available to the hearer of the speech act
- explore alternate payoff structures, such as those seen in Figure 9.3
- integrate the decision tree of Figure 9.2 with trust games like Figure 9.1 and Figure 9.3.

Requests and Fragile Trust Games

We first note that we could also construct a different variant of the trust game to model variations of requests. Consider cases where there is a cost to the

receiver for helping, $-h$, rather than a reward r, as seen in Figure 9.3. This could even be more plausible than the original trust game and potentially present a greater discount threshold to overcome in the repeated instances. We construct the symmetric game for this purpose in Table 9.1.

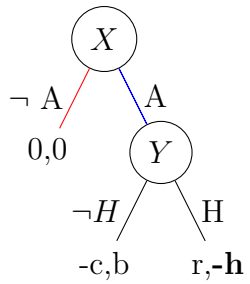

Figure 9.3: Fragile Trust Game: Backward induction gives us a Nash Equilibrium of $(\neg A, \neg H)$ as expected.

Given this more plausible model of interaction in the fragile game, we will apply the notions and results from the previous chapters to generalizing the model for a population, repeating the interaction within a relationship, and examining the preferences interacting with sympathetic payoffs.

Beginning with this basic model in the asymmetric, one-shot case with standard utility functions, we see a familiar result. If we use backward induction on the game in Figure 9.3, we see that it is irrational for Y to help, and therefore X would pay the cost of asking $-c$. With this in mind, it behooves X to not ask and thereby keep payoffs at $(0, 0)$. One way to get around this is to symmetrize the game, as seen in Table 9.1. Looking at this game, we see the asymmetric case's corresponding Nash Equilibrium. I.e. the action $\neg A \neg H$ is a best response to itself and thus constitutes the symmetric Nash Equilibrium in the case of a one-shot scenario.

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

Table 9.1: Symmetric Population Fragile Trust Game Formalized: Nash Equilibrium of $(\neg A \neg H, \neg A \neg H)$ is also the mutual minmax profile as it yields the lowest utility of the mutual best responses. Given the parameters, the two underlined outcomes could both feature in the repeated game.

We can contrast the fragile model to the Symmetric Trust Game from our previous chapter where we symmetrized the game in Figure 9.1, as seen in Table 9.1. We can see from the start several results. The most striking of which is that among the two potentially socially beneficial outcomes (AH, AH) and

	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 9.2: Symmetric Population Trust Game Formalized: Nash Equilibrium of $(\neg A\neg H, \neg A\neg H)$.

$(A\neg H, A\neg H)$ in Table 9.1, it is not apparent which one is obviously better. Under a high cost to helping h, we might see the true fragility of making the request. In other words, there could be a case where the friendly strategy AH could be a Nash Equilibrium in a repeated game but not necessarily an evolutionarily stable strategy, as a population playing strictly $A\neg H$ against itself would outcompete a population playing strictly AH. We also see that we will require a steeper discount factor to sustain friendly outcomes than before. Taking the Imitator and Punishment dynamics from before against naïve strategies, we have the following constraints on the discount levels required to sustain the symmetric outcome of (AH, AH) in Table 9.1:

$$\delta_{Imi} > \frac{b+h}{r+c}$$
$$\delta_{Pun} > \frac{b+h}{r+b+c}$$

These constraints on the discounts differ in the numerator: $b+h$ vs. $b-r$, the symmetric reward seen in the last chapter. Observing that $b+h = b-(-h)$ we see that we are comparing the benefit of helping to that of not helping once more. As this numerator is larger$(b+h > b-r)$, we see that a more costly task for the receiver should require a relationship highly invested in future interactions. We should make a further observation that in all cases we have posited that $b > r$. Thus in the case of the imitation dynamics we need that $\frac{b+h}{r+c} < 1$ or $b+h < r+c$ for the discount to be meaningful. It may not be the case that this is always true. As we know $b > r$ and that $h > c$ is possible, depending on how much X would be at a loss without help, it might be the case that this will never happen. Thus we see that the imitation dynamics may never be enough to sustain the cooperative moves in the fragile version of the Symmetric Trust Game seen in Table 9.1.

In the case of punishment, we need that $\frac{b+h}{r+b+c} < 1$ or that $b+h < r+b+c$. This equates to $h < r+c$. We can observe that this condition can be more easily satisfied. Notice that this also gives the speaker a good metric on how costly the speech act could be based on his belief of the cost to the hearer for helping. That said, if we have that $h > r$, we will lose (AH, AH) as a potential outcome, as it will not outperform a payoff of $(0, 0)$. [1]

When considering the figure in **??**, we see first that the asymmetric outcomes

[1] E.g. *I don't know HOW I can repay you.*

fall outside of the region of socially feasible equilibria. We next can see that the outcome of (AH, AH), which gives a payoff of $(1, 1)$, is outperformed by $(A\neg H, A\neg H)$, which gives a payoff of $(2, 2)$. Note that this is an inversion of the pattern from the Symmetric Trust Game and gives us a reason for the difficulty of sustaining the cooperative outcome without punishment. I.e. the pull of the non-cooperative action $A\neg H$ is too strong. We claim that it should be easier(in the sense of less patience required) to find repeated game strategies that promote the symmetric $(A\neg H, A\neg H)$ outcome, although we will not perform that calculation here.

Let us now compare the feasible regions and symmetric payoffs in the two games from Figure 9.4. If we look at the feasible region from the fragile game in **??**, we see that the symmetric outcome that lies furthest from the origin along the line of highest group benefit the insistent outcome of $(A\neg H, A\neg H)$. We then see that the outcome in the middles of the region is the friendly outcome (AH, AH). We note this as the Nash equilibria must outperform the min-max profile, it could be the case that a high cost of hellping would make this case equivalent in perfomance or worse than $(\neg A\neg H, \neg A\neg H)$, and thus not an equilibrium. As such, we have a very pessimistic outlook for this game. In order to remedy this, we may have to appeal to other mechanisms to promote the desired outcomes. In contrast, our region from Figure 9.4 with similar payoffs generates a much more beneficial situation. Here we find that the friendly outcome of $(AH, AH)$is the greatest in group welfare and that some of the asymmetric pure strategy outcomes, $A\neg H$ vs. AH, survive the elimination of points that do not outperform the mutual minmax.

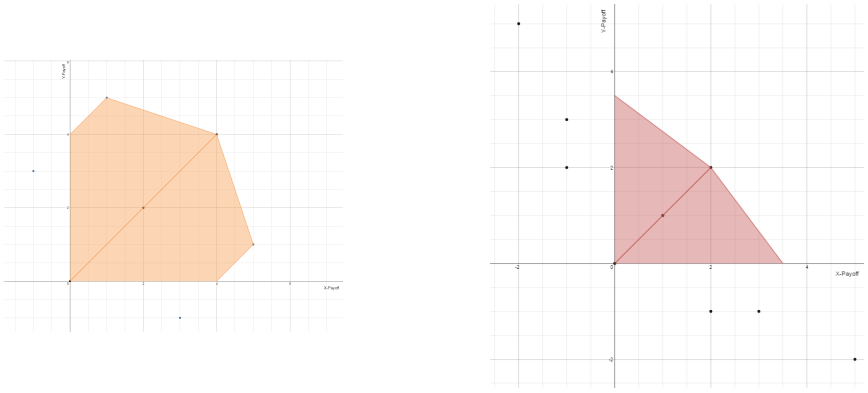

Figure 9.4: Feasible regions for canonical Symmetric Trust Game and fragile Symmetric Trust Game from the previous chapter with $r = 2$, $b = 3$, $c = 1$, $h = 1$.

Asymmetric Knowledge and Asymmetric Payoffs

It may be the case that the payoffs differ radically based on prior information, and we want to reflect that in this work. For instance, if the speaker has imperfect information about the actual state of the world then we would have a different game. In this scenario, we can imagine the case that a beggar asks a wealthy

passer-by for money. If it is not apparent as to whether the passer-by is wealthy, the speaker has imperfect information about the state of the world when asking him for money. This sets the stage for us discussing answers to requests as a matter of both ability and volition. It will form the basis of a later chapter, and there we will see further the impact of sympathy on this exchange:

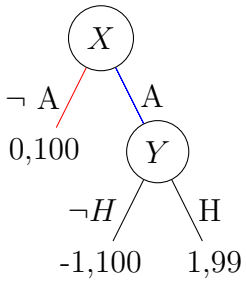

Figure 9.5: Beggar's Basic Trust Game

For the present moment, we can imagine a scenario of a passer-by with $100 in his pocket, of which the beggar asks for $1. The passer-by can help or not, and if not helped the beggar suffers an opportunity cost. We can also consider a symmetrized version of this game as before, with the table in Table 9.3, where we imagine a population of potential inequalities.

	AH	$\neg AH$	$A\neg H$	$\neg A\neg H$
AH	100,100	1,99	98, 101	-1, 100
$\neg AH$	99,1	100,100	99,1	0,0
$A\neg H$	101,98	1,99	99,99	-1, 100
$\neg A\neg H$	100,-1	0,0	100,-1	100,100

Table 9.3: Symmetric Beggar's Trust Game: Nash Equilibrium of $(\neg A\neg H, \neg A\neg H)$.

Despite the success of the previous chapter for showing the benefit of tit-for-tat style imitation and grim trigger style punishment, we should remark that a beggar has little in the way of power over his addressee. Thus in the coming chapters we will have to appeal to sympathy to account for this brand of cooperation.

9.2 Requests with Gratitude: Gratitude Isn't Gratis

Within a discourse, we have multiple moves that can take place [Asher and Lascarides, 1998, 2001]. For instance, depending on one's role in the relationship and the discourse, there can be asymmetries of action space and information space. As we mentioned *uncorrelated asymmetry* earlier, we could have a case where the two interactants know their separate roles but have an otherwise symmetric relationship. This is most likely to happen in a dominance relationship. Notice

however that this also fits in with turn-taking in a conversation. Only one speaker can hold the conversational floor at a time, and yet both will, over the course of a conversation, have some element of setting the conversation's agenda. This asymmetry folds the complex nature of speech acts into a simplified model of message cost for the sender, which we highlight here.

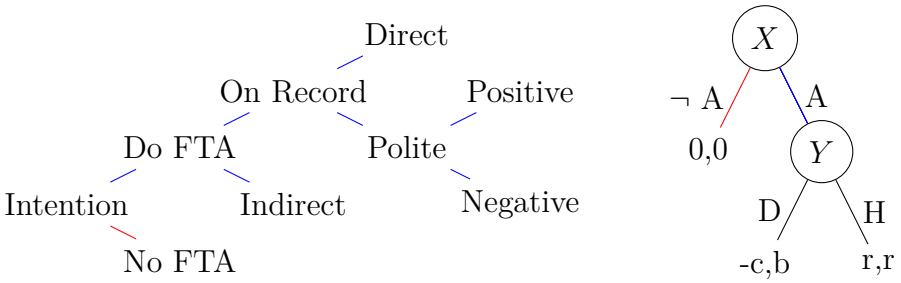

Figure 9.6: Note that trust games on the surface lose the complexity of speaker choice for the benefit of hearer choice and utility functions.

Extending Trust Games

In an individual speech act, a natural asymmetry comes into play. Speakers have different roles, and the model of a requester who can thank his helper is presented below. We thus connect our work from last chapter and some of the concepts from Quinley and Ahern [2012]. The idea is that we want to model a more complicated discourse resembling the *Centipede* games seen before. In this case, a sender plays the trust game with the receiver, but the sender can also reply in thanks. In effect we have a nested trust game similar to the fragile trust game seen in Figure 9.3. On a basic level, we can extend the fragile trust game and have the interaction seen in Figure 9.7.

If we perform backward induction on the game in Figure 9.7 as before, we get the classic result. Thanking is more costly than not doing so, therefore helping is more costly than not doing so, and therefore asking is more costly than not doing so. This again gives us that we should expect no requests in one-shot scenarios. In contrast, we might rather expect the tree to play out according to a dialogue like the following:

> **Xavier:** I hate to bother, but could you help me with this paperwork?
> **Yves:** Sure. I filled it out last week.
> **Xavier:** Thanks! I owe you one!

Here we can make two observations (or claims). The first is that we can see sequential reciprocity playing out in the dialog. The second is that we also have a marker of future reciprocity in the thanks. Assuming a future discounting and/or a case of indefinite repetition, one claim might be that it is not so important as to guarantee future reciprocity as it is to publicly claim it to preserve the relationship. We say this as an agent's future could be uncertain and that *a little courtesy goes a long way*, or as said in the eponymous *A Little Thanks Goes a*

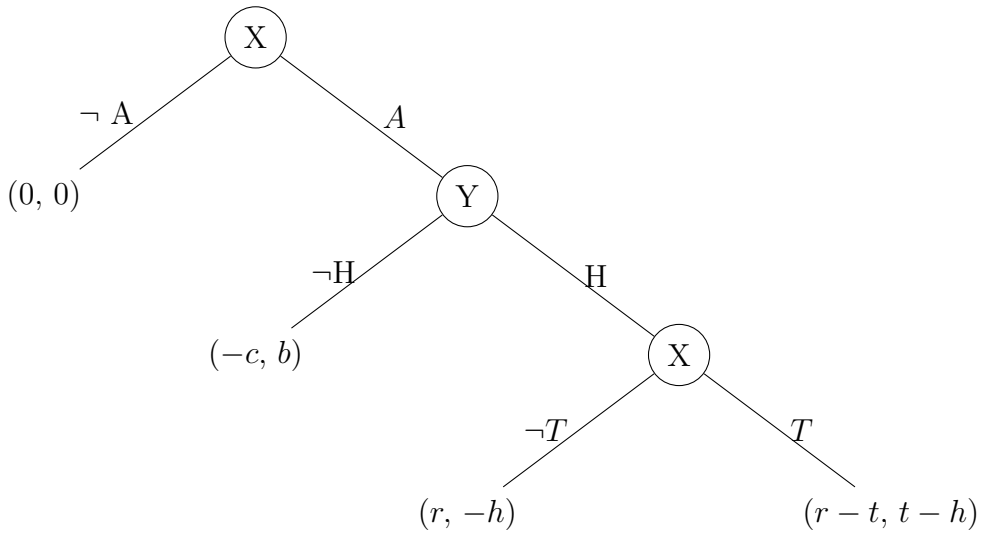

Figure 9.7: Extended Fragile Trust Game: X can choose to Ask (A) Y, who can choose to Help (H). X can choose to Thank (T).

Long Way Grant and Gino [2010]. We can model this by saying that the mere act of thanks in the present has a utility value equivalent to the expected value of some future benefit in a far-off horizon, as it is now common knowledge of the indebtedness of the recipient. E.g. it might be the case that a sufficiently impatient player would ignore the possibility of a future return on his favor if his interlocutor said thanks.

We see in English other forms of reciprocity indicated in formulaic conventions of gratitude. Notice that several of these both give implications of a measured cost being assessed (e.g. appreciate) and make common knowledge a future debt (e.g. obliged) to be repaid through a repeated interaction. These include:

· *Much obliged.*

· *I owe you one*

· *I appreciate that.*

· *What can I do to repay you?*

· *I won't forget this.*

The question now is whether we can construct a model predicting the appearance of these in natural language consistent with the data. We now proceed to analyze constraints for eliciting certain equilibria. There are a few options here:

· Consider a repeated, symmetrized game in a population as done in Ch.7-8,

· Explore the repeated asymmmetric version of the game,

· Introduce sympathy payoffs and relational preferences as done in Ch.6,

· Introduce more complicated payoff system including face payment, seen in section 9.3

We will proceed by exploring these options first independently and then combining them. One of our goals is to see the impact of sympathy on patience required to sustain the cooperative outcomes. Another goal is to investigate the geometry of the feasible region seen in the folk theorem's predictions to validate our analytical results. In all of these games, we have actions that are generally seen as *cooperative* in the sense that they give us options that are more beneficial to the group. We will see what parameters of sympathy and patience allow for these outcomes.

In these coming sections, we will follow a general trend of investigation. This will first allow us to investigate the impact of sympathy on what players would do in the one-shot game. Next, we will examine the impact sympathy has on the patience required to sustain various outcomes in the repeated game with its normalized utility. We should note that in cases of zero sympathy, we have the standard utility function. I.e.

$$V_i(0) = U_i$$

This means that we can verify our results by setting $s = 0$ in the cases without sympathy. This also means that we will omit some of the calculations, to be considered in an appendix. We now proceed to the symmetric model of the fragile trust game extended.

Fragile Trust Games: Extended, Symmetrized, Repeated

	ATH	$A\neg TH$	$\neg AH$	$AT\neg H$	$A\neg T\neg H$	$\neg A\neg H$
ATH	r-h,r-h	r-h-t,r+t-h	r-t,t-h	t-c-h,r-t+b	-c-h,r+b	-c,b
$A\neg TH$	r+t-h,r-h-t	r-h,r-h	r,-h	t-c-h,r-t+b	-c-h,r+b	-c,b
$\neg AH$	t-h,r-t	-h,r	0,0	t-h,r-t	-h,r	0,0
$AT\neg H$	r-t+b,t-c-h	r-t+b,t-c-h	r-t,t-h	b-c,b-c	b-c,b-c	-c,b
$A\neg T\neg H$	r+b,-c-h	r+b,-c-h	r,-h	b-c,b-c	b-c,b-c	-c,b
$\neg A\neg H$	b,-c	b,-c	0,0	b,-c	b,-c	0,0

Table 9.4: Symmetric Extended Fragile Trust Game : X(Row) and Y(Col) play two Extended Trust Games simultaneously with each other. Notice that a player who does not play A will also not have the choice of $T \vee \neg T$.

Let us consider the Extended Fragile Symmetric Trust Game as seen in Table 9.4. Once again, we have the Nash Equilibrium of $(\neg A\neg H\ ,\neg A\neg H)$, so now we ask how we can make the various symmetric population outcomes into equilibria by making the actions best responses to themselves in a repeated game. A consideration here is what strategy subsets should be promoted as norms. For instance, we claim that not thanking after helping should be a grosser violation

of norms than not helping after being asked [Bartlett and DeSteno, 2006, Tsang, 2006, Gintis, 2005, 2000] for the reason that it is essentially a low-cost version of reciprocity.

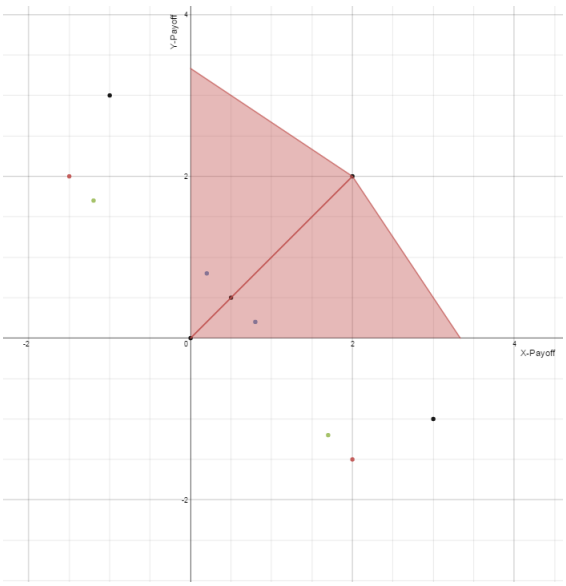

Figure 9.8: Feasible region for the extended fragile Symmetric Trust Game with $r = 2$, $b = 3$, $c = 1$, $h = 1$, $t = .5$. Outcomes including thanking are in black. For $t = 0$, this is the fragile Symmetric Trust Game.

We can return to our use of the folk theorem and consider the figure in Figure 9.8. What we notice when considering the outcomes with thanking is that they often fall on the same line as their related actions. We also notice two new outcomes in the interior region, generated by the actions ATH and $A\neg TH$. This means that these actions are justifiable in the repeated instances of the game. We will now consider the various dynamics of imitation and punishment, combined with elements of strategic restriction towards analyzing the game.

Imitation and Punishment Dynamics: ATH vs.$A\neg T\neg H$

We begin with the imitation dynamics pitting ATH against $AT\neg H$. Notice that we need to respect the inequality of $b > r > h > c > t$. Notice also that as the roles are symmetric, we will not tag the discount values with an agent's label. E.g. we will write δ and not δ_X.

$$U(ATH) > U(A\neg T\neg H) \Rightarrow r - h > (1-\delta)(r + b + \delta \sum_{j=0}^{\infty}(b-c)\delta^j) \Rightarrow \delta > \frac{h+b}{r+c}$$

This will never happen as $h + b > r + c$ based on our inequality above. This means we need an alternate mechanism for promoting cooperation. Consider now the punishment dynamics, which would promote a player towards punishing an unhelpful and thankless partner:

$$r - h > (1 - \delta)(r + b + \delta \sum_{j=0}^{\infty}(-c)\delta^j) \Rightarrow \delta > \frac{h + b}{r + b + c}$$

As we have the larger denominator, this quantity will be potentially reachable for patient players as we already have that $r + c > h$, and thus the discount will never be greater than 1. Nonetheless, the high reward for breaking both norms shows us that even under a punishment dynamics, the players will have to be very patient.

Imitation and Punishment Dynamics: ATH vs.$A\neg TH$

We can observe again by inspection of the table above that the imitation dynamics will never promote ATH vs.$A\neg TH$. We now turn to the punishment dynamics. In this case we choose the punishing action of $\neg A\neg H$, although $A\neg T\neg H$ would yield a worse outcome for this strategy.

$$r - h > (1 - \delta)(r + t - h + \delta \sum_{j=0}^{\infty}(-c)\delta^j) \Rightarrow \delta > \frac{t}{r - h + t + c}$$

We can see in this inequality, as t is a very cheap signal, that we don't require very patient players to sustain this norm. We can also compare this to the case above, where we have agents following one norm, but not another in this case, and thus the level of patience required is much less to sustain the cooperative outcome.

Imitation and Punishment Dynamics: ATH vs.$AT\neg H$

We now turn to the case of the other violation of norms: choosing not to help. If we compare ATH vs.$AT\neg H$ under the imitation dynamics, we see that we get a similar case to not helping before:

$$r - h > (1 - \delta)(r - t + b + \delta \sum_{j=0}^{\infty}(b - c)\delta^j) \Rightarrow \delta > \frac{h + b - t}{r + c - t}$$

Because of the relative weights of the payoffs, this will not give us the cooperative outcome for any value of δ. This leads us to the punishment dynamics:

$$r - h > (1 - \delta)(r - t + b + \delta \sum_{j=0}^{\infty}(-c)\delta^j) \Rightarrow \delta > \frac{h + b - t}{r + b + c - t}$$

This does sustain the cooperative outcome of ATH. Notice again this is a potentially higher discount value than the one used for maintaining thankfulness alone. What we have shown here is that the norm of thankfulness is *easier* to

sustain through repeated interaction, and thus it should come more as a surprise when someone *does not thank* another for a favor than when someone *does not grant* a favor. We have also shown that **only** the punishment dynamics promote the cooperative outcomes in our case.

Strategic Restrictions and Equilibria

We will now perform strategic restrictions in the manner done before to find subsets strategically equivalent to other known games. For instance, if we restricted the game to always helping and always asking, we would have a zero-sum scenario. The question is what strategies we should eliminate. E.g. should we eliminate not asking? Not thanking? In the following sections, we will consider eliminations that lead to new equilibria or that give us productive counterexamples as to why some equilibria persist despite mechanisms that might go against them.

Eliminate Not Asking

	ATH	$A\neg TH$	$AT\neg H$	$A\neg T\neg H$
ATH	r-h,r-h	r-h-t,r+t-h	t-c-h,r-t+b	-c-h,r+b
$A\neg TH$	r+t-h,r-h-t	r-h,r-h	t-c-h,r-t+b	-c-h,r+b
$AT\neg H$	r-t+b,t-c-h	r-t+b,t-c-h	b-c,b-c	b-c,b-c
$A\neg T\neg H$	r+b,-c-h	r+b,-c-h	b-c,b-c	b-c,b-c

Table 9.5: Symmetric Extended Fragile Trust Game : Eliminate Not Asking. Note that the action $A\neg T\neg H$ strictly dominates its counterparts. Notice as well that $r - h < b - c$ for all permutations of their values.

If we consider the game seen in Table 9.5, we see that we have several Nash equilibria, all of which involve the $\neg H$ action. Also notice that this the desired state of (ATH, ATH) is the worst symmetric state in terms of group benefit $b > r > h > c > t \Rightarrow r - h < b - c$. Have we therefore removed the most credible threat of punishment from the players? How do the punishment and imitator dynamics fare here for promoting outcomes other than the equilibria in the one-shot game? Imitation and punishment both prove useless. To see why, recall that the equilibria in the repeated game should fare better than the mutual minmax, in this case the equilibria that yield $b - c$ as a payoff to both players. Now consider the feasible outcomes that could be so: there are none, as seen in Figure 9.9. Thus we should not expect the stability of the friendly outcomes.

Eliminate Not Helping

Consider what would happen in eternally helpful society. There would be no credible threat of not helping the thankless. This is what we see in the table here inTable 9.6. In this case the strategy $A\neg TH$ dominates the others, and the notion of thanks becomes a wasted cost. Does

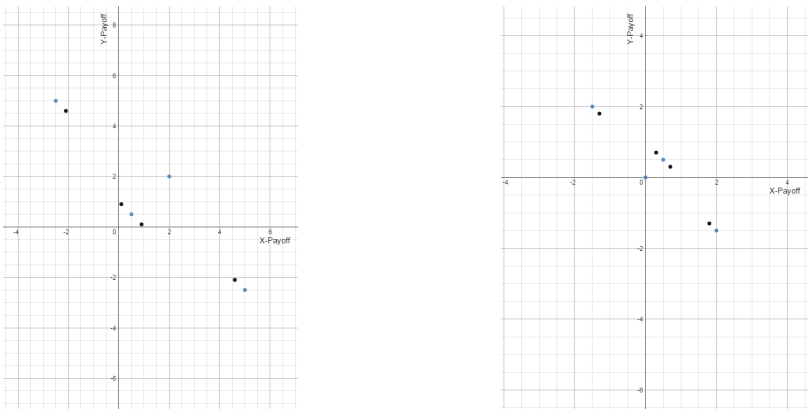

Figure 9.9: Symmetric Extended Fragile Trust Game : Eliminate Not Asking(Left) vs. Not Helping(Right). There are no feasible equilibria, as $b - c$ and $r - h$ are both the mutual minmax payoffs and the most group-beneficial joint payoffs.

	ATH	$A\neg TH$	$\neg AH$
ATH	r-h,r-h	r-h-t,r+t-h	r-t,t-h
$A\neg TH$	r+t-h,r-h-t	r-h,r-h	r,-h
$\neg AH$	t-h,r-t	-h,r	0,0

Table 9.6: Symmetric Extended Fragile Trust Game :Not Helping Eliminated. Nash Equilibrium of ($A\neg TH$, $A\neg TH$))

The Nash Equilibrium of ($A\neg TH$, $A\neg TH$) seen in Table 9.6 gives us the mutual minmax payoff of $r - h$ for both players. It is for this reason that we will not see any new equilibria in the repeated game predicted by the graph in Figure 9.9.

Eliminate Not Thanking

We can see in this case a subset of the strategies strategically equivalent to the Prisoner's Dilemma when we compare $AT\neg H$ and $\neg A\neg H$. This should give us a Nash Equilibrium of ($\neg A\neg H$, $\neg A\neg H$). On the other hand, there is a non-dilemma when examining the upper left quadrant of ATH and $\neg AH$. We can further separate these two scenarios by eliminating further actions.

It now is the case that we have two pure strategy outcomes in the feasible region, seen in **??**, the symmetric cases ATH and $AT\neg H$ against themselves. We now consider the prospective discounts provided by the imitation and punishment dynamics that would promote the friendly outcome of ATH. As ($AT\neg H$, $AT\neg H$) is along the farthest frontier of group welfare, this may be the reason that the imitator dynamics fails; imitation gives us the highest paying outcome. Now for the punishment dynamics:

	ATH	$\neg AH$	$AT\neg H$	$\neg A\neg H$
ATH	r-h,r-h	r-t,t-h	t-c-h,r-t+b	-c,b
$\neg AH$	t-h,r-t	0,0	t-h,r-t	0,0
$AT\neg H$	r-t+b,t-c-h	r-t,t-h	b-c,b-c	-c,b
$\neg A\neg H$	b,-c	0,0	b,-c	$\boxed{0,0}$

Table 9.7: Symmetric Extended Fragile Trust Game: Not Thanking Eliminated. We see the familiar Nash Equilibrium of ($\neg A\neg H$, $\neg A\neg H$).

$$r - h > (1-\delta)(r - t + b + \delta \sum_{j=0}^{\infty} -c\delta^j) \Rightarrow \delta_{PUN} > \frac{h+b-t}{b+r+c-t}$$

We now consider whether this is plausible. Under the values in the diagram from **??**, we have that $\delta_{PUN} > \frac{h+b-t}{b+r+c-t} \approx 0.74$. Thus for very patient players, we can achieve the friendly outcomes.

	ATH	$\neg AH$
ATH	r-h,r-h	r-t,t-h
$\neg AH$	t-h,r-t	0,0

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

Table 9.8: Symmetric Extended Fragile Trust Game :Multiple Eliminations

If we further eliminated $\neg H$ or H, we would have the non-dilemma posed by ATH vs. $\neg AH$ or the Prisoner's Dilemma. In the first case, we could also have a Nash Equilibrium of (ATH, ATH) for the right choices, as seen in Table 9.8.

Unified Trust Game

A similar game tree can be found in Quinley and Ahern [2012]. To unify that account with our notation, this game tree now more closely resembles the notation seen in previous chapters and the payoff structure found in Quinley and Ahern [2012]. The issue now is to consider its variations in symmetrization, sympathy, and repetition.

Quinley and Ahern [2012] approached this with a basic notion of reputation which we advance here in tandem with the work done in the last chapter. If we symmetrize the game, we get a 6×6 matrix (Table 9.9) :

Here we make the claim that among the payoffs and costs, the ones consisting purely of face costs and not of action should have the lowest value. Among these, we often see that a formula for a request like *I hate to bother you but could you do this?* should be more costly than a sincere thanks like *I really appreciate it*, not only because of length or production costs but also because of potential face loss. Thus we have $c > t$.

In terms of actions, we also remark that the cost of doing the action on one's own $-a$ should be greater than the cost to the one helping $-h$, so $a > h$. Last

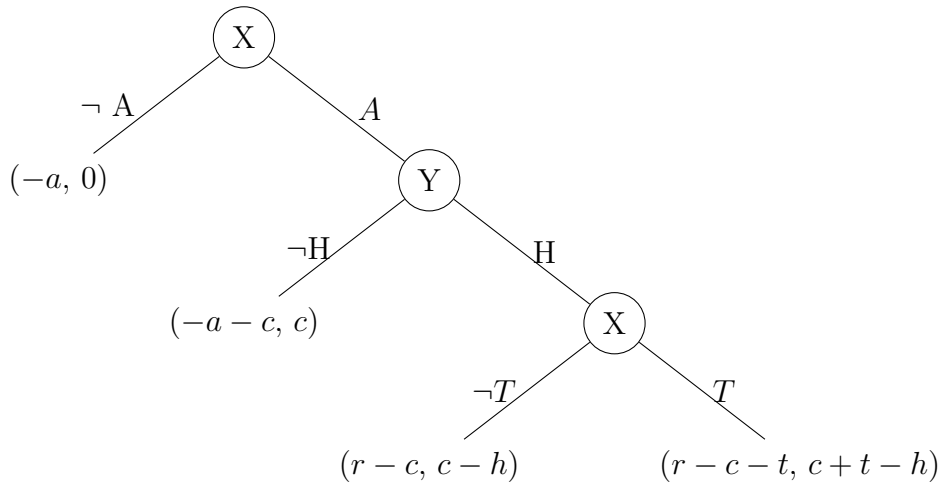

Figure 9.10: Extended Unified Trust Game: Player X can choose to Ask (A) something from Player Y, who can then choose to Grant (G) the favor. Player X can choose to Thank (T) or not Thank ($\neg T$) player Y.

	ATH	$A\neg TH$	$\neg AH$	$AT\neg H$	$A\neg T\neg H$	$\neg A\neg H$
ATH	r-h,r-h	r-h-t,r+t-h	r-c-t,c+t-a-h	t-a-h,r-t	-a-h,r	-a-c,c-a
$A\neg TH$	r+t-h,r-h-t	r-h, r-h	r-c,c-a-h	t-a-h,r-t	-a-h,r	-a-c, c-a
$\neg AH$	c+t-a-h, r-c-t	c-a-h,r-c	-a,-a	c+t-a-h,r-c-t	c-a-h,r-c	-a,-a
$AT\neg H$	r-t, t-a-h	r-t, t-a-h	r-c-t, c+t-a-h	-a,-a	-a,-a	-a -c, c-a
$A\neg T\neg H$	r,-a-h	r,-a-h	r-c,c-a-h	-a,-a	-a,-a	-a-c, c-a
$\neg A\neg H$	c-a,-a-c	c-a,-a-c	-a,-a	c-a, -a-c	c-a, -a-c	-a,-a

Table 9.9: Symmetric Extended Unified Fragile Trust Game : X(Row) and Y(Col) play two Extended Trust Games simultaneously with each other. Notice that a player who does not play A will also not have the choice of $T \vee \neg T$.

we claim that the reward for having the task done should also be greater than the cost of helping. Thus $r > h$, otherwise we might see no reason for multiple people to ask each other for help.[2] We thus posit the inequalities:

$$a \geq r > h > c > t$$

We can now consider the set of feasible equilibria in the repeated game. One thing we can notice is that the game itself yields lower payoffs in general to the players given the steep penalty a for not being able to get help. We also notice that several payoff points fall within the feasible region, and these are both symmetric and asymmetric strategy profiles. We also see that as the most socially beneficial state is (ATH, ATH), we should have a lower threshold of

[2]The technical reason is that this action profile is less efficient than the min-max profile and thus could not be an equilibrium in a repeated interaction.

patience to overcome to promote the cooperative outcomes, as compared to the fragile game. In addition, we can see the implications of reducing thanking to zero. As described in Figure 9.11, following the outcomes in black to their nearest blue neighbors would correspond to eliminating the thanking action. This would generate a region similar to the ones we have seen before, as in the Symmetric Trust Game. In that event, the only surviving pure strategy equilibrium in the repeated game would be (ATH, ATH), as the border outcomes do not strictly outperform the mutual minmax profile of $(\neg A\neg H, \neg A\neg H)$.

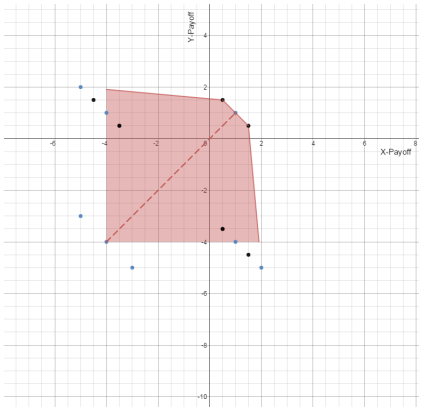

Figure 9.11: Feasible region for the unified Symmetric Trust Game with $a = 4$, $b = 3$, $c = 1$ $t = .5$. Outcomes including thanking are in black.

Imitation vs. Punishment

We could first examine the *Friendly Imitator* strategy as seen before to compare when the friendly strategy would outcompete a naïve player playing a selfish strategy of $A\neg T\neg H$. This would occur for a discount parameters predicted by the inequality

$$r - h > (1 - \delta)(r + \delta \sum_{j=0}^{\infty} -a\delta^j) \Rightarrow \delta > \frac{h}{r + a}$$

Observing these inequalities and the fact that we could claim the punishing strategy for all players is as before $\neg A\neg H$ as it always produces negative payoffs for plays against it, we could then attempt the *Friendly Punisher* against a selfish strategy of $A\neg T\neg H$. This would occur for a discount parameters predicted by the inequality

$$r - h > (1 - \delta)(r + \delta \sum_{j=0}^{\infty} (-a - c)\delta^j \Rightarrow \delta > \frac{h}{c + r + a}$$

We thus have that for a higher cost of helping, a lower cost of the initial action, or a lower reward, we will require more patient players for the cooperative outcome

to emerge. An interesting thing about these inequalities is the disappearance of the other factors seen in the previous analyses as the symmetry in the process gave us several cancellations. We can also observe that $r+a = r-(-a)$. Thus we can consider that the discount level required to maintain the friendly outcomes will be lower as the reward r outpaces the cost of performing the action on one's own $-a$. Last, we see that as we invoke the punishment dynamics, we require less patient players for cooperative outcomes to emerge. This limits our discussin of the unidfied trust game, for in the next section, we focus on adapting its structure to face payment and face threats.

Results

We saw in this section that the extended forms of the trust games, coupled with repetition and symmetrization, gave us chances at cooperation where there was none in the one-shot game. We also saw that the fragile trust game presents a unique set of challenges, as its nominally cooperative outcomes are outperformed by those violating norms. In terms of strategic elimination, we see a common theme: eliminate the credible threat of opting out in one stage of the game, and the other agent should choose to break the norm in the next. We saw this in both cases earlier. Eliminate $\neg A$ and it seems irrational to help H. Eliminate $\neg H$ and it seems irrational to thank T. Eliminating strategies from these games can either accelerate the chance of cooperation or eliminate it entirely, when strategies that are credible punishments disappear.

These models give us explanatory power for three phenomena:

- As eliminating $\neg H$ makes T unnecessary, we have reason for the formulaic responses of *de nada* in Spanish, *Mach nichts!* in German, or *No problem* in English as markers of positive politeness where there is low social distance. Does this also predict highly inflected formula for thanking when there is a low probability of helping?

- As outcomes involving thanking make the utilities closer to symmetric, we have a way of increasing the fairness of an exchange without repetition.

- As the parameters involving the cost of the request influenced the probability of helping more than the face payment of thanks, we have reasons for why a request should be a more elaborate speech act.

There can be a different motivator than material utility however. Agents being observed or with a chance of having their behavior reported may be concerned about face threats. We now address some of the work from Quinley and Ahern [2012] that inspired the strategically similar unified game. We then follow it up with mechanisms of repetition.

9.3 Trust, Face, and Context

We proceed to outline the model of requests found in Quinley and Ahern [2012]. This game models a sender who *can* perform a task but at a greater personal

cost than it would take the receiver. He can ask a receiver for help and later choose to thank him for the subsequent help. For each action we identify the following payoffs according to the sequential play, written $< S_X, S_Y, S_X >$ for the sequential play of strategies on each player's part. We see the game tree given in Figure 9.12, strategically similar to the one seen in Figure 9.10.

We now adapt developments seen in Quinley and Ahern [2012], Quinley [2011] towards incorporating face into the possible action sequences of a trust game. The action sequences follow, with explanations for the various utilities.

- $< \neg A_X >$: X pays a cost of performing the task himself; Y is unaffected .
  $U_X = -c_x$, $U_Y = 0$
- $< A_X, \neg H_Y >$: X pays Y a small face payment, but as Y does not help, X still pays a cost of performing the task himself.
  $U_X = -c_x - f_r$, $U_Y = f_r$
- $< A_X, H_Y, \neg T_X >$: X pays Y a small face payment and receives a benefit from Y's help, Y receives a face payment and performs the task at some nominal cost.
  $U_X = b_x - f_r$, $U_Y = f_r - c_y$
- $< A_X, H_Y, T_X >$: X pays Y a small face payment in the request and receives a benefit from Y's help; Y receives a face payment and performs the task at some nominal cost; X then pays Y face again through thanking Y.
  $U_X = b_x - f_r - f_t$, $U_Y = f_r + f_t - c_y$

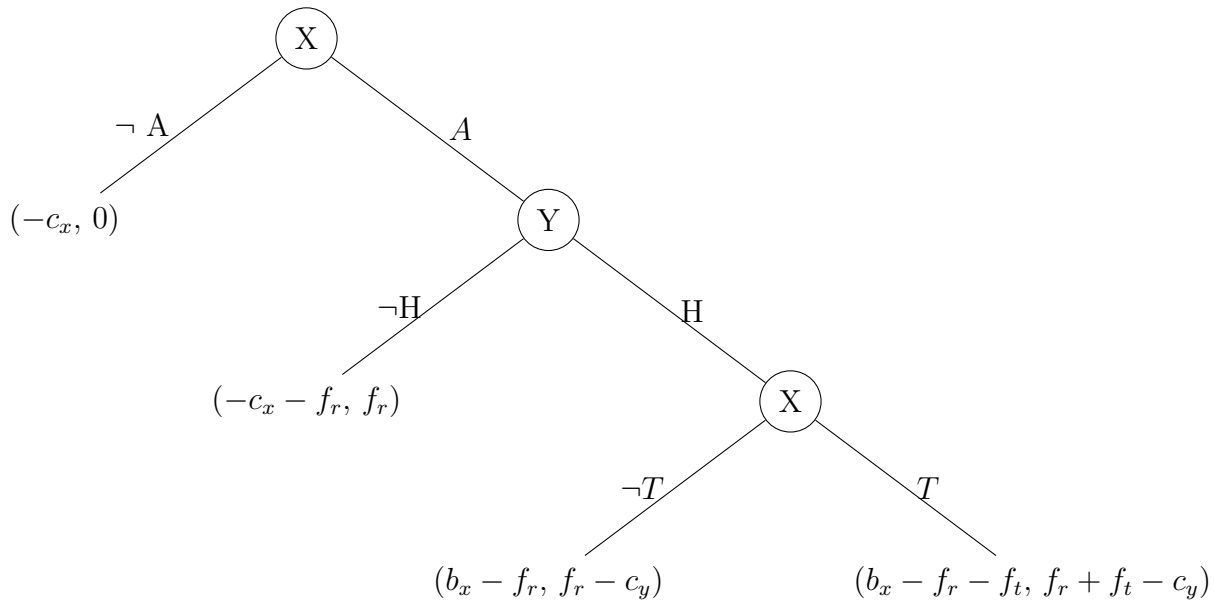

Figure 9.12: Unobserved Request Game with Face: Player X can choose to Ask (A) something from Player Y, who can then choose to Help(H). Player X can choose to Thank (T) or not Thank ($\neg T$) player Y.

Just as we have seen before via backward induction, this game tree shows that thanking is more costly, and thus X should not thank, therefore Y should not grant the favor, therefore X should not ask. This pessimistic outcome we could resolve through symmetrizing the game and investigating repeated outcomes with punishment strategies. We claim this would be sufficient. However, we also claim that another internal mechanism might push this through: face.

Asymmetric Repetition: Discounted Repetition Outperforming Alternatives

Since X acts last, there is still not enough incentive for him to play the cooperative move. However, if we repeat the interaction, there might be. We can discount the future interactions and compute sufficient conditions for satisfying the folk theorem. We should show that

- in some cases, there exists a discount value sufficient to promote cooperation for each agent past his next best alternative to the uncooperative action and
- for other cases, it may be impossible to promote the cooperative actions merely through repetition of the asymmetric game.

Let us first consider our previous cases: the basic trust game and the fragile trust game in Figure 9.17. Assuming trigger strategies, we can consider that if there exists a round at which Y would become unhelpful, then that should be equivalent to being unhelpful in the first round[Mailath and Samuelson, 2006]. In the basic case, we had that $\delta_Y > \frac{b-r}{b}$ promoted the cooperative outcome for Y, whereas the fragile game had no such result. There did not exist a discount value for Y to induce helping unless we symmetrized the game or introduced sympathy. As we assumed that the action to help should be more costly than the speech act, we had the case that *actions speak louder than words.*

We can now move forward to the game with face in Figure 9.12. With an agent Y that would punish X for not thanking, the claim is that since we are discounting on an infinite horizon, a defection at some point in the future is equivalent to considering defecting on the first round and then considering the utility for punishment in the subsequent rounds. Let us label the discount value for X as δ_X in the final step. We should have: In the more complicated case with face-payment seen in Figure 9.12, we have:

$$b_x - f_r - f_t > (1-\delta)(b_x - f_r + \delta \sum_{j=0}^{\infty}(-c_x - f_r)\delta^j) \Rightarrow \delta_X > \frac{f_t}{b_x + c_x}$$

This should give us that as the difference between the benefit and the cost increases, we should expect less patient players required to sustain thanking. This would also follow for cases where the cost of thanking goes down. In contrast, as the cost of thanking goes up, we would expect that X would have to be more patient with Y.

Might we expect a countervailing punishment for X of not asking for help? In some cases, yes, but in our setup above the burden is primarily on Y through the various costs required to help, which should still be more than verbal face payments. Let us consider his situation in the discounted folk theorem under punishment of $\neg A$ against a thankful type for X.

$$f_r + f_t - c_y > (1-\delta)(f_r + \delta \sum_{j=0}^{\infty} 0\delta^j) \Rightarrow \delta_Y > \frac{c_y - f_t}{f_r}$$

We see here that as the level of thankfulness and courtesy in the request go up, we should expect a more cooperative Y. This is only under the specious assumption that $c_y < f_t + f_r$, which may not be the case. It appears then that there is no reasonable incentive as of yet for Y to help, even in the repeated case.

But what about the case where there is no thanks? We claim that without this norm we might not see optimal behavior. Since we stipulated that $c_x > f_r$, it should be the case that the thankfulness constraint gives Y a stronger incentive to help. Otherwise we would require impossible levels of patience δ. To see why, consider the case of an thankless X:

$$f_r - c_y > (1-\delta)(f_r + \delta \sum_{j=0}^{\infty} 0\delta^j) \Rightarrow \delta_Y > \frac{c_y}{f_r}$$

Repetition with Reputation

We can also consider the case where an outside observer might witness the participants. One way to model this is a social context multiplier m, where uncooperative actions diminish the utility of the agent who does not act cooperatively. As asking for help is not a norm, we do not include it in the observation-induced penalties. In this case, consider the tree in Figure 9.13.

We now consider the effect of the observation-induced multiplier on the actions of the players. Without repetition, where t is $Pr(T)$, we need that:

$$m(b_x - f_r) < b_x - f_r - f_t \Rightarrow mf_r < t(f_r + f_t - c_y) + (1-t)(f_r - c_y)$$

This should hold under the assumption that the action of helping is more costly for Y than the speech act of a request: $c_y > f_r$. The important thing to see here is that we have sustained cooperation without repetition. Under repetition, we should have lower threshold of discounting required. We would have the following for X:

$$b_x - f_r - f_t > (1-\delta)(m(b_x - f_r) + \delta \sum_{j=0}^{\infty} (-c_x - f_r)\delta^j) \Rightarrow \delta_X > \frac{m(b_x - f_r) + f_r + f_t - b_x}{c_x + f_r + m(b_x - f_r)}$$

For the case of a thankful X against Y who would not ask after not receiving help, we would have:

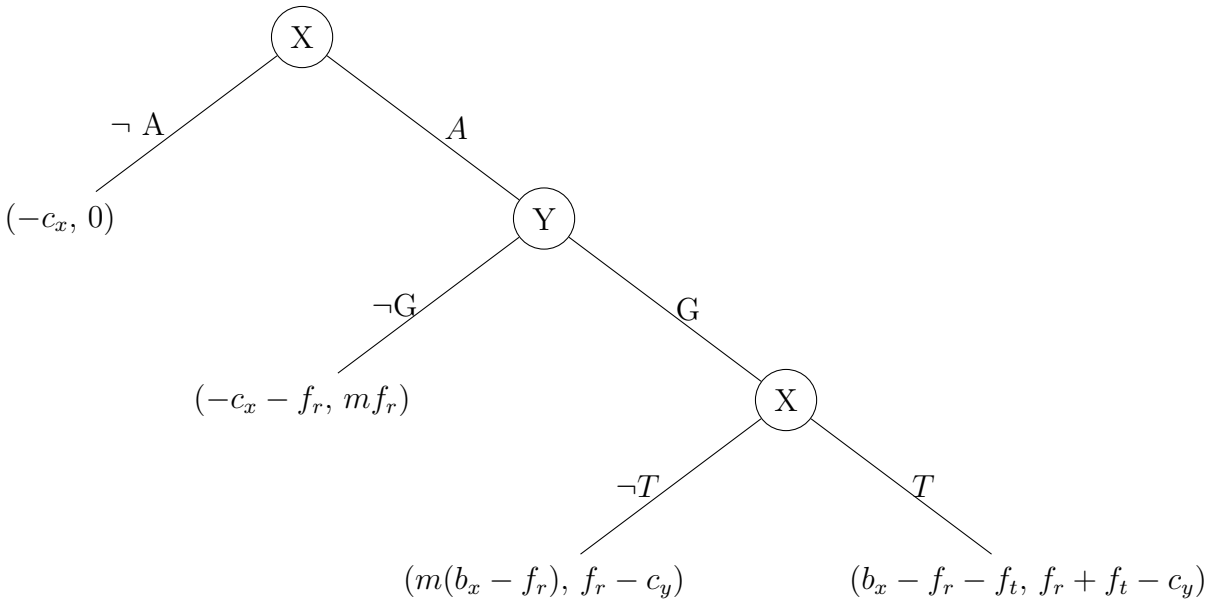

Figure 9.13: Observed Request Game with Face: Player X can choose to Ask (A) something from Player Y, who can then choose to Grant (G) the favor. Player X can choose to Thank (T) or not Thank ($\neg T$) player Y.

$$f_r + f_t - c_y > (1-\delta)(m(f_r) + \delta \sum_{j=0}^{\infty}(0)\delta^j) \Rightarrow \delta_X > \qquad \frac{mf_r - f_r - f_t + c_y}{mf_r}$$

Results and Reflection

How do the versions of the asymmetric game compare with observation, face, and repetition involved? Because

- the cost of helping for Y is greater than the cost of the speech acts
- the penalty against X for not being able to get help is greater than zero and
- Y is at a loss when helping,

we see that we need more stringent conditions to make Y play cooperative moves.

δ_X	Obs	δ_X Unobs
No Rep	YES	NO
Rep	$\frac{m(b_x-f_r)+f_r+f_t-b_x}{c_x+f_r+m(b_x-f_r)}$	$\frac{f_t}{b_x+c_x}$

δ_Y	Obs	δ_Y Unobs
No Rep	YES	NO
Rep	$\frac{mf_r-f_r-f_t+c_y}{mf_r}$	NO

Table 9.10: Comparison of face-augmented trust games with repetition. Note that by adding observation, we can guarantee the existence of cooperative outcomes. In the case of Y, who has the most cost, this may not work otherwise.

In some cases, this asymmetric game leaves something to be desired. What we see here is that

- Because the action $c_y > f_t + f_x$, we find it more difficult to sustain the norm of helping than of thanking. Note still however that with the case of thanks we have a possible solution when $c_y - f_t < f_x$
- The asymmetric game with thanks and face therefore still gives us no hope of cooperation in the discounted case unless $c_y < f_t + f_r$, which we claim is not always realistic
- Thus one option is to consider the game with symmetric roles, something already seen previously
- Another option is to consider the game with sympathy.

This is not so much discouraging as it is revealing. We need other mechanisms to promote the desired outcomes in the more costly actions, even when considering repetition. We have seen this already in repetition, that if we have symmetric roles, a threshold of patience often exists, at least under threat of punishment if not always under imitation. Are real humans so heartless and self-serving? We hope not, and thus we invoke sympathy preferences in accordance with Sally [2000, 2001]. .

9.4 Reciprocity Versus Repetition The Case of Other-Regarding

We now proceed to adapt our work on sympathy and other-regarding from Chapter 6 to the basic asymmetric trust games, the symmetrized versions, and the extended versions towards justifying classes of cooperative behavior in the absence of repetition. We then add in repetition to compare the levels of patience required by sympathetic players to sustain cooperation. As discussed before, there exists a wealth of theoretical work on Rabin [1993], Fehr and Schmidt [1998], Levine [1998], and behavioral Fehr and Schmidt [2001], Camerer [2003]

and neurobiological evidence Fehr [2008] of other-regarding preferences. Here we adapt the notion of sympathy as advanced by Sally [Sally, 2000, 2001] to explain the observed behavior. The central notion is that agents make choices based on the welfare of those around them. More formally, we return to our notion of a *sympathy distribution* over the payoffs of all the agents involved in the game. For each agent in a particular interaction, there is a distribution function, σ_i , such that $\sum_{j=1}^{n} \sigma_i(U_j) = 1$, which determines how much agents will weigh the welfare of others in their decisions. To recall, the perfectly self-interested agent of classical economics is such that $\sigma_i(U_j) = 0$ for all $j \neq i$. A selfless (purely altruistic) agent would be such that $\sigma_i(U_i) = 0$.

As the conversations or speech acts in which we are interested typically contain only two people, here we consider the limiting case of a single interlocutor. We can rewrite this function in a similar way, figuring the iconic s as the sympathy

parameter as done before. As we have used δ previously for a discount value, we also wish to avoid confusion between the repeated games notation and Sally's notation for sympathetic payoffs. This gives us:

$$V_i = (1 - s) \cdot U_i + s \cdot U_j \tag{9.1}$$

Sympathy and Repetition in Asymmetric Trust Games

In the following examples, we will follow a similar process for examining the interaction of sympathy with repetition. We will first compute the sympathy preferences required to promote cooperation in the one-shot scenarios. We will then see how sympathy affects the minimum discount values in repeated asymmetric games. Let us consider first the basic, generalized trust game and its counterpart, the fragile trust game. We will later move onto the three-stage extended games.

We can see in each of these games that in the one-shot version, according to the classical notion of utility, that each game will fail to produce cooperative behavior if we examine them with backwards induction. There must however be other motivations for cooperative behavior. We can see the breakdown for the various motivations in Table 9.4.

We should also make a comment on notation that we will denote the discount parameter mediated under sympathy as $\delta_X(s)$ or $\delta_Y(s)$. The idea is that sympathy is an additional input that should influence the discount value required to sustain the cooperative outcome. We will see several instance that show where players sympathetic to each other require less patience. Note that $\delta_i(0) = \delta_i$ for either player, i.e. we should be able to reduce the discount parameter to the one seen without sympathy by substituting $s = 0$. Further, note that we will use the classical utility function $U(\cdot)$ in cases without sympathy and $V(\cdot)$ in cases with sympathy. If we compare this to the games seen before, we can compute discount values for X to thank Y were we to undergo asymmetric repetition with opt-out punishment. We can then compute discount values for Y to help X. We will compare these discounts in the sympathetic and classical cases.

	No Sympathy U	Sympathy V
No Repetition	No Cooperation	Cooperation With s
Repetition	Cooperation With δ	Cooperation With $\delta(s)$

Table 9.11: In the four permutations of the games under inclusions of sympathy and repetition, we can find incentives for cooperative behavior. Even in some cases these may not be enough. Note that sympathy can influence how patient a player might be over the long run.

Although sympathy and repetition can provide incentives for cooperative behavior, this may not be enough in every case. In particular, the fragile trust game presents obstacles to cooperation because of the steep cost for Y to pay for helping. In order to simplify the reading of these next few sections, we will stick to the following objectives:

- Determine the minimum sympathy s for promoting the cooperative actions; e.g. $s > \frac{b}{b+c}$ OR
- Determine the minimum patience, or discount, δ for promoting the cooperative actions; e.g. $\delta > \frac{b-r}{b}$.

In the cases where there is already sympathy present, this discount δ will be thought of as $\delta(s)$, where we note that the discount δ is a *function* of s. These cases compared to those without sympathy should be consistent with values of $s = 0$. As all of these situations in the repeated game amount to solving linear inequalities based on the *one-shot deviation principle*, where we want it to be the case that the utility for a cooperative action C is higher than the utility of playing a selfish action D in the first round and dealing with the response R to that action in the later rounds, we will thus omit the calculations, recalling that the utility function V is a function of both the outcome of the game and the sympathy s:

$$V_i(s) = (1 - s) \cdot U_i + s \cdot U_j \tag{9.2}$$

This means we will calculuate cases where $\delta(s)$ is greater than some bound. We do this through the *one-shot deviation principle*, seen in the repeated game payoffs:

$$V_i(CC) > (1 - \delta)(V_i(DC) + \delta \sum_{j=0}^{\infty} V_i(DR)\delta^j) \tag{9.3}$$

In our games, the cooperative actions are asking A and helping H, while the responses can include several types of actions. In general, we base our analysis on the conditions promoting these actions. We will work through the analysis in detail for the basic trust game fromFigure 9.14 and omit the calculations for the other cases. In some cases, we will leave the game tree and diagrams of the region of socially feasible equilibria in the repeated game. As every game where there is no sympathy $s = 0$ corresponds to the original game, we will typically depict only the sympathetic cases. A full list of results follows at the end of the section.

Basic Trust Exemplar: Repetition and Sympathy

When we look at the basic trust game, we can investigate the patience and sympathy conditions required to sustain trustworthiness. First, we begin with the sympathy condition. Considering here the general case, we can perform backward induction on a player Y to see the minimum condition of sympathy required for him to play H in the game Figure 9.14. This occurs when

$$V_Y(H) > V_Y(\neg H) \Rightarrow s_Y > \frac{b - r}{b + c}$$

Figure 9.14: Trust Games: General and sympathetic payoffs.

We thus see that the sympathy threshold will be lower when the benefit of not helping compared to the reward of helping is small or when the difference between the benefit of not helping and the cost of being not helped is large. Another interesting note to compare might be under what sympathy value does X have an incentive to ask against an unhelpful Y. This occurs when

$$V_X(A) > V_X(\neg A) \Rightarrow s_X > \frac{c}{b+c}$$

This gives us the case that as the message cost goes up or the benefit of not helping goes down, we have a larger sympathy value required to promote the initial move. While the first criterion seems natural, perhaps the second deserves more attention. Paying attention to the costs means that more costly messages might not be sent if there is not sufficient sympathy on the part of the players.

Basic Trust Game Repeated with Sympathy

Previously, in Chapter 8 and section 9.3 , we saw that for a given discount value, $\delta_Y > \frac{b-r}{b}$, we can achieve the cooperative move on the part of Y in the repeated case. That we can achieve this outcome should align with the diagram in Figure 9.14. Recall that for Y, we wanted that

$$U_Y(H) > U_Y(\neg H) \Rightarrow \delta_Y > \frac{b-r}{b} = 1 - \frac{r}{b}$$

Now to the sympathetic case of repetition, whose feasible region we see in Figure 9.15. We should be able to see that the prior case is a reduction to $s = 0$. For Y, we want that

$$V_Y(H) > V_Y(\neg H) \Rightarrow \delta_Y(s) > \frac{b(1-s) - cs - r}{b(1-s) - cs} = 1 - \frac{r}{b(1-s) - cs}$$

While this notation may obscure what's happening, consider first that $b(1-s) - cs < b$ and that when we found the discount value for Y before in section 9.3, it was the case that $\delta > \frac{b-r}{b}$. Here we see that since $b(1-s) - cs < b$, we have now that the minimum value of patience required under sympathy is less than when

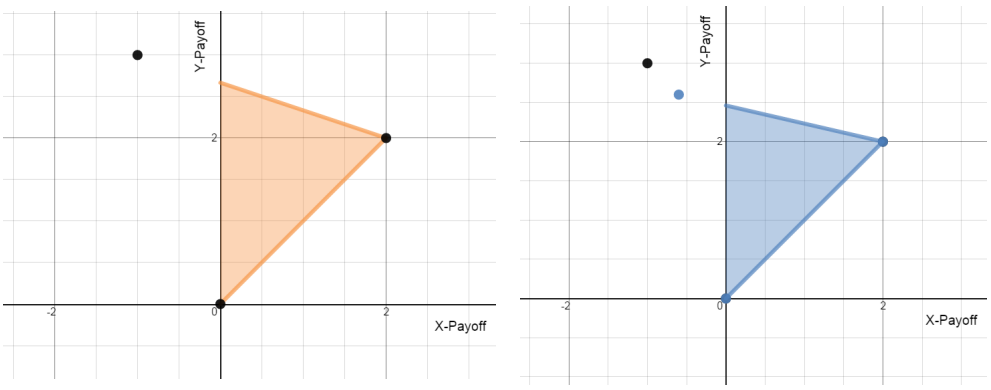

Figure 9.15: Here we see the feasible regions of equilibria for the basic trust game with $r = 2$, $b = 3$, $c = 1$, with sympathy $s = 0$ and $s = .1$. The point altered by sympathy is in blue, while the original corresponding to the sequence $A_X, \neg H_Y$ that gave a payoff of $-c, b$ is in black.

no sympathy is present. To see why, notice we can compare these two discounts δ_s, δ and see that with $\delta(s)$ we are subtracting a larger fraction from one, and thus we have:

$$\delta_Y(s) > 1 - \frac{r}{b(1-s) - cs}; \delta_Y > 1 - \frac{r}{b} \Rightarrow \delta_Y(s) < \delta_Y$$

Hence we see that in this case there does exist a value of sympathy that would promote cooperation and that sympathy lowers the patience required for agents to cooperate. Could we have an overabundance of sympathy that would skew the feasible region of payoffs? Yes, as we see in Figure 9.15.

Fragile Trust Games with Sympathy

In the fragile trust game, where Y pays an extra cost for helping, we should expect higher levels of sympathy than before. Here we see the two contrasted in Figure 9.16

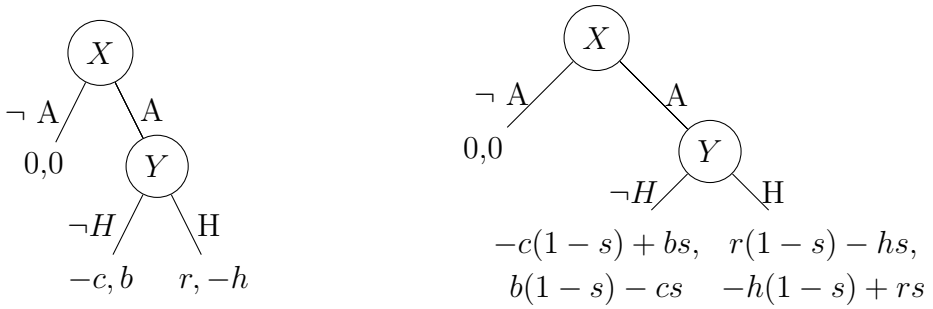

Figure 9.16: Fragile Trust Games: General and sympathetic payoffs.

Fragile Trust Game Repeated with Sympathy

First, consider the fragile trust games from Figure 9.16. Without sympathy, there does not exist a discount value that would promote the cooperative move for Y, as the threat of X opting out would in fact help Y. We found in section 9.3 that assuming this were possible would lead to a contradiction in the result $\delta > \frac{b+h}{b}$. Geometrically, this is because the pure strategy outcomes lie outside the feasible region of equilibria. We can see these as the black dots in Figure 9.17. What about in the sympathetic case? One option we have is to consider the feasible region altered by sympathy, as seen in Figure 9.17.

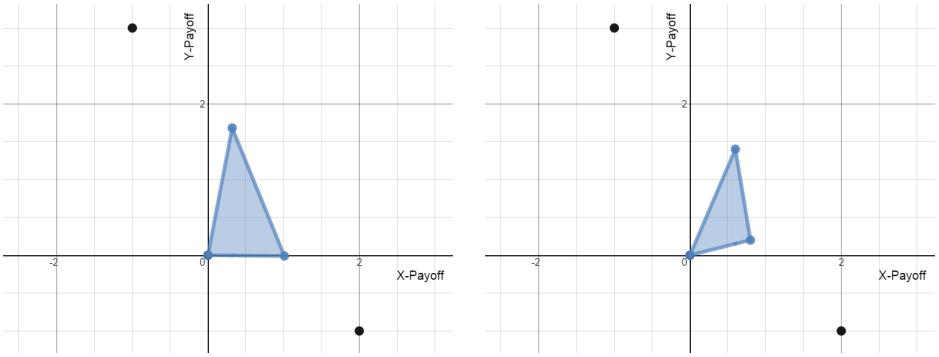

Figure 9.17: Feasible regions for the fragile trust game with parameters $r = 2$, $b = 3$, $c = 1$, $h = 1$, with sympathy $s = \frac{1}{3}$ and $s = .4$. The points altered by sympathy are in blue, while the originals are in black. For A_X, H_Y to become an equilibrium in the repeated game with sympathy, we must have at least $s > \frac{h}{r+h} = \frac{1}{3}$.

Extended Trust Games with Sympathy

Here we consider sympathetic payoffs on the extended trust games: basic, fragile, and unified. In this subsection, we will consider the extended version of the basic model, the *extended trust game*. Results under variations of sympathy and repetition follow at the end of the section.

The sympathetic payoffs can affect the feasible region, as seen in Figure 9.19. Note that as the outcome involving thanking has a transferable utility, it is equally as good for the group as the one without it. This is why both feature on the same frontier and why both move closer together with sympathy. We can see this in the game tree from Figure 9.18, where adding the utility of the players gives us $2r$ in both of the final outcomes.

Extended Fragile Trust Game

Just as before, we will examine the extended fragile game as to whether sympathy or repetition can promote the cooperative outcomes. As the game pictured in Figure 9.20 under $s = 0$ is the same as the original game without sympathy, we can use the tree in Figure 9.20 to proceed.

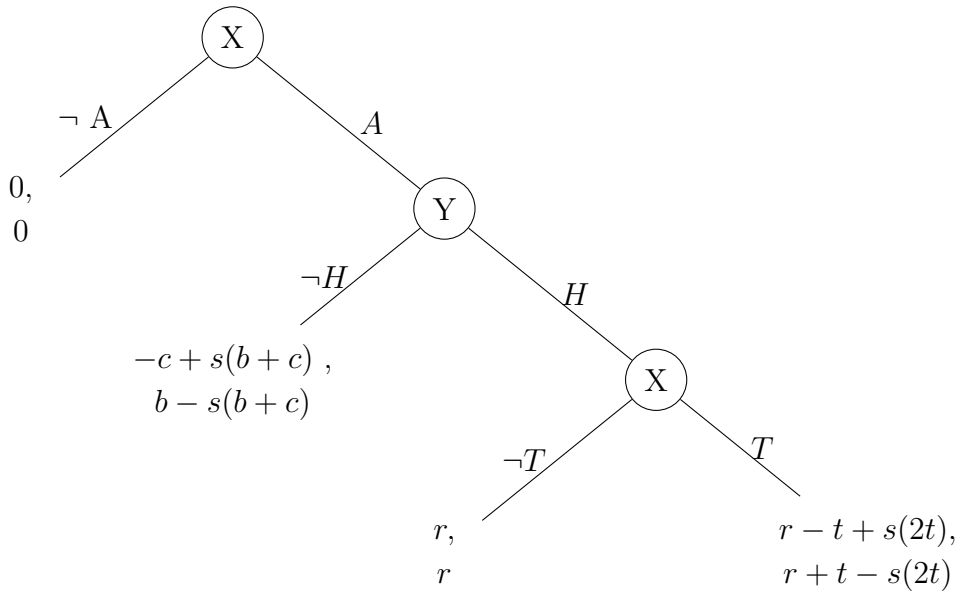

Figure 9.18: Extended Trust Game (Sympathy): Player X can choose to Ask (A) Player Y, who can then choose to Grant (G) the favor. Player X can choose to Thank (T) or not Thank ($\neg T$) player Y.

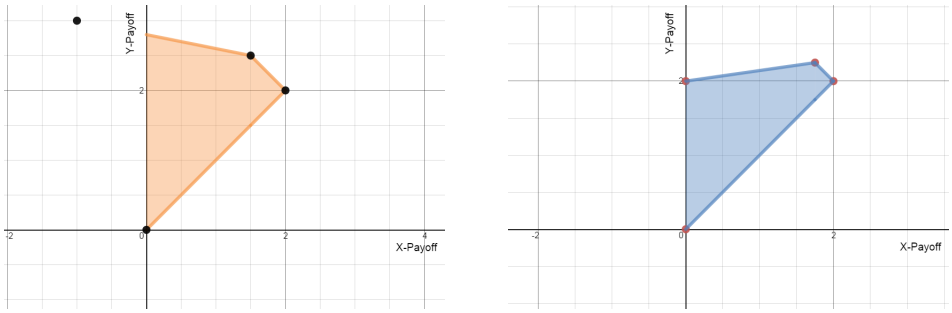

Figure 9.19: Feasible regions for the extended trust game with $r = 2$, $b = 3$, $c = 1$, $t = .5$, with no sympathy and with sympathy $s = .25$. $s = .5$ will collapse the region, while $s > .25$ allows the outcome of $A\neg HT$ to become an equilibrium.

Extended Fragile Trust Game Repeated with Sympathy

Under repetition, we want to know if the discount thresholds should be smaller with sympathy. Consider the feasible region predicted by the folk theorem in Figure 9.21. We can observe that the cooperative outcome lies outside of the feasible region, and thus we should not expect there to exist a discount threshold sufficient to promote cooperation on the last move.

First we begin with no sympathy. When considering the figure in Figure 9.21, we can see that the two outcomes where Y helps X certainly outperform the mutual minmax for player X, so there does exist a discount value where X would be thankful. However, upon closer inspection of the figure in Figure 9.21, we see that there should be no pure action pairing that would benefit Y beyond the

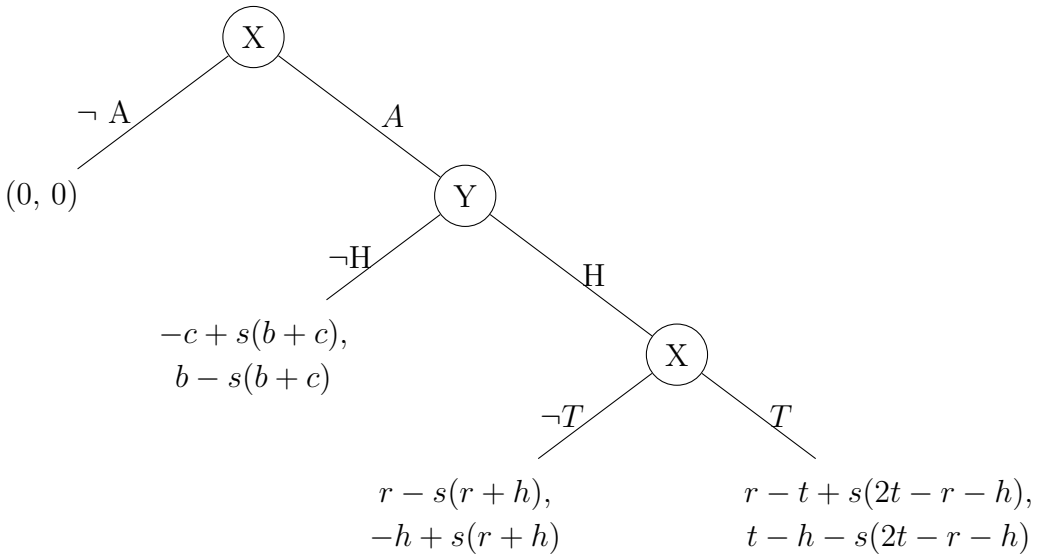

Figure 9.20: Extended Fragile Trust Game with Sympathy: X can choose to Ask (A) Y, who can choose to Grant (G) the favor. X can choose to Thank (T).

mutual minmax payoff of $(0, 0)$.

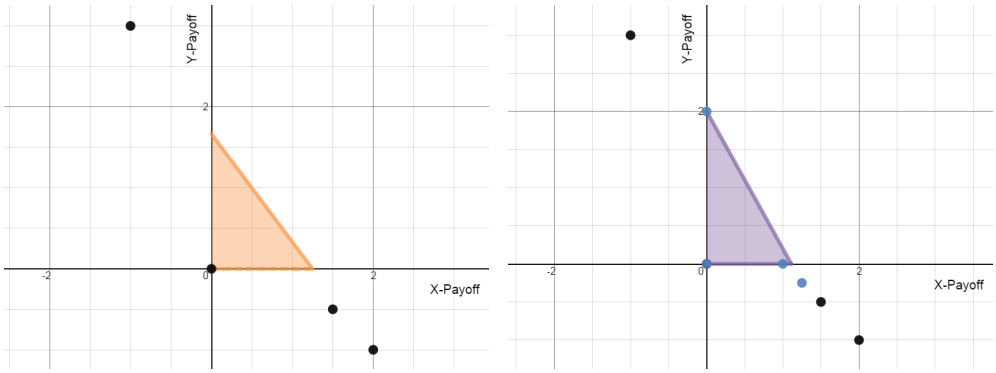

Figure 9.21: Feasible regions for the extended fragile trust game with parameters $r = 2$, $b = 3$, $c = 1$, $t = .5$, with sympathy $s = 0$ and $s = .25$. AHT will become feasible before $AH\neg T$, when $s = .25$. This is because the major hindrance to Y is $-h$, keeping his payoff below zero.

We can also see that the final outcomes are equally good for the group based when we consider the game tree from Figure 9.20. I.e. adding the utility of the players gives us $r - h$ in both of the final outcomes.

Unified Trust Game

Now we consider the unified trust game. This is in lieu of the trust game with face (section 9.3) from Quinley and Ahern [2012], as depicted in Figure 9.22. We prefer the cleaner notation from the unified trust game, in that we can account

for the face-based actions through the payoff modifications t and c. Results follow that detail the constraints on sympathy and discounting that will promote the cooperative outcomes in one-shot and repeated cases.

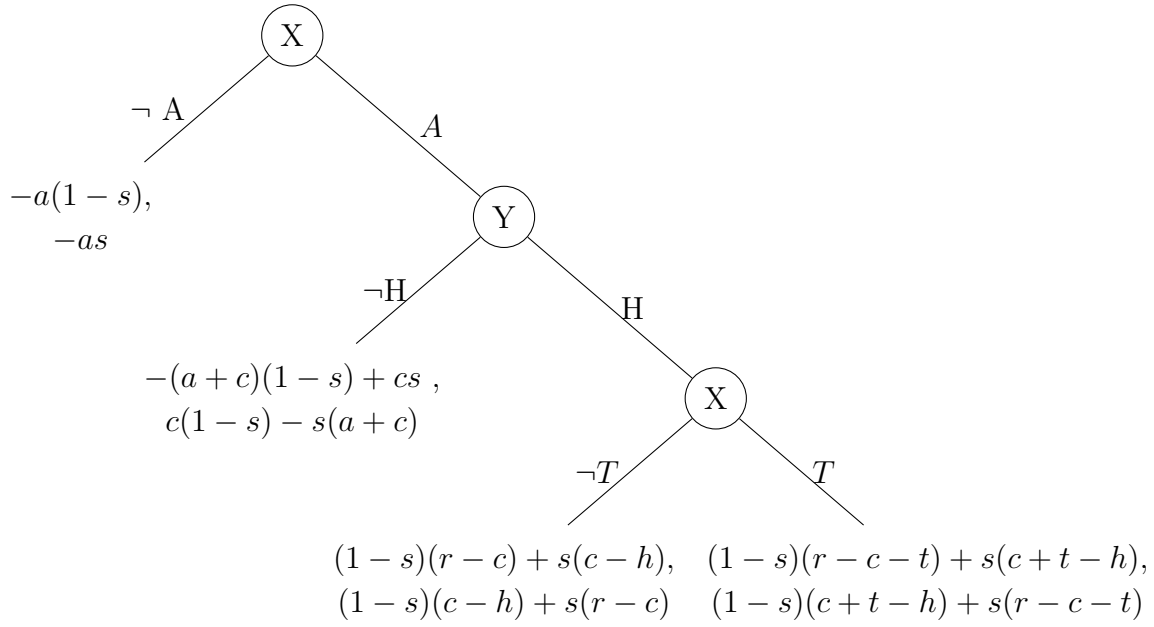

Figure 9.22: Unified Trust Game with Sympathy: Player X can choose to Ask (A) something from Player Y, who can then choose to Grant (G) the favor. Player X can choose to Thank (T) or not Thank ($\neg T$) player Y.

Results on Sympathy and Discounting

In this section we examined the two types of thresholds required for the cooperative move in every case: sympathy s and patience δ. Our first question was how the game variations of payoff alterations and extended forms affected the sympathy levels required to sustain the cooperative outcomes in the one-shot scenarios. Note that we considered patience as given by δ potentially to be a function of sympathy in the later cases. Thus our second question was how sympathy affected the discount values required to sustain the cooperative outcomes. I.e. are sympathetic partners more likely to cooperate even in cases where they are less future-oriented or less likely to see each other?

Basic Results on Sympathy

We saw in our analysis previously that the fragile game requires more sympathy on the part of Y to sustain the cooperative outcome than the basic game does, as seen in Table 9.30. This is not surprising, as this game gives Y a payoff of $-h$ vs. r for helping in the one-shot scenario. Thus we should have it be the case that there must be an additional incentive to promote cooperation when taking the loss.

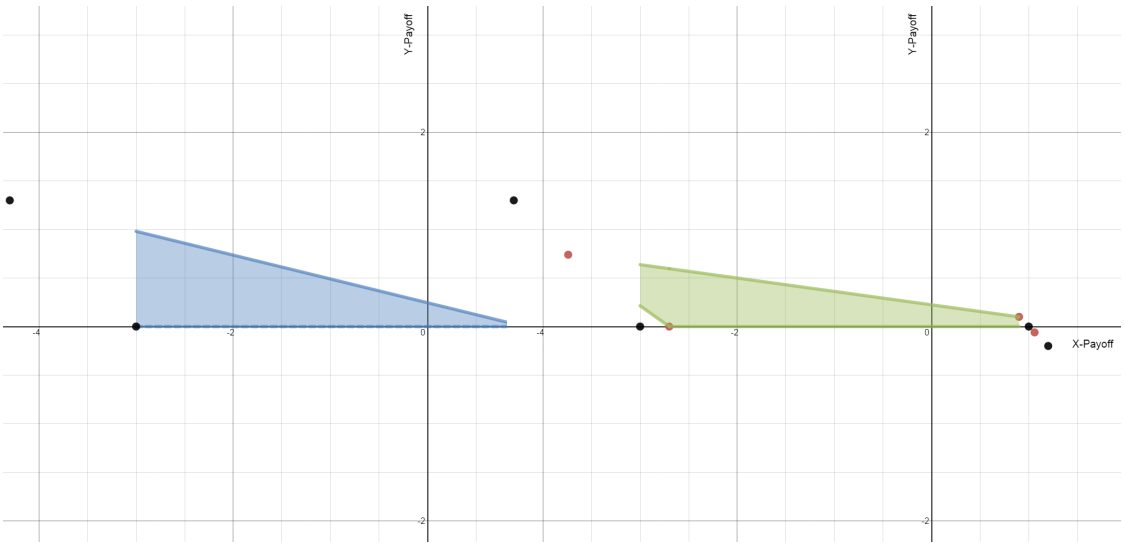

Figure 9.23: Feasible regions for the unified trust game with $a = 3$, $r = 2.5$, $h = 1.5$, $c = 1.3$, $t = .2$ and with sympathy $s = 0$ and $s = .1$. Beyond certain sympathy thresholds, this region loses its convexity. Sympathetic outcomes are in red.

Game vs Action	A_X	G_Y
Basic	$s > \frac{c}{b+c}$	$s > \frac{b-r}{b+c}$
Fragile	$s > \frac{c}{b+c}$	$s > \frac{b+h}{b+h+c+r}$

Table 9.12: Sympathy Thresholds in Two-Stage Trust Game: In each entry, the sympathy required to promote the cooperative outcome is given.

How do the constraints on sympathy in the extended versions compare to their counterparts? We see the results in Table 9.13. In the case of the basic game, we saw that we have a plausible constraint on the sympathy required for Y to help X. It was then the case that if $c + 2r > b$, then we should expect a greater amount of sympathy required to help the thankless X, as seen in the entry under $(G_Y|\neg T_X)$. In the fragile game, we saw a surprising result, that the sympathy required for Y to help a thankless X was in fact less than that required to help a thankful X. From the table, this is because $\frac{b+h-t}{b+h+c+r-2t} > \frac{b+h}{b+h+c+r}$. In the unified game, we see again the natural result that the sympathy required to help a thankful X is less, as $\frac{h}{a+r+h} > \frac{h-t}{a+r+h}$.

In these variants on the game, we see at least several trends

- In the extended versions, the constraint for X that $s > \frac{1}{2}$ is perhaps unrealistic except in the case of very close association
- It is not always the case that a thankful X means less sympathy is required for Y to sustain the cooperative moves.
- Realistic levels of sympathy may not always be enough to promote the desired outcomes, and thus additional mechanisms may be required.

Game vs. Action	T_X	$(G_Y \mid T_X)$	$(G_Y \mid \neg T_X)$
Basic	$s > \frac{1}{2}$	$s > \frac{b-(r+t)}{b+c-2t}$	$s > \frac{b-r}{b+c}$
Fragile	$s > \frac{1}{2}$	$s > \frac{b+h-t}{b+h+c+r-2t}$	$s > \frac{b+h}{b+h+c+r}$
Unified	$s > \frac{1}{2}$	$s > \frac{h-t}{a+r+h}$	$s > \frac{h}{a+r+h}$

Table 9.13: Sympathy Thresholds in Extended Trust Games: In each entry, the sympathy required to promote the cooperative outcome is given. The constraints for Y condition on the type of X. Note that if $t = 0$, the constraints are equivalent.

Results on Discount Values

One question we would like to ask is whether the sympathetic versions of the games give us a lower discounting level required to sustain the cooperative outcomes. Does this gives us that with a higher level of sympathy, we should maintain lower thresholds of patience required for cooperation? Note that is analogous to asking whether agents with higher levels of sympathy would help each other even with a low probability of future interaction. If this is the case, we have a natural analog to the concept of *social distance* seen in Brown and Levinson [1987]. I.e. we can think of sympathy decreasing with levels of social distance in a function like $s = \frac{1}{d+1}$. First we consider the results of the two-stage games in Table 9.14, comparing the discount values achieved sans sympathy to those with. Note that in these tables, it will always be the case that the values found without sympathy are equivalent to $s = 0$.

Game vs. Action	G_Y	$G_Y(s)$
Basic	$\delta > \frac{b-r}{b}$	$\delta(s) > \frac{b(1-s)-cs-r}{b(1-s)-cs}$
Fragile	$\delta > \frac{b+h}{b}$	$\delta(s) > \frac{(b+h)(1-s)-s(r+c)}{b(1-s)-cs}$

Table 9.14: Discount Thresholds in Two-Stage Trust Game. Here we see the discount threshold required to sustain the cooperative action G. Note that the fragile trust game has an impossible discount threshold.

As we saw previously that $\frac{b-r}{b} > \frac{b(1-s)-cs-r}{b(1-s)-cs}$, we know that for the basic game in Table 9.14, we have that sympathy induces lower levels of discounting required to sustain the cooperative outcome. In the fragile case, it was not possible to sustain the cooperative outcome, even through repetition, as $\frac{b+h}{b} > 1$. Sympathy made that possible, with the constraint that $s > \frac{h}{r+h}$. Note further that we do not in any case include a discount value that would make X choose the asking action, as there is no credible threat of punishment if he does not in the asymmetric game. This is analogous to the claim from Bicchieri [2006] that trustworthiness is an enforceable norm but trust is not.

Now we move on to the extended game results seen in Table 9.15, Table 9.17 ,and Table 9.16. First we consider the results for X in Table 9.15. In the games without sympathy, we see that cooperation is possible as $\frac{t}{a+r} < 1$ and $\frac{t}{r+c} < 1$. For the basic game extended, it should be clear that the discount value is less in

Game vs. Action	T_X	$T_X(s)$
Basic	$\delta > \frac{t}{r+c}$	$\delta > \frac{t(1-2s)}{r+c}$
Fragile	$\delta > \frac{t}{r+c}$	$\delta > \frac{t(1-2s)}{r+c-s(b+c)+s(r+h)}$
Unified	$\delta > \frac{t}{a+r}$	$\delta > \frac{t(1-2s)}{a+r-s(a+r+h)}$

Table 9.15: Discount Thresholds for X in Extended Trust Game with and without Sympathy

the sympathetic version of the game. What about the fragile version? So long as $3r + h + c > b$, we will have the patience required under sympathy is less. In the unified version, we require that if $r + a > h$, then we will also have the patience required for X under sympathy is less than that without.

Game vs. Action	$G_Y \mid T_X$	$G_Y \mid \neg T_X$
Basic	$\delta > \frac{b-r}{b}$	$\delta > \frac{b-r}{b}$
Fragile	$\delta > \frac{b+h-t}{b}$	$\delta > \frac{b+h}{b}$
Unified	$\delta > \frac{h-t}{c}$	$\delta > \frac{h}{c}$

Table 9.16: Discount Thresholds for Y in Extended Trust Game without Sympathy. Note that we are comparing against the two types of X. These values should be equivalent to the cases where $s = 0$.

Game vs. Action	$G_Y \mid T_X(s)$	$G_Y \mid \neg T_X(s)$
Basic	$\delta(s) > \frac{b-r-t-s(b+c-2t)}{b-s(b+c)}$	$\delta(s) > \frac{b-r-s(b+c)}{b-s(b+c)}$
Fragile	$\delta(s) > \frac{b+h-t+s(2t)-s(b+c+r+h)}{b-s(b+c)}$	$\delta(s) > \frac{b+h-s(b+c+r+h)}{b-s(b+c)}$
Unified	$\delta(s) > \frac{h-t-s(h+r+a-2t)}{c(1-2s)}$	$\delta(s) > \frac{h-s(h+r+a)}{c(1-2s)}$

Table 9.17: Discount Thresholds for Y in Extended Trust Game with Sympathy. Note that we are comparing against the two types of X. The cases where $s = 0$ are equivalent to the cases without sympathy in the previous table.

How should we interpret the table of results for Y in Table 9.17? First we should note that it is not always the case that repeated interaction or sympathy alone lead to the desired outcomes in the extended games. Second, we reserve that the plausible condition of $s < \frac{1}{2}$ should be respected in most cases. In the fragile case, we see a highly pessimistic picture. It appears that in the cases without sympathy there is a clearly insurmountable level of patience required. In the cases with sympathy, we may need to rely on constraints like message cost or the cost of helping to satisfy the conditions.

In the basic case, we see there is a plausible discount value that will promote cooperation, even in the absence of sympathy. The sympathy values make cooperation more plausible, as the discount value required with no sympathy is higher

for all values of r, b, c, s. We also note in the basic game that the discount value required for cooperation against a thankful opponent is less so long as $s < \frac{1}{2}$, which should fit with our previous condition as well as intuition. Thus we see less patience required to sustain cooperation with a thankful partner.

In the case of the extended fragile and unified games, we see that there is no chance of a discount value that is less than one without sympathy. In the fragile game, we have that the natural result that the patience required for cooperation is less for sympathetic players, so long as $\frac{b}{h} > \frac{c}{r}$, a result independent of the sympathy parameters. Note that this may not be guaranteed however. Is the cooperative move possible however? We saw earlier that for sufficient levels of sympathy it is, given that $s > \frac{h}{r+h}$, then we will have that $\delta(s) < 1$. In our previous analysis for the fragile game, we found the surprising result that the discount value of the game with sympathy was less than the game without it for Y but not for X. This means that a sufficiently patient Y might choose the cooperative move were he sympathetic to X.

In the unified game, we see that given the plausible condition $r + a > h$, the natural result emerged: the required discount value for Y in the unified game was less with sympathy. In this case if $s > \frac{c-h}{2c-h-r-a} = \frac{h-c}{h+r+a-2c}$, then we have that the cooperative move is possible. [3] In both the fragile and unified game, we see once again that so long as $s < \frac{1}{2}$, the patience required against a thankful player is less than against an thankless one.

Sympathy in Repeated Games

Basing the discount values on the sympathetic outcomes allows us to reverse-engineer the sympathy conditions that would make cooperation possible in the cases where it previously was not. This also allows us to place a ceiling on the cases where too much sympathy might give agents reason to act irrationally. For instance, an outcome might originally be a feasible Nash Equilibrium but an excess of sympathy could remove it from the repeated game. [4] The most interesting decision point to consider is whether Y should help X in the extended version of the trust game. We can therefore compare decisions within the game, as seen in Table 9.31.

Game vs. Action	$G_Y \vert T_X(s)$	$G_Y \vert \neg T_X(s)$
Basic	$s < \frac{r+t}{2t}$	$r > 0$
Fragile	$s > \frac{h-t}{r+h-2t}$	$s > \frac{h}{r+h}$
Unified	$s > \frac{h-(t+c)}{r+h+a-2(t+c)}$	$s > \frac{h-c}{r+h+a-2c}$

Table 9.18: Sympathy across opponent type for Y in extended games. Discounts on repeated interaction give us sympathy conditions necessary to promote the cooperative moves and guarantee that the discounts are plausible, i.e. $\delta < 1$. The basic game had trivial constraints, while the others had minimum values. These should be equivalent in the event that $t = 0$.

[3] Note that we rearranged the inequality and the fraction, as $2c - h - r - a < 0$.

[4] Something like a lost cause.

We see a few interesting predictions. First note that the basic extended game already had the case that the cooperative outcome was in the feasible region. Thus we have that we want constraints that keep it that way. With thanks, we have $s < \frac{r+t}{2t}$. This is trivially true, as $r > t$. Notice that without thanks, we want that $r > 0$. This is also trivially true.

Now to the fragile and unified cases, where the initial circumstances did not provide that the cooperative outcomes without sympathy could be achieved in the repeated cases. Thus we wanted to consider that if $\delta(s) < 1$, then what value should s_Y have? For the fragile game, we had that $\delta(s) < 1 \Rightarrow s > \frac{h-t}{r+h-2t}$ or $s > \frac{h}{r+h}$. In the case of our numerical payoffs, this amounted to $s > .25$ or $s > \frac{1}{3}$. In the unified game, we found that we wanted $s > \frac{h-(t+c)}{r+h+a-2(t+c)}$ in the case of a thankful opponent or $s > \frac{h-c}{r+h+a-2c}$ in the case of a thankless opponent. In the case of our numerical payoffs, this was $s > \frac{1}{6}$ and $s > .079$, a very plausible set of constraints.

As X had the final move in the one-shot extended game, we saw that these games required $s_X > \frac{1}{2}$. How does this compare to the sympathy values elicited in the repeated game? Table 9.13 gives us that each of these games in the repeated versions would trivially work for X, as each of them had payoffs for X that outperformed the minmax profile. Thus we see that any sympathy value would suffice. More important to consider is that it was Y who needed sympathetic payoffs to increase his utility.

s_X	One-Shot	Repeated
Basic	$s > \frac{1}{2}$	$s > 0$
Fragile	$s > \frac{1}{2}$	$s > 0$
Unified	$s > \frac{1}{2}$	$s > 0$

Table 9.19: Sympathy Thresholds for X in Extended Trust Games: One-Shot vs. Repeated. Without repetition, the sympathy required for X was unrealistic. With repetition, the cooperative outcome was always beneficial for X but often not for Y.

The next question should be how do the sympathy values for Y compare to the games without repetition? We saw earlier that the constraints for the basic game can be satisfied with purely sympathy or purely repetition. Combining the two gives us a trivial constraint on the amount of sympathy required to maintain $\delta(s) < 1$. For the fragile game this was not the case. This game required either a high level of sympathy or a high discount value. In the one-shot case, the sympathy required was $\frac{b+h}{b+h+c+r}$ vs. $\frac{h}{r+h}$. In the figure depicting the feasible region and parameters from Figure 9.17, we have that $\frac{b+h}{b+h+c+r} = .57$ vs. $\frac{h}{r+h} = .33$, giving us that the game with repetition allows a more realistic sympathy value.

Last, we compare the sympathy constraints seen in the one-shot versions of the extended games with the repeated versions. The important thing to note is that the repeated versions of the fragile and unified games required both sympathy and repetition to ensure the cooperative outcomes.

s_Y	One-Shot	Repeated
Basic	$s > \frac{b-r}{b+c}$	$r+1>0$
Fragile	$s > \frac{b+h}{b+h+c+r}$	$s > \frac{h}{r+h}$

Table 9.20: Sympathy Thresholds for Y in Two-Stage Trust Game: In each entry, the sympathy required to promote the cooperative outcome is given across the variation in game type. We see that the basic game has a trivially true condition independent of s. For our parameters, we have that $\frac{h}{r+h} < \frac{b+h}{b+h+c+r}$.

s_Y vs. T_X	One-Shot	Repeated
Basic	$s > \frac{b-r-t}{b+c-2t}$	$r+1>0$
Fragile	$s > \frac{b+h-t}{b+h+c+r-2t}$	$s > \frac{h-t}{r+h-2t}$
Unified	$s > \frac{h-t}{a+r+h}$	$s > \frac{h-t-c}{a+r+h-2(c+t)}$

Table 9.21: Sympathy Thresholds for Y in Extended Trust Games(Thankful Opponents): In each entry, the sympathy required to promote the cooperative outcome is given across the variation in game type against a thankful X. The basic game has a condition independent of s. For our parameters, the repeated games require less sympathy.

For the fragile game, we see two interesting patterns. Across game types, the sympathy required in both of the one-shot cases was substantially higher, roughly two to one ($\approx .58 : .33$ and $\approx .57 : .25$). Second, we see a twist: within the one-shot game, the sympathy required against a thankful opponent was more, ($\frac{b+h-t}{b+h+c+r-2t} \approx .58$, $\frac{b+h}{b+h+c+r} \approx .57$), whereas in the repeated case it was less ($\frac{h}{r+h} \approx .33$, $\frac{h-t}{r+h-2t} \approx .25$. The result in the repeated game seems more natural, and it may be due to slight variation in the parameters that gives us the mildly surprising result in the one-shot case.

s_Y vs.$\neg T_X$	One-Shot	Repeated
Basic	$s > \frac{b-r}{b+c}$	$--$
Fragile	$s > \frac{b+h}{b+h+c+r}$	$s > \frac{h}{r+h}$
Unified	$s > \frac{h}{a+r+h}$	$s > \frac{h-c}{a+r+h-2c}$

Table 9.22: Sympathy Thresholds for Y in Extended Trust Games(Thankless Opponents): In each entry, the sympathy required to promote the cooperative outcome is given across the variation in game type against a thankless X. The basic game has a condition independent of s. For our parameters, the repeated games require less sympathy.

The unified game presents a more intuitive picture. First, both of the repeated games require less sympathy than their one-shot counterparts. Second, both types of games required less sympathy for a thankful opponent. In the parameters

stipulated before from Figure 9.22, for the one-shot game, we had $\frac{h}{a+r+h} \approx .26$ and $\frac{h-t}{r+h+a} \approx .19$. For the repeated case, we had $\frac{h-c}{r+h+a-2c} = .17$ and $\frac{h-(t+c)}{r+h+a-2(t+c)} \approx .08$.

In sum, the sympathy parameter gave us the possibility of cooperation in the asymmetric cases, even when there was none even in the repeated version of the asymmetric game. As according with Sally [2001] and Camerer [2003], we claim that sympathy, repetition, and patience provide a psychologically plausible model that delivers results even in seemingly intractable games. We now move into the impact of sympathy with the repeated symmetric games.

9.5 Sympathy and Symmetry

We have yet to see the full-scale interpretation of sympathetic payoffs on the symmetric trust games. We begin that in this section. Our goals are threefold:

- discover thresholds of sympathy that promote new pure strategy equilibria in the feasible region,
- find thresholds of sympathy and patience that promote the cooperative outcomes, and
- explore the connection between sympathy and patience values δ.

Two-Stage Games: Basic and Fragile

We will begin with the two-stages games under the dynamics of imitation and punishment. Our first goal will be to see what discount values are brought about by the sympathetic payoffs. Our subsequent goal will analyze whether these outcomes are feasible and what sympathy values they require. When appropriate, we will use the derived sympathy values to reconstruct diagrams of the feasible regions under sympathy.

Under the two dynamics, we will investigate the mechanisms for enforcing the *friendly* outcome of (AH, AH), where we will see when AH is a best response to itself. In imitation, a player imitates his rival should that rival defect from the strategy, much like *tit-for-tat*. Under punishment dynamics, a player will punish his rival by playing the strategy that would minimize the score of a defector, much like *Grim Trigger*. If the players are sympathetic, the asymmetric payoffs will be altered. The symmetric payoffs remain the same. We will then consider the discount threshold required as a function of the sympathy involved.

As the calculations required follow the same scheme, we will limit our explanation to the following example of the Symmetric Trust Game with sympathy and the fragile game, as its payoffs present an interesting barrier to overcome. The table of results for the remaining games will follow at the end of the section.

The Symmetric Trust Game with Sympathy

We begin with the basic game, whose table we see in Table 9.23. In this case we will examine when (AH, AH) yields a higher payoff than playing $A\neg H$ first

	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 9.23: Symmetric Trust Game: Nash Equilibrium of ($\neg A\neg H$,$\neg A\neg H$).

and then being imitated, with players having sympathetic payoffs. Note that this should align with $s = 0$ and our result from earlier where $\delta_{IMI} > \frac{b-r}{r+c}$. To indicate the discount threshold as a function of sympathy, we will write $\delta(s)$.

$$U(AH) > U(A\neg H) \Rightarrow \delta_{IMI}(s) > \frac{b - r - s(b+c)}{r + c - s(b+c)}$$

Now to punishment. Note that this should align with $s = 0$ and our result from earlier where $\delta_{IMI} > \frac{b-r}{b+r+c}$.

$$U(AH) > U(A\neg H) \Rightarrow \delta_{PUN}(s) > \frac{b - r - s(b+c)}{b + r + c - 2s(b+c)}$$

Is it now the case that the discounts are more forgiving with sympathy? Yes, and by a large margin. First, we see $\delta_{IMI}(s) < \delta_{IMI}$, where $\delta_{IMI} \approx 0.3$ and $\delta_{IMI} \approx 0.2$ for a small sympathy value of $s = 0.1$. Likewise, we have that $\delta_{PUN}(s) < \delta_{PUN}$, where $\delta_{PUN}(s) \approx 0.10$ and $\delta_{PUN} \approx 0.10$. Note that the swing is not as dramatic in the case of punishment in absolute terms, as there is a small "benefit" to being punished in the sympathetic case, given by sb in the long run. Then again, the ratio of the two is greater still in the case of punishment, so in relative terms the swing is larger.

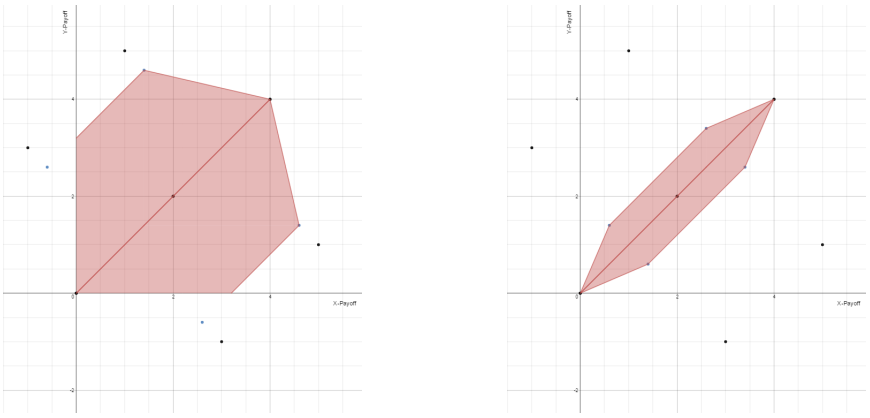

Figure 9.24: Here we see the feasible region of equilibria for the Symmetric Trust Game $r = 2$, $b = 3$, $c = 1$,and with sympathy $s = .1$ and $s = .4$ The sympathetic outcomes are in blue.

We can see the feasible region for this game with its sympathetic convex hull in Figure 9.24, but calculations indicate that for sympathy values above approximately $s \approx 0.23$, we will have negative discount values predicted by our punishment and imitation dynamics. This level of sympathy should give us that we have a guarantee that our preferred strategy will be a best response to itself. The interpretation is that any level of patience should suffice for such high levels of sympathy.

Given that the sympathetic discount values are lower in general, we now derive the sympathy conditions under which a discount value can promote cooperation in the first place. Thus in every case, we will check what happens when $\delta < 1$. First we begin with imitation and punishment in the Symmetric Trust Game. Note again that $\delta_{IMI}(s)$ indicates the discount as a *function* of s, not as a multiple of it.

$$\delta_{IMI}(s) < 1 \Rightarrow b < 2r + c$$

$$\delta_{PUN}(s) < 1 \Rightarrow s < \frac{2r + c}{b + c}$$

In the first case, our constraint is not dependent on s, and thus a high enough reward or costly enough speech act would promote the cooperative action. In the second case, we have an easily satisfiable condition, as $\frac{2r+c}{b+c}$ is greater than one. Now to the fragile case, where cooperation has proven previously difficult.

The Fragile Symmetric Trust Game with Sympathy

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

Table 9.24: Symmetric Fragile Trust Game: Nash Equilibrium of ($\neg A\neg H$,$\neg A\neg H$). Given the parameters, the two underlined outcomes could both feature in the repeated game.

We now proceed to the fragile game, whose table we see in Table 9.24. In this case we will examine when (AH, AH) yields a higher payoff than playing $A\neg H$ first and then being imitated, with players having sympathetic payoffs. Note that this should align with $s = 0$ and our result from earlier where $\delta_{IMI} > \frac{b+h}{r+c}$. Earlier this led to an impossible discount threshold to cross. Here we see the following similar case.

$$U(AH) > U(A\neg H) \Rightarrow \delta_{IMI}(s) > \frac{b+h-s(r+h+b+c)}{r+c-s(r+h+b+c)}$$

Now to punishment. This aligns with $s = 0$ and our result from earlier where $\delta_{IMI} > \frac{b+h}{b+r+c}$.

$$U(AH) > U(A\neg H) \Rightarrow \delta_{PUN}(s) > \frac{b+h-s(r+b+h+c)}{b+r+c-s(r+h)-2s(b+c)}$$

Is it now the case that the discounts are more forgiving with sympathy? In this case we get a shocking "No"! Whereas the case of punishment without sympathy presented a plausible discount, in this case $\delta > \frac{2}{3}$, we have that increasing the sympathy parameter increases the required discount in the fragile game. As a result, the discount under imitation remains impossible.

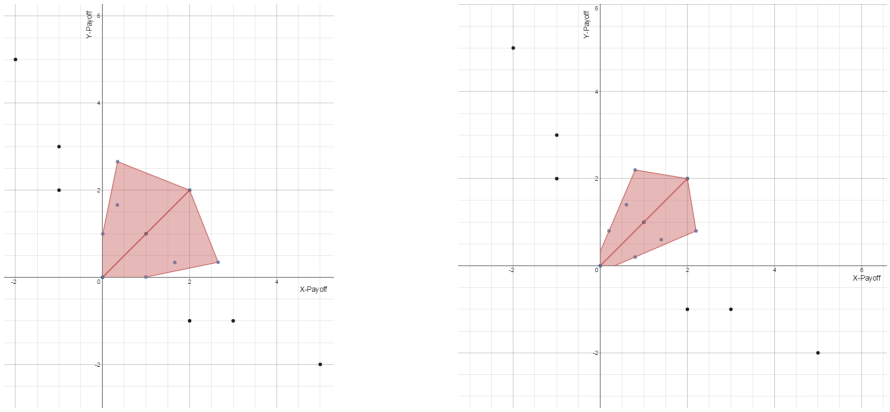

Figure 9.25: Here we see the feasible region of equilibria for the fragile Symmetric Trust Game $r = 2$, $b = 3$, $c = 1$, $h = 1$, and with sympathy $s = .33$ and $s = .4$ The sympathetic outcomes are in blue.

We now derive the sympathy conditions under which a discount value might promote cooperation, as our constraints impose harsh boundaries on the discount values.

$$\delta_{IMI}(s) < 1 \Rightarrow b + h < r + c \Rightarrow \perp$$

$$\delta_{PUN}(s) < 1 \Rightarrow s < \frac{r+c-h}{b+c}$$

While the first case gives an expected contradiction, we have that the second provides a maximum bound on the sympathy allowed. In this case, $\frac{r+c-h}{b+c} = .5$, and thus we have a plausible constraint. It is interesting to note that in the other cases, we are looking for a minimum requirement on sympathy, but in this one

we have an upper limit.

Extended Symmetric Trust Game with Sympathy

	ATH	$A\neg TH$	$\neg AH$	$AT\neg H$	$A\neg T\neg H$	$\neg A\neg H$
ATH	2r,2r	2r-t,2r+t	r-t,r+t	r+t-c,b+r-t	r-c,b+r	-c,b
$A\neg TH$	2r+t,2r-t	2r,2r	r,r	r+t-c,b+r-t	r-c,b+r	-c,b
$\neg AH$	r+t,r-t	r,r	0,0	r+t,r-t	r,r	0,0
$AT\neg H$	b+r-t,r+t-c	b+r-t,r+t-c	r-t,r+t	b-c,b-c	b-c,b-c	-c,b
$A\neg T\neg H$	b+r,r-c	b+r,r-c	r,r	b-c,b-c	b-c,b-c	-c,b
$\neg A\neg H$	b,-c	b,-c	0,0	b,-c	b,-c	$\boxed{0,0}$

Table 9.25: Extended Symmetric Trust Game: A player who does not play A will also not have the choice of $T \vee \neg T$. Nash Equilibrium of $(\neg A\neg H, \neg A\neg H)$

We consider the game whose table we see in Table 9.25. In this case we will examine when (ATH, ATH) yields a higher payoff than playing $A\neg T\neg H$ first and then being imitated, with players having sympathetic payoffs. Note that this should align with $s = 0$ and our result from the two-stage game earlier where $\delta_{IMI} > \frac{b-r-s(b+c)}{r+c-s(b+c)}$. Results follow at the end of the section.

This case is exactly the same under sympathetic payoffs as the two-stage game, as the symmetric outcomes eliminate the effect of thanking the other player. We can observe the feasible regions in the figures below, Figure 9.26 and Figure 9.27.Since the discounts are identical under the sympathetic parameters, we will not have to investigate the sympathy conditions implied by the discount.

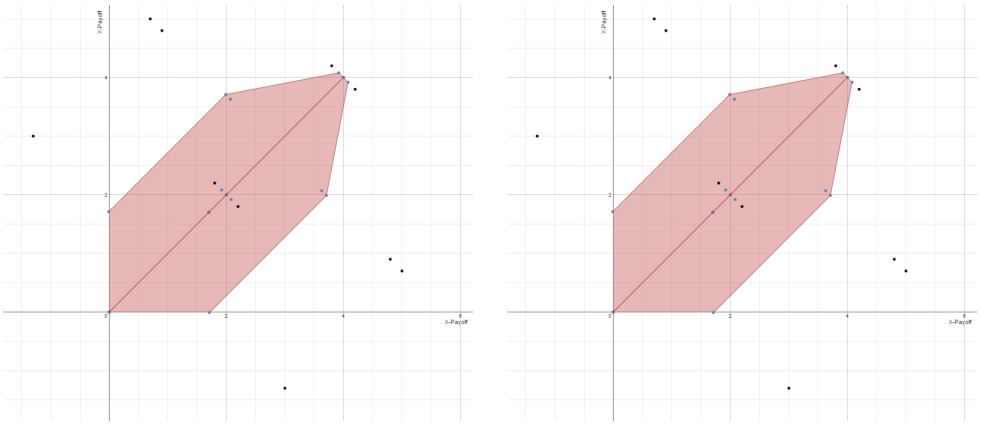

Figure 9.26: Feasible region for the Extended Symmetric Trust Game $r = 2$, $a = 3$, $c = 1.3$, $t = .2$ and with sympathy $s = .3$ and $s = .4$ The sympathetic outcomes are in blue, while the originals are in black.

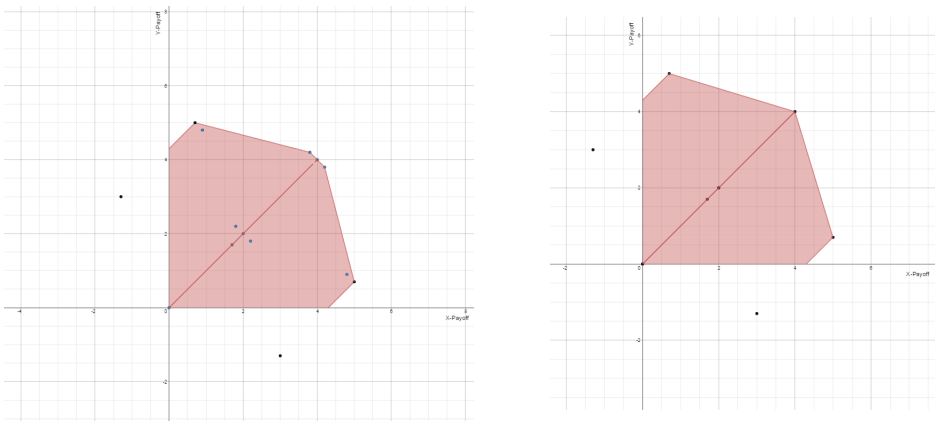

Figure 9.27: Feasible regions for the Extended Symmetric Trust Game and the original Symmetric Trust Game, both with $r = 2$, $a = 3$, $c = 1.3$, $t = .2$. Outcomes from the extended version involving thanks are in blue.

	ATH	$A\neg TH$	$\neg AH$	$AT\neg H$	$A\neg T\neg H$	$\neg A\neg H$
ATH	r-h,r-h	r-h-t,r+t-h	r-t,t-h	t-c-h,r-t+b	-c-h,r+b	-c,b
$A\neg TH$	r+t-h,r-h-t	r-h,r-h	r,-h	t-c-h,r-t+b	-c-h,r+b	-c,b
$\neg AH$	t-h,r-t	-h,r	0,0	t-h,r-t	-h,r	0,0
$AT\neg H$	r-t+b,t-c-h	r-t+b,t-c-h	r-t,t-h	b-c,b-c	b-c,b-c	-c,b
$A\neg T\neg H$	r+b,-c-h	r+b,-c-h	r,-h	b-c,b-c	b-c,b-c	-c,b
$\neg A\neg H$	b,-c	b,-c	0,0	b,-c	b,-c	$\boxed{0,0}$

Table 9.26: Symmetric Extended Fragile Trust Game : A player who does not play A will also not have the choice of $T \vee \neg T$.

Extended Fragile Symmetric Trust Game with Sympathy

We now proceed to the extended fragile game, whose table we see in Table 9.26. If we examine when (ATH, ATH) yields a higher payoff than playing $A\neg T\neg H$ first and then being imitated, with players having sympathetic payoffs, then we should see this will play out exactly as before in the game without the thanking move. We thus have that for the imitation dynamics we will have $\delta_{IMI} > \frac{b+h}{r+c}$ when $s = 0$ and $\delta_{IMI}(s) > \frac{b+h-s(r+h+b+c)}{r+c-s(r+h+b+c)}$. Earlier this led to an impossible discount threshold to cross. Here we have the same.

The punishment dynamics for $s = 0$ gives us $\delta_{PUN} > \frac{b+h}{b+r+c}$ and $\delta_{PUN}(s) > \frac{b+h-s(r+b+h+c)}{b+r+c-s(r+h)-2s(b+c)}$. Note that both of these cases are identical to the ones before, as the symmetric game eliminates outcomes with thanking in these outcomes. Since the discounts are identical under the sympathetic parameters, we will not have to investigate the sympathy conditions implied by the discount.

Unified Symmetric Trust Game with Sympathy

Last, we consider the Unified Symmetric Trust Game, whose table we find in Table 9.27 and whose feasible region we see in Figure 9.29. Results follow at the

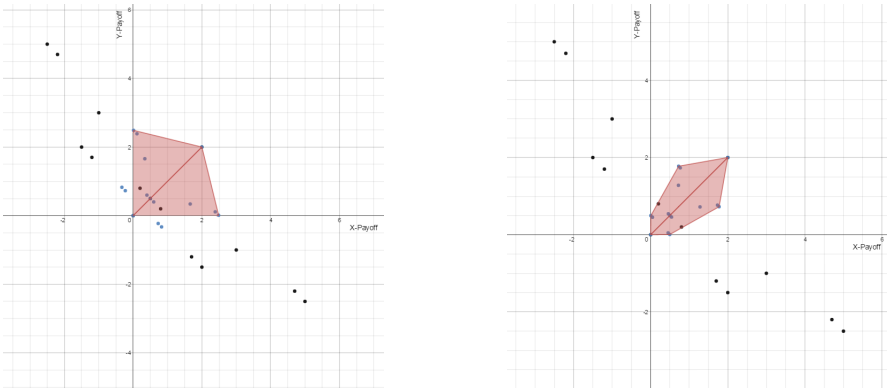

Figure 9.28: Feasible region for the extended fragile Symmetric Trust Game $r = 2$, $b = 3$, $c = 1$, $h = 1.5$, $t = .3$ and with sympathy $s = .33$ and $s = .43$ The sympathetic outcomes are in blue, while the originals are in black.

	ATH	$A\neg TH$	$\neg AH$	$AT\neg H$	$A\neg T\neg H$	$\neg A\neg H$
ATH	r-h,r-h	r-h-t,r+t-h	r-c-t,c+t-a-h	t-a-h,r-t	-a-h,r	-a-c,c-a
$A\neg TH$	r+t-h,r-h-t	r-h, r-h	r-c,c-a-h	t-a-h,r-t	-a-h,r	-a-c, c-a
$\neg AH$	c+t-a-h,r-c-t	c-a-h,r-c	-a,-a	c+t-a-h,r-c-t	c-a-h,r-c	-a,-a
$AT\neg H$	r-t, t-a-h	r-t,t-a-h	r-c-t, c+t-a-h	-a,-a	-a,-a	-a -c, c-a
$A\neg T\neg H$	r,-a-h	r,-a-h	r-c,c-a-h	-a,-a	-a,-a	-a-c, c-a
$\neg A\neg H$	c-a,-a-c	c-a,-a-c	-a,-a	c-a,-a-c	c-a,-a-c	-a,-a

Table 9.27: Unified Symmetric Trust Game: A player who does not play A will also not have the choice of $T \vee \neg T$.Nash Equilibrium of $(\neg A\neg H, \neg A\neg H)$

end. We also derive the sympathy conditions under which a discount value might promote cooperation.

These results are consistent with $s = 0$, as $\delta_{IMI}(s) > \frac{h}{a+r}$ and $\delta_{PUN}(s) > \frac{h}{a+c+r}$. Although it should be clear that we have $\delta_{IMI} > \delta_{PUN}$, how does sympathy affect the parameters? Under the parameters $r = 2$, $a = 3$, $c = 1$, $h = 1.5$, $t = .5$, as seen in Figure 9.29, we have the same result with sympathy, where $\delta_{IMI} = .30$, $\delta_{PUN} = .25$, and $\delta_{IMI} = .19$, $\delta_{PUN} = .16$ under $s = .10$. This holds for all values of s, and thus we have that $\delta_{IMI}(s) > \delta_{PUN}(s)$. We also see that when $s = .23$, as seen on the left in Figure 9.29, we have that the discount threshold hits zero, giving us a guarantee that players can be cooperative in the long run.

Results

This section featured two lines of inquiry: discount values required to sustain cooperation in repeated interaction, and sympathy constraints derived from those discounts. Two interesting patterns were that

1. The extended games had discount values independent of the t-value for thanking, and thus for repeated interaction resembled their simpler counterparts.

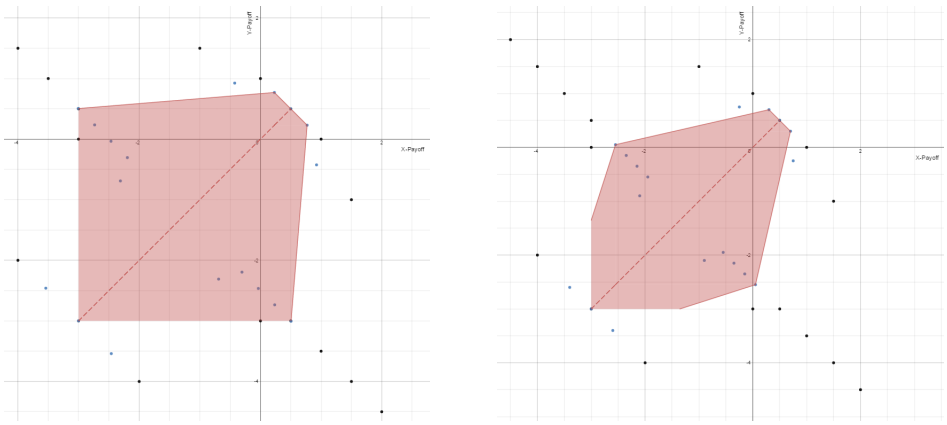

Figure 9.29: Feasible region for UnifiedSymmetric Trust Game $r = 2$, $a = 3$, $c = 1$, $h = 1.5$, $t = .5$; with sympathy $s = .23$ and $s = .3$ The sympathetic outcomes are in blue, while the originals are in black.

2. The constraints needed to promote the possibility of a viable discount value were often independent of the s-value of sympathy.

Discounts

Game vs. Strategy	Imitation	Punishment
Basic	$\delta_{IMI}(s) > \frac{b-r-s(b+c)}{r+c-s(b+c)}$	$\delta_{PUN}(s) > \frac{b-r-s(b+c)}{b+r+c-2s(b+c)}$
Fragile	$\delta_{IMI}(s) > \frac{b+h-s(r+h+b+c)}{r+c-s(r+h+b+c)}$	$\delta_{PUN}(s) > \frac{b+h-s(r+b+h+c)}{b+r+c-s(r+h)-2s(b+c)}$

Table 9.28: Discount Thresholds in Two-Stage Symmetric Trust Game with Sympathy: Discount threshold required to sustain the cooperative profile AH. The fragile trust game has an impossible discount threshold.

Game vs. Strategy	Imitation	Punishment
Basic	$\delta_{IMI}(s) > \frac{b-r-s(b+c)}{r+c-s(b+c)}$	$\delta_{PUN}(s) > \frac{b-r-s(b+c)}{b+r+c-2s(b+c)}$
Fragile	$\delta_{IMI}(s) > \frac{b+h-s(r+h+b+c)}{r+c-s(r+h+b+c)}$	$\delta_{PUN}(s) > \frac{b+h-s(r+b+h+c)}{b+r+c-s(r+h)-2s(b+c)}$
Unified	$\delta_{IMI}(s) > \frac{h-s(a+r+h)}{a+r-s(a+r+h)}$	$\delta_{PUN}(s) > \frac{h-s(a+r+h)}{a+c+r-s(a+r+h+2c)}$

Table 9.29: Discount Thresholds in Extended Symmetric Trust Game with Sympathy: Discount threshold required to sustain the cooperative profile ATH. The fragile trust game has an impossible discount threshold. The basic and fragile game have identical discounts to their simpler versions.

We see that in the basic and unified games, we have that cooperation can emerge without sympathy. In the fragile case this is not so, and it is only the case with punishment that we can overcome the negative payoffs that would lead to different outcomes.

Sympathy Values

Game vs. Strategy	Imitation	Punishment
Basic	$b < 2r + c$	$s < \frac{2r+c}{b+c}$
Fragile	$b + h < r + c$	$s < \frac{r+c-h}{b+c}$

Table 9.30: Sympathy Constraints in Two-Stage Symmetric Trust Game: In each entry, the sympathy required to promote the cooperative outcome is given. When the result is independent of sympathy, no s-value will appear..

Game vs.Strategy	Imitation	Punishment
Basic	$b < 2r + c$	$s < \frac{2r+c}{b+c}$
Fragile	$b + h < r + c$	$s < \frac{r+c-h}{b+c}$
Unified	$h < a + r$	$s < \frac{a+c+r-h}{2c}$

Table 9.31: Sympathy Constraints in Extended Symmetric Trust Game: When the result is independent of sympathy, no s-value will appear. The basic and fragile game have sympathy constraints equivalent to their simpler versions.

In contrast to our previous work on the asymmetric games, we saw that these games often had parameters that made the sympathy values irrelevant. This was often because the symmetric outcomes are not affected by sympathy, and thus the payoffs stay the same. Sympathy is not the only mechanism affecting decision rules however. We now move to a more general framework of tolerance for outcomes that fall within the prescribed boundaries of a relationship.

9.6 Relational Equilibria and Repeated Trust Games

In this section we apply the relational equilibria from Chapter 6 towards understanding decision rules and repeated game outcomes in the Symmetric Trust Game and their subsequent applications to face-threatening acts. We wish to build on the work in Grinberg et al. [2012], where the experimenters test human subjects in a repeated Prisoner's Dilemma with various decision heuristics. These heuristics were motivated by the relational models in Fiske [1992] of dominance, recirocity, communality, and value-matching. We have several goals here:

- Refine the mathematical sophistication and descriptive power of the social relations
- Give graphical interpretations of the boundaries of outcomes allowed under each type of relationship

- Analyze the payoffs in the repeatedSymmetric Trust Game according to the relational models.

- Examine the impact of sympathy on the feasible region of the repeated Symmetric Trust Game and the relational boundaries.

- Incorporate applications of *face* and *social distance*

In Grinberg et al. [2012], subjects in an experimental Prisoner's Dilemma made decisions using rules based on payoff allocations related to the social roles in Fiske [1992]. The payoffs to the agents were doled out according to the following rules:

- Total Group Payoff: Communality

- Unequal Shares: Dominance

- Equal Shares: Reciprocity

- Proportionality: Value-Matching

Thus agents in a communal relationship based their decisions on the group outcome $U_X + U_Y = n$; agents in a dominance hierarchy based decisions on a final split knowing $U_X > U_Y$, where X was more powerful than Y; and agents in a reciprocal relationship based their decisions on a final payoff where $U_X = U_Y$. Those in the value-matching relationship receive a fraction of their payoff kU_X or kU_Y. This last relationship type is strategically equivalent to the classical notion of utility. In contrast, we will focus on the first three types of relationships, remarked by Fiske [1992] as being more universal than value-matching, a relationship type based on cost-benefit analysis of interaction. Fiske claims this relationship type is only seen in modern societies, and, more to our point, it resembles the agent-based modeling of classical economics, on which we base this entire work.

As seen in Chapter 6, we expanded on this somewhat, giving these notions a more flexible apparatus. This should allow for mechanisms like *fuzzy mind*, forgiveness, or *trembling hand* to enter the picture [Fehr and Gachter, 2000, Nowak, 2006, Selten, 1983]. The overall picture is that agents in a certain type of relationship should stay within the bounds of the prescribed payoffs *on average*. This gives us a way to understand the prevalence of certain outcomes. This also allows us to incorporate an idea like *tolerance* or *satisficing*, where agents might choose actions or accept outcomes that outperform a certain threshold, even though they do not maximize utility Schwartz et al. [2002].

These next few sections expand the geometric approach seen in Chapter 6 towards identifying tolerable outcomes for agents in various relationship types. Each relationship type specifies a further partition of the feasible 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 a path towards an acceptable outcome in the feasible region of equilibria and 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.

We now discuss these relationships and the subsequent applications to the Symmetric Trust Game and examples of requests.

Amicable Equilibria

The relationship of communality is marked by an interest in achieving the common good [Kameda et al., 2005, Fiske and Tetlock, 1997], as seen in *The Theory of Moral Sentiments* [Smith, 2010, Gintis, 2005]. To that end, we have posited a model similar to the experimental design in Grinberg et al. [2012] that preserves outcomes in restricted subsets of the feasible region. The story doesn't end there however. As we claim that communality can vary across degrees, we would like to posit that some relationships may require more stringent outcomes for the group than others.

In accordance with satisficing, we now make a refinement of the amicable equilibrium seen before in Chapter 6. With this definition, we can make claims on the extent to which a decision rule accounting for group benefit will promote cooperative 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.*

Here we can see that the presence of amicable equilibria rule out various outcomes that would not make sense under a relationship of communality, where two agents would be acting in the group's interest[Grinberg et al., 2012, Sally, 2000]. This might allow us to rule out outcomes that were equilibria in the one-shot game if it were not the case that $U_X + U_Y \geq k$. More importantly, this might also allow us to rule out large sections of the feasible region in the analysis of the repeated game. This is crucial as much of the equilibrium analysis has provided reasons for why a certain outcome might be stable but not reasons why that outcome would be preferred/ dispreferred. As we base this definition on a degree of communality or a tolerance for acceptable outcomes, the flexibility of agents towards eliminating certain outcomes may depend on their initial tolerance.

We can also see that off-diagonal outcomes in a symmetric game must lie along the same line of sympathetic payoffs, as we see in **??**. These *sympathetic frontiers* can rule out certain pure strategy equilibria in the Symmetric Trust Game and its variants. This means that agents in a communal relationship would want to outperform outcomes along these lines in some cases, depending on their interest in the group's outcome [Sally, 2002, Álvarez and Hurtado, 2012].

Do we want sympathetic frontiers that will rule out potentially exploitative behavior? In fact, no. Rather we want lines that rule out mutually destructive behavior. In our case, we can consider the outcomes that are the worst in group performance, and a sufficiently great sympathetic frontier of $U_X + U_Y = k$ would rule them out. We might also see that low social distance should protect the agents from these outcomes, as close friends might not accept outcomes that are detrimental to them as a whole. The connection between sympathy and social distance should therefore be explored.

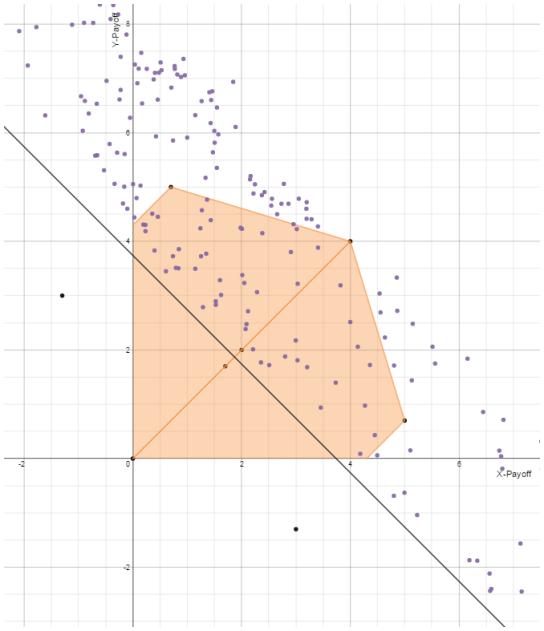

Figure 9.30: Amicable Outcomes in the Symmetric Trust Game: Here we amicable outcomes for a boundary of $k = 3.75$. Only those within the feasible region would be *amicable equilibria*.

Reciprocity, Fairness, and Egalitarian Equilibria

Fairness means balancing the differences in outcomes [Fehr and Schmidt, 2001, 1998, Tabibnia et al., 2008, Tabibnia and Lieberman, 2007] As outcomes improve, this preference should diminish in absolute terms in accordance with the law of diminishing returns [Camerer, 1997, Rabin, 1993, Eika, 2000] This is why we normalize by the sum of the utilities. Although communality might rule out some highly unequal outcomes, we might see relationships in the case of reciprocity to include these outcomes as a credible threat. In Chapter 6, we gave a metric for the tolerance of unequal outcomes as the difference in outcomes compared to the total utility. This gave us a measure of inequality:

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

We used this measure to define a partition of the feasible region of equilibria according to the folk theorem. Provided outcomes are sufficiently close to one another in comparison to the overall group outcome, a fairness tolerance should mean that agents would rule out the outcomes that were outside of these boundaries.

Definition 2. *For a given inequality tolerance T, we say an outcome in a repeated game is a T-**egalitarian outcome** if the respective payoff profile (U_X, U_Y) satisfies $\frac{|U_X - U_Y|}{U_X + U_Y} \leq T$. A T-**egalitarian equilbrium** is an egalitarian outcome in the feasible region of repeated game payoffs subject to the folk theorem.*

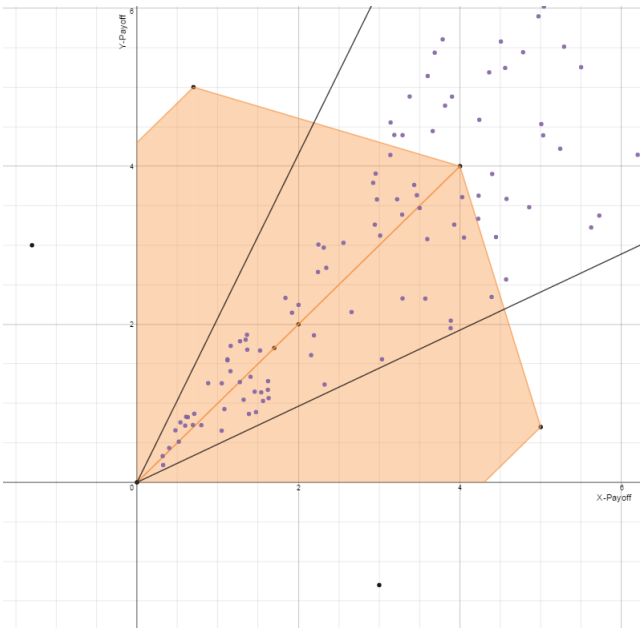

Figure 9.31: Egalitarian Outcomes in the Symmetric Trust Game: Here we see egalitarian outcomes for an inequality tolerance of $T = .35$. Only those within the feasible region would be *egalitarian equilibria*.

When we have a sympathetic preference like sympathy, there is a difference between the concrete payoff and the perception of that payoff in a strategy profile. In the case of a reciprocal relationship, this can explain why an outcome that may seem unfair to an observer may in fact be perceived as fair by the participants, as seen in Figure 9.32. This can also give us a rationale for why seeming face-loss may not be perceived as such when occurring between partners with high degrees of sympathy towards each other.

Sympathy and Fairness

Some interactants may be willing to overlook perceived slights if they feel a great deal of sympathy for each other. This could mean that an outcome that is *prima facie* unfair might not be judged as such. Is this compatible with our framework? Yes! Consider that the feasible region of equilbria under sympathy brings asymmetric outcomes more closely together, as seen in Figure 9.32. E.g. if Y benefits

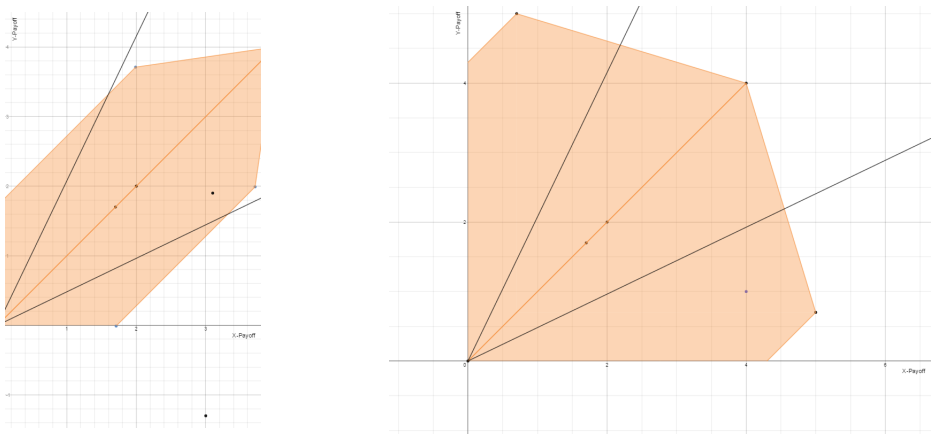

Figure 9.32: Perceived Outcomes (Left) vs. Concrete Outcomes(Right): Here we see the feasible region of egalitarian equilibria for the Symmetric Trust Game with $r = 2$, $b = 3$, $c = 1$, $h = 1.5$, $t = .3$ and with a tolerance of $k = .3$. Notice on the left that we see the region as dictated by $s = .3$, whereas the right diagram depicts the actual space of outcomes. The unfair outcome is $(4, 1)$ but under sympathy $s = .3$ it becomes $(3.1, 1.9)$.

at the expense of X, a sympathetic X will judge this more favorably. Moreover, a sympathetic Y will not feel as good about the imbalance in outcomes. In terms of requests, an example of might be an office romance, where two individuals nominally in a reciprocity relationship might go out of their way to help each other. A one-sided example would also be plausible under this scenario, where one agent might say *I'd be happy to help.*

On the other hand, if there is a low degree of sympathy between two agents in a repeated reciprocity relationship, an agent requiring an outcome that is not egalitarian may call attention to this, so as to make clear the fiuture payment of a debt. This may occur in a situation like *I know this is inconvenient but could you help me with this?*. The speaker is calling attention to the fact that the payoff is inherently negative for the hearer should he comply. To remedy this, he might say *Thanks! I owe you one!*

Dominance and Despotic Equilibria

Dominance relationships stand in sharp contrasts to reciprocal ones. Arguably, they form a mutually exclusive subset of the action space. This is because a dominance relationship should prototypically serve the one with more power. We thus turn to the last type of outcome and equilibrium: despotic. Recall from our work on preferences that we had:

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.*

The gist is that in a repeated interaction between two individuals of unequal power, we expect that the overall pattern of behavior, in this case the utilities derived from infinite horizon repetition with discounting, will favor the individual with more power. In particular, if the agents X and Y have relative power in the ratio $A : B$ where $A + B = 1$ and $A > B$, we should see that the outcomes respect the inequality $\frac{U_Y}{U_X} < \frac{B}{A}$. This gives us the inequality from the defiintion and provides a simple way to graph this, as seen in Figure 9.33. As the outcomes will fall towards the agent with more power, we see the opposite case as well, where $\frac{U_Y}{U_X} > \frac{B}{A}$.

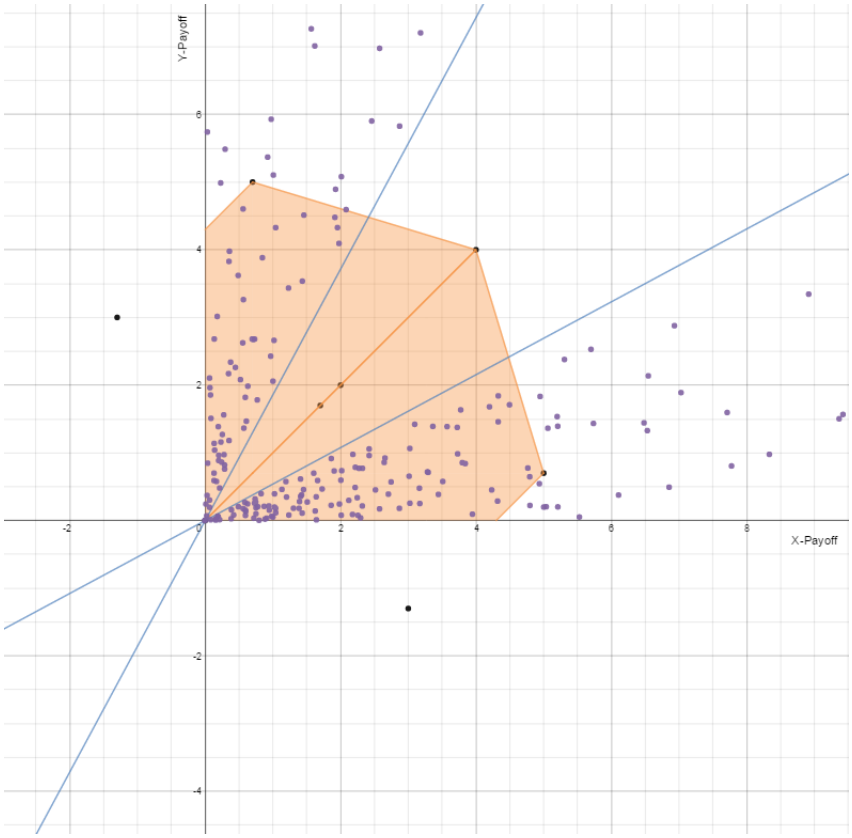

Figure 9.33: Despotic Outcomes in the Symmetric Trust Game: Here we see despotic outcomes for power imbalance of A:B. Only those within the feasible region would be *despotic equilibria*. These two sets of outcomes reflect distinct events, where in one case X is more powerful and in another where Y is more powerful.

Sympathy and Dominance

To follow up on the previous discussion of fairness, we can sometimes think of dominance as a case of too much sympathy. This might occur between a parent and a child, where the parent with the high degree of sympathy perceives constantly helping the child as a fair outcome. Figure 9.32 can then be seen through the observer's eyes, and we have that the evaluation of the agents is in fact different from the reality.

We might also have the case that one agent, through the filter of sympathy, perceives the relationship to be one of fairness. This agent will tolerate outcomes that are inherently unfair, leading an unbiased observer to note that the relationship is in fact one of dominance. This can occur, for instance, when one agents admits a gross inability to accomplish a certain task without help by evoking sympathy. For example, *I know I can never repay you, but could you please give me a dollar so I can feed my family?*

If a superior does not feel strong sympathy towards a subordinate, the subordinate may need to invoke a personal feeling of displeasure with troubling the superior, so as to maintain the claim that the act of the request itself is one of negative utility for the speaker, e.g. *I hate to ask, but can you sign this form?*.

Fairness vs. Dominance

One natural claim to make by inspecting the figures in Figure 9.31 and Figure 9.33 is that these types of outcomes seem mutually exclusive. We can thus compare the two expressions:

$$\frac{|U_X - U_Y|}{U_X + U_Y} = k$$

and

$$\frac{U_X}{U_Y} = \frac{A}{B}$$

Solving the first for U_Y and assuming $U_X > U_Y$, we have

$$\begin{aligned} -\frac{k-1}{k+1} &= \frac{B}{A} = \frac{1-A}{A} \\ \Rightarrow k &= 2A - 1 = A - (1 - A) \end{aligned}$$

Looking back at this and noting that $A + B = 1$, we should not be surprised that the slope of the boundary line for dominance $m = \frac{B}{A}$ is related to the tolerated difference in outcome between the two players, as predicted. Conversely, we have that the tolerated, normalized difference in outcomes $\frac{|U_X - U_Y|}{U_X + U_Y}$ should be equivalent to the relative power imbalance $A - (1 - A)$. What does this mean in terms of deriving the boundaries for those in a dominance relationship? We simply compare their relative power normalized. Notice this gives us exactly one of the criteria for determining face threats in the equation posited by Brown and Levinson [1987] as $P(H, S)$, the power of the hearer over the speaker.

How does the relative imbalance of power affect a relationship of communality? Notice that in contrast to the reciprocity relationship, communality is open to imbalances of power, so long as they favor the group. This is why a communal relationship can include outcomes that might also fall under a dominance or fairness relationship.

Social Distance and Sympathy

How can we quantify the relationship between social distance among agents and how they would implement decision procedures among each other? We want it to be the case that agents unknown to each other, or whom we cannot connect in a social network, should have effectively an unreachable distance between them. We also want this to correspond to no sympathy between them. On the other hand, we want it to be the case that an agent who cares about another agent entirely must in fact be treating them like himself, with an effective zero social distance. One option is therefore to let:

$$d = \frac{1}{s} - 1 = \frac{1-s}{s}$$

This allows us to compare my self-regard $1-s$ to my sympathy s for the other agent. At a level of no sympathy, I should treat this person like a stranger, for whom I have an immeasurable social distance, or $d = \infty$. In the case of a close friend, my sympathy toward them should be higher. And in the case of playing against myself, a sympathy value of 1 will give us a social distance of zero.

One advantage of this definition is that it is bidirectional. This means that we can think of sympathy as a function of social distance. Given a certain social distance between two agents in a network, we could use this to predict a sympathy component to their decisions. We are open to a revision of this model, as we might also have $d = c\frac{1-s}{s}$ for some proportionality constant c. For instance, this constant might involve the number of agents in a social network. We thus have a connection between sympathy and the second determiner of face threats in Brown and Levinson [1987], social distance.

As mentioned before, sympathy changes the *perception* of the agents as to how they evaluate the various outcomes, as seen in Figure 9.32. For example, if $s_X = .3$ and $V_X(U_X, U_Y) = (1-s)(U_X) + s(U_Y)$, then $V_X(1,4) = .7 + 1.2 = 1.9$ and $V_X(2,1) = 1.4 + .3 = 1.7$. This case would give us that a player X with this level of sympathy would prefer the first outcome over the second. We can even notice that in absolute terms, the second outcome is more "fair" according to the metric as seen in Figure 9.32.

Could it be the case that social distance and sympathy could vary across the relationship types? We claim yes, as one might feel a high degree of sympathy for a subordinate or a low degree of sympathy for a colleague. Further, the difference between close friends and acquaintances is not so much in the type of relationship as the distance and sympathy between the two. Interestingly, inattentiveness to social distance might also lead a superior to lose face, as a high level of sympathy for a subordinate might make him act in a way that overly benefits his subordinate to his own detriment. This would give the appearance of an inappropriate level of social distance [Goffman, 1967].

Face and Place

One component of *face* is doing what is expected, given the social role in a relationship. [Strebel, 1996, Goffman, 1967, Heffetz and Frank, 2008]. We can then think of *negative face* as the probability that one can maximize one's own utility and *positive face* as the probability that others will show sympathy. What this means for positive face concerns is that agents in a reciprocal relationship should not choose outcomes that favor one tremendously over the other, agents in a communal relationship should not choose outcomes that make the group substantially worse off, and agents in a dominance relationship should not choose outcomes that harm the superior. Geometrically, we can use this to identify points in the feasible region that agents would avoid in a one-shot interaction. We can extend this to note that sufficiently patient agents in a repeated interaction might tolerate these outcomes, with the provision that the imbalance later be corrected.

If we call attention to an outcome that crosses the lines of the prescribed relationship, we may initiate repeated interaction to avoid losing face. Notice this is especially crucial when initiating an action that is an immediate face threat to the other person, like a request. If the speaker is in a reciprocal relationship, this may be one that steps outside of that tolerance. If the speaker is subordinate in a dominance relationship, this may be one that overly favors the speaker over the hearer. Notice however in a communal relationship that the gravity of a request is not as heavy, for the need to correct an outcome that favors at least one member does not exist. Consider the interchange below:

- **Xavier:** Can you get me a drink?
- **Yves:** Sure!
- **Outcome:** $(+1, -1)$

In a communal relationship, there would not be need for further redress of the situation, although we might expect the interchange to continue:

- **Xavier:** Hey thanks man! Can I get the next round?
- **Yves:** Thanks buddy, but I'm not keeping score. It will even out at the end of the night.
- **Outcome:** $(-1, +1) \Rightarrow (0, 0)$

Were we in a reciprocal relationship, the interchange might go like this:

- **Xavier:** Do you think you can get the drinks and I can get the food?
- **Yves:** OK, that makes sense.
- **Outcome:** $(0, 0)$

or

- **Xavier:** Oh no! I forgot to visit the bank today. Do you mind helping me with my bill? I can pay you back tomorrow.
- **Yves:** Sure that works.
- **Outcome:** $(+1, -1)$

Notice that in the second example we have a redress of face concerns coupled with an invitation to repeated interaction towards balancing the outcome. In a dominance relation, we might see something like:

- **Xavier:** Mr. Smith, I hate to bother you, but I just realized I forgot my wallet. Is there a way you could pick up my drink?
- **Yves:** I suppose so. Just pay me back tomorrow buddy!
- **Outcome:** $(+1, -1)$

This might be resolved as:

- **Xavier:** Mr. Smith, sorry to interrupt you, but I wanted to thank you for last night. Here's $5 for the drink.
- **Yves:** Certainly.
- **Outcome:** $(0, 0)$

This interaction could be more nuanced depending on the situation, for in the dominance relation the superior has a chance to acquire positive face, whether indirectly by reputation or directly by refusing repayment of the debt. We might also see the superior use positive politeness towards the subordinate, as in *No problem buddy.*

9.7 Conclusion

This chapter has traced a variety of models that approximate the dynamics behind requests. It has explored symmetric and asymmetrtic trust games, altered their payoff structures, and examined these findings under mechanisms like sympathy and repetition. It has further given more flexible and quantifiable measures of sympathy, social distance, and fairness than seen in Brown and Levinson [1987].

Several patterns have emerged. They are:

- Cooperation may not be possible without at least one of repetition or sympathy. Their combination typically leads to higher likelihood cooperation through an inverse relationship between patience required δ and the sympathy the agents have towards each other.
- In cases where cooperation is difficult to maintain, punishment may be required. Without the threat of punishment, cooperation may be unsustainable.

- Decision heuristics based on social relations and tolerance of outcomes allow us to rule out strategic profiles that would be admissible under the folk theorem. This gives natural predictions about tolerance for seeming face violations.

Bibliography

Andrés Álvarez and Jimena Hurtado. 'out of sight, out of mind': Modern economics, social interactions, and smith's sympathy. Documento CEDE, (2012-01), 2012.

Nicholas Asher and Alex Lascarides. Questions in dialogue. Linguistics and philosophy, 21(3):237–309, 1998.

Nicholas Asher and Alex Lascarides. Indirect speech acts. Synthese, 128(1-2): 183–228, 2001.

Monica Y Bartlett and David DeSteno. Gratitude and prosocial behavior helping when it costs you. Psychological science, 17(4):319–325, 2006.

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

Penelope Brown and Stephen C. Levinson. Politeness: Some universals in language use. In E.N. Goody, editor, Questions and Politeness, Cambridge papers in social anthropology; no. 8, pages 56–310. Cambridge University Press, New York, 1978.

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.

Colin F Camerer. Progress in behavioral game theory. The Journal of Economic Perspectives, pages 167–188, 1997.

Kari H Eika. Economic man and his social preferences. 2000.

Ernst Fehr. Social preferences and the brain. In Paul W. Glimcher, Ernst Fehr, Colin Camerer, and Russell Alan Poldrack, editors, Neuroeconomics: Decision Making and the Brain, chapter 15. Academic Press, 2008.

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`.

Ernst Fehr and Klaus M. Schmidt. Theories of fairness and reciprocity - evidence and economic applications. CEPR Discussion Papers 2703, C.E.P.R. Discussion Papers, February 2001. URL http://ideas.repec.org/p/cpr/ceprdp/2703.html.

Alan Page Fiske and Philip E Tetlock. Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. Political psychology, pages 255–297, 1997.

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

H. Gintis. Game theory evolving: A problem-centered introduction to modeling strategic interaction. Princeton University Press, 2000.

Herbert Gintis. Moral sentiments and material interests: The foundations of cooperation in economic life, volume 6. MIT press, 2005.

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

Adam M Grant and Francesca Gino. A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. Journal of personality and social psychology, 98(6):946, 2010.

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

Ori Heffetz and Robert H Frank. Preferences for status: Evidence and economic implications. HANDBOOK OF SOCIAL ECONOMICS, Jess Benhabib, Alberto Bisin, Matthew Jackson, eds, 1:69–91, 2008.

Tatsuya Kameda, Masanori Takezawa, and Reid Hastie. Where do social norms come from? the example of communal sharing. Current Directions in Psychological Science, 14(6):331–334, 2005.

David K. Levine. Modeling altruism and spitefulness in experiment. Review of Economic Dynamics, 1(3):593–622, July 1998. URL http://ideas.repec.org/a/red/issued/v1y1998i3p593-622.html.

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.

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.

Barry Schwartz, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, and Darrin R Lehman. Maximizing versus satisficing: happiness is a matter of choice. Journal of personality and social psychology, 83(5):1178, 2002.

Reinhard Selten. Evolutionary stability in extensive two-person games. Mathematical Social Sciences, 5(3):269–363, 1983.

Adam Smith. The theory of moral sentiments. Penguin, 2010.

Thomas Sowell. Property rights. Townhall. com. http://www. townhall. com/columnists/thomassowell/ts20010809. shtml (accessed December 23, 2003), 2001.

Helen Spencer-Oatey. Culturally speaking: Culture, communication and politeness theory. Continuum International Publishing Group, 2008.

Paul Strebel. Why do employees resist change? Harvard business review, 74(3): 86, 1996.

Golnaz Tabibnia and Matthew D Lieberman. Fairness and cooperation are rewarding. Annals of the New York Academy of Sciences, 1118(1):90–101, 2007.

Golnaz Tabibnia, Ajay B Satpute, and Matthew D Lieberman. The sunny side of fairness preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychological Science, 19(4):339–347, 2008.

Jo-Ann Tsang. The effects of helper intention on gratitude and indebtedness. Motivation and Emotion, 30(3):198–204, 2006.