Beliefs, Relationships, and Equality: An Alternative Source of Discrimination in a Symmetric Hiring Market via Threats

Serafina Kamp, Therese Nkeng, Vicente Riquelme, and Benjamin Fish

University of Michigan, Ann Arbor, MI, USA

Abstract. Machine learning has grown in popularity to help assign resources and make decisions about users, which can result in discrimination. This includes hiring markets, where employers have increasingly been interested in using automated tools to help hire candidates. In response, there has been significant effort in attempting to understand and mitigate the sources of discrimination in these tools. However, previous work has largely assumed that discrimination is the result of some initial *unequal distribution of resources* across groups: one group is on average less qualified, there is less training data for one group, or the classifier is less accurate on one group, etc. Recent work on relational equality have suggested that there are other sources of discrimination that are non-distributional, namely inequality in social relationships. Here, we demonstrate how discrimination can arise from a non-distributional source: We provide subgame perfect equilibria in a simple sequential model of a hiring market with Rubinstein-style bargaining between firms and candidates that exhibits discriminatory outcomes, yet there was no initial unequal distribution of resources across groups of candidates or firms. We provide the range of possible expected payoffs to the firms and candidates at equilibrium, including asymmetric payoffs where some candidates receive less. This is the result of asymmetric strategies where firms successfully take advantage of those candidates, resulting in discrimination. Thus, we show that we must look beyond the distribution of resources to understand sources of discrimination in machine learning.

Keywords: Algorithmic Fairness · Machine Learning · Hiring Markets · Relational Equality · Computers and Society

1 Introduction

Machine learning (ML) algorithms are marketed to make more efficient and data-driven decisions that improve human decision-making. These automated tools have become increasingly popular in recommendation systems, classification problems, and resource assignment amongst other areas [7, 10, 30]. With their growing popularity, it has become clear that these models can engender discrimination against members of certain demographic groups [4, 18, 38, 41].

Such discrimination can exist in various forms and types, but what they all have in common is the harm that affects individuals and the inequities they perpetuate. Generally, this discrimination is evidenced by the unfair distribution of resources, such as skill sets, data, and predicted labels [5, 6, 23, 29]. For example, unequal amounts of training data across demographic groups can then result in unequal accuracy. When different groups have different knowledge about strategic actions to take to ensure a positive label, this can also result in unequal accuracy. Moreover, models with initial resource inequalities will generate feedback loops that recreate and amplify the disparities in resources and outcomes [7, 35, 41]. The causality, nature, and degree of this discrimination varies, leading to a rich research area on formalizing, measuring, and mitigating bias present in socio-technical systems [7, 12, 22, 37].

To prove the existence of such discrimination, one must measure and define differences across groups, individuals, experiences, or outcomes [6, 29, 37]. The measurements of fairness that researchers have developed span across ML and economics including group, individual, and causal fairness alongside taste-based and statistical discrimination [5, 8, 33, 44]. These works have focused on equalizing accuracy, error, or outcomes between groups defined by sensitive attributes such as race or gender [25, 27, 43, 48]. For example, *equalized odds* [23] requires equal true and false positive rates across groups. Enforcing this constraint can be used to correct situations where, for example, the distribution over training examples is different across groups (e.g. there are more examples for the majority group than a minority group), and the classifier learns to favor the majority group. Such results have been seen in many domains like resume screening, criminal justice, healthcare, hiring markets, and more [13, 17, 21, 32, 40, 46]. However, this implies that if the distributions of the groups are the same, the optimal unconstrained classifier will not lead to discrimination, at least as defined by equalized odds, and adding in the equalized odds constraint won't help. In this work, we show the possibility of discriminatory outcomes even when there are *no* distributional differences across individuals, let alone groups. Here, we show there is work to do *beyond correcting training distribution inequities* in ensuring fairness in machine learning.

In economics, notions of fairness also present similar limitations in addressing discrimination. Statistical discrimination, for example, is a model of discrimination in markets that shows how it can be economically "rational" to use sensitive attributes as a noisy signal of underlying worker productivity. It would then follow that reducing such friction by equalizing worker skill sets would eliminate discrimination by eliminating those cases where sensitive attributes are a useful signal for productivity. However here, building on previous work [21], we show discrimination can still occur when both signals for productivity and underlying productivity are equal across all workers, eliminating statistical discrimination as a source. We show *equalizing skill sets among workers* is not enough to prevent discrimination from occurring.

The above examples reveal that these fairness metrics rely on one key normative assumption – that the source of discrimination is due to an unequal

distribution of resources. This view of equality is commonly known as distributional equality [15]. However, there are other views, such as relational equality [3], that highlight other aspects of equality. Relational equality concerns itself with flattening social hierarchies and ensuring equal social relationships throughout society. Notably, the quality of social relationships cannot be reduced to the distribution of resources, though they may be related [49]. Given these alternative viewpoints on how to ensure equality [3, 9], we should expect to see examples of discrimination which have not been caused by the unequal distribution of resources. Existing fairness metrics in ML only help when there is an unequal distribution of resources of some kind, and so to prevent discrimination in ML and other algorithmic decision making, we first need to identify alternative causes of discrimination.

In this work, our main contribution is to provide such an example where discrimination is possible even though all agents are identical to each other. Our example is a bargaining game, representing a simple hiring market. The agents in our market include firms looking to fill a job position and candidates applying for these positions. Once a firm and candidate are matched (the candidate applies for the firm's position), they participate in wage negotiation: rounds of bargaining to determine how to split the surplus that will be generated as a result of the candidate's employment. Since there are many firms and candidates in our market, every agent may choose to end the bargaining at any point without agreement, and take an endogenously determined outside option, i.e. they return to the pool and get re-matched with another agent. We ensure that our market is initially symmetric in terms of resources among all agents, which will be explained when we formally introduce our market in Section 3.

Hiring markets give us a useful arena to showcase unfairness among similar groups of people and is an area where major ML investments are being made [13, 21, 26], though our results hold for bargaining games in other domains. Specifically, we show that there exists subgame perfect equilibria (SPE) in the bargaining game with discriminatory outcomes. A set of strategies is in SPE when there are no beneficial deviations from the strategies on or off the equilibrium path [51][1]. This means that utility-maximizing agents can have incentives to collectively discriminate, regardless of whether those agents use automated decision making systems or not. Our results in Section 4 give two kinds of discriminatory outcomes at equilibrium: One where a group of candidates gets more of the surplus than another group of candidates and another where the firms get more of the surplus than the candidates. We show that the source of discrimination *is not distributional* and instead comes from (correct) beliefs about an agent's ability to make threats during bargaining. This builds upon previous work [21] which showed how beliefs of agents in a market could be the source of discrimination, although in that work beliefs only needed to be correct on the equilibrium path.

In Section 2 we review related works on fairness metrics, bargaining and markets, and non-distributional notions of equality. In Section 3, we introduce our bargaining game. Section 4 describes equilibria in this game and the range

[1] See [51] for more on the game theoretic concepts used in this work.

of payoffs possible to each agent. We conclude in this section by highlighting the need for non-distributional notions of equality to improve our understanding of discrimination.

2 Related Work

2.1 Fairness Metrics in Machine Learning

Much previous work has been done to conceptualize and create measurements of fairness in algorithmic decision making, and machine learning in particular. Many metrics, both statistical and causal, have been proposed to define and measure unfairness across individuals and groups in classifiers and beyond [14, 23, 33, 37, 53]. This includes equalized odds, which as mentioned above, implicitly makes the assumption that the source of discrimination is from resource distribution, like enough training data for each group. The assumption that the source of discrimination is in the distribution of training examples underpins current fair machine learning metrics [21], not just in statistical metrics like equalized odds, but also causal ones like counterfactual fairness [33]. If the distribution over features does not change across counterfactual values of a sensitive attribute, then ensuring counterfactual fairness would not necessarily address any discrimination that results. While these metrics associate fairness with equalizing accuracy or the like, many works recognize that there may be differences between groups due to *historical bias* that lead to discriminatory outcomes even when an algorithm is accurate with respect to its training data [24, 37, 47]. Specifically, feedback loops are well understood in this context where machine learning algorithms amplify initial resource allocation differences that exist historically [1, 36]. However, mitigating historical biases are generally context-dependent and not well understood [6, 37, 46]. Historical bias could overlap with the non-distributional discrimination that we aim to study: We are interested in discriminatory outcomes that are not the result of initial asymmetric resource distribution and some historic biases may be the result of this phenomena. However, importantly, the discrimination we are interested in cannot be reduced to historical bias since we will show that this discrimination can arise at any time from symmetric resource distribution settings.

2.2 Discrimination Analysis in Bargaining Games

As discussed in the introduction, we build off of Fish and Stark's [21] work on non-distributional sources of discrimination in a bargaining game, but provide a different mechanism: Rather than beliefs about outside options alone, we also look at an agent's ability to make threats during bargaining. Because we are using SPEs, agents can only make credible threats such that their beliefs about other agents' outside options must be correct at equilibrium. That work used a weaker notion of equilibrium and did not have fully symmetric and identical agents.

There are several similar works to ours which highlight undesirable outcomes of bargaining models. Their motivations and outcomes vary, but this includes work that showcases how rational agents with complete (symmetric) information could be Pareto-inefficient where both parties would be better off in a different equilibrium [20, 34, 52]. There is also work that analyzes and aims to correct the disparity in wages across workers [11, 19, 26], focused largely on statistical discrimination, which features distributional inequality in the form of disparity in skills across workers at equilibrium. Rather than seeking to correct statistical discrimination, we show that discriminatory outcomes still arise, *even when skill level is assumed to be the same between workers.* Thus, addressing statistical discrimination is insufficient to preventing discriminatory outcomes in hiring markets.

There are several works that consider bargaining games with multiple equilibria [28, 45, 50], as we do here. Models that result in multiple equilibria are useful for demonstrating the existence of discrimination, because it enables the possibility of both an unfair outcome and a better alternative that's still incentive-compatible. However, these works all have some asymmetry between agents including exogenous costs, who gets to propose first, and outside option values [28, 45, 50]. Our work directly extends the scenario modeled by Ponsati and Sákovics [45], which features a two-player bargaining game with exogenous outside options. We extend their model by endogenously modeling an agent's outside option, allowing for more than two agents in our market, and most importantly, imposing symmetry between the bargaining agents. We show in our extended model that a range of payoffs at SPE still exists. Similarly to us, Agranov et al. [2] demonstrate the existence of asymmetric outcomes at SPE with initially symmetric agents. They do so in a multi-lateral budget allocation setting, but it is a setting where no exit or outside options are permitted. Thus the equilibria leave all agents powerless to negotiate by conditioning their strategies on any past deviations from equilibrium. In our setting, agents are allowed to retain power, at least a priori, through the existence of outside options. Moreover, the equilibrium strategies we find are "match-stationary". That is, each agent's strategy for the current bargaining match does not condition on agent behavior, including deviations, from previous bargaining matches.

2.3 Critiques of Distributional Equality

There are a few works that consider alternative sources and presentations of discrimination beyond resource distribution. Several works in political philosophy discuss the shortcomings of a purely distributional view of equality and identify relational equality as a way to address the gaps [3, 16, 49]. Similarly, we show that a purely distributional view of fairness, like current fairness metrics support, misses instances of discrimination that theories of relational equality may better explain. In the literature on machine learning, Birhane [9] draws attention to the complexities of fairness questions and discusses how relational ethics might be a useful framework. Kasy and Abebe [31] similarly recognize the failures of current fairness metrics to capture certain forms of discrimination and they model the

ways in which power affects the outcome of an algorithm. Neuhäuser et al. [39] investigate the effects of behavioral interventions (such as the homophily preferences for nodes in the network) during the growth of networks on the visibility of minority populations. However, these works are not focused on identifying non-distributional sources of inequality, particularly in markets.

3 The Market

3.1 The Model

Our market is an extensive-form game [51] where candidates and firms are matched and take turns proposing and responding to surplus splits in a bargaining game. We make the normative assumption that all firms and candidates are equally skilled and entitled to the surplus. This assumption highlights that there exist equilibria where candidates and firms may play different strategies for reasons that are not merit based and that result in unequal payoffs. We say two agents are the same "type" if they play the same strategy. In the case of two types of candidates and one type of firm, we use p to denote the probability of a firm matching with one type of candidate. For convenience, we assume our market remains a constant size with an equal number of firms and candidates at all time steps and that the composition of our market remains unchanged such that p does not change over time.

Each agent chooses a profile of behavioral strategies before they enter the market, consisting of the actions an agent would take when they are the proposer and responder – including their opt out strategies. The strategy profiles are also known as a finite state machines in our infinitely repeated game [42]. Their specific strategies will be explained in more detail in Section 4 and we will show that the strategy profile is in SPE.

In this market, there are discrete time steps and during each time step there is a matching and a bargaining phase. During the matching phase, unmatched firms and candidates are matched using a particular process. We do not assume any particular mechanism for matching, but one example of the matching process could be that firms and agents are matched with some α i.i.d. probability such that all agents are expected to be matched in $\frac{1}{\alpha}$ time steps.

Once two agents are matched, they enter the bargaining phase. During the bargaining phase, firms and candidates participate in an extensive form Rubinstein-style bargaining game [42] to determine the split of the surplus, which we normalize to 1. Both the firm and the candidate will have an equal probability of proposing first during each bargaining game that occurs within our market. Both agents also have the ability to opt out during bargaining and match with a new bargaining partner. These features are important because otherwise the agent that gets to propose first has an advantage over the other agent [51] and having an outside option gives power to both agents. Note that our game is complete information and is different from classic Rubinstein bargaining because we include outside options for both agents as in [45]. So, we have now enforced that

our market is initially symmetric because no agent is more entitled to the surplus than another, no agent has a proposal advantage, each agent has an outside option, and our game is complete information.

In the bargaining phase of a single game where agent i is the first proposer and agent j the responder, agent i will propose a split of the surplus $(1-y, y)$ for some $y \in [0, 1]$. As the responder, agent j can either reject or accept the proposal. If they accept, they both leave the market and agent i and j receive a utility of $1 - y$ and y, respectively. Two new agents (a firm and a candidate with the same strategies as those leaving) would then enter the market in the next time step to keep our market size and composition constant. If the responder rejects, then either agent can decide to opt out of the negotiation. As in Ponsati and Sákovics [45], the order in which the agents opt out will not affect the equilibria, so we assume they decide simultaneously. When either agent chooses to opt out, both agents must pay a cost of $0 \leq \tau \leq 1$ and are sent back into the market pool to be re-matched in some future time. Here, τ corresponds to the waiting and matching cost incurred when an agent decides to re-enter the market. Dissimilar to Arganov et al. [2], their next bargaining partner will not condition on the actions in their previous match (such as deviations from their strategy). Finally, if both agents decide not to opt out, negotiations continue in the next time step where the current responder becomes the new proposer and the current proposer becomes the new responder. After each time step during bargaining, a discount factor of $0 \leq \delta \leq 1$ is applied to the payoff of all agents. That is, if two agents i and j have been bargaining for $t + 1$ time steps and agree on a surplus split of $(1 - y, y)$, agent i gets $\delta^t(1 - y)$ and agent j gets $\delta^t y$. Note that δ is only applied within a bargaining match, so agents that enter (or re-enter) the market at a later time step only incur δ when they begin bargaining. This is because τ accounts for both the stochastic and costly process of applying for (or trying to fill) a job and the variable amount of time it can take to get to the wage negotiation stage. The market timeline as well as the specifics of the bargaining phase can be visualized in Figure 1.

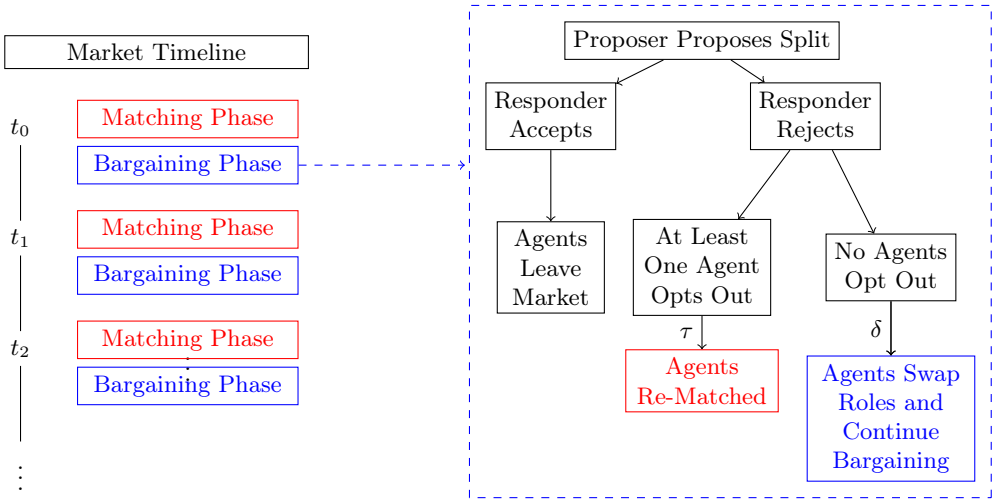
During any given bargaining game each agent has an expected payoff of that game which we will call W_{ij} for agent i when playing agent j. Note that, $W_{ji} = 1 - W_{ij}$ since they are bargaining over a surplus normalized to 1. We will use U_i as the expected outside option for agent i, interpreted as the expected payoff for i at the point in time they opt out of their current negotiation. We only consider strategies where W_{ij}, W_{ji}, and U_i do not depend on time.

To compute the outside option of an agent (U_i), we will need τ and the expected payoff to agent i across all possible agents that i can bargain with which we will notate by

$$W_i = \mathbb{E}_j[W_{ij}]$$

Thus,

$$U_i = \tau W_i$$



(a) Market Timeline Overview (b) Bargaining Phase Overview

Fig. 1: Figure 1a shows the progression of the market over time. Figure 1b shows the progression of one round of bargaining within a time step. Note that agents incur a cost of τ when at least one opts out and they must be re-matched in future matching phase. They incur a cost of δ when no agents opt out and they continue bargaining in the bargaining phase in the next time step.

4 Results

In this section, we will focus on the case of one type of firm and two types of candidates which we will call c_1 and c_2 candidates. As above, p will be the probability of a firm matching with a c_1 candidates. It is sufficient to consider only this case to demonstrate the existence of equilibria with unequal payoffs at equilibrium among agents who are otherwise indistinguishable. Here, c_1 and c_2 are left as arbitrarily different categories to show that these results hold for any categorization of candidates – by (intersections of) sensitive attributes or not. Qualitatively similar results will hold for markets with m kinds of firms and n kinds of candidates for any finite m and n, but we omit such results as they are not necessary to demonstrate the existence of these equilibria.

We now state our main theorem which formalizes the conditions necessary to create strategies that are in SPE for each possible combination of payoffs to different agents. This theorem implies that there exist strategies where different types of agents receive different values in the payoff range at equilibrium. Since we assumed no agent is entitled to more of the surplus than the others, it is exactly these asymmetric outcomes that we can consider discriminatory. We will highlight two cases in the discussion and use ideas from relational equality to analyze where the discrimination comes from. Thus, we will show how relational equality can be useful in the design of fair ML.

Theorem 1. *If $\tau \leq \frac{\delta^2}{1+\delta}$, then for any $p \in [0,1]$ and any w_1, w_2 that satisfy*

$$w_k \leq \frac{1}{2}\left(\frac{1+\delta-2\tau}{1-\tau}\right) \text{ for } k \in \{1,2\}, \tag{1}$$

$$w_1 \geq \frac{1}{2}\left(\frac{1-\delta+2\tau(1-p)w_2}{1-\tau p}\right), \tag{2}$$

$$\text{and } w_2 \geq \frac{1}{2}\left(\frac{1-\delta+2\tau p w_1}{1-\tau(1-p)}\right), \tag{3}$$

there exists an SPE where the firms obtain an expected payoff of $W_f = pw_1 + (1-p)w_2$, the c_1 candidates get an expected payoff of $W_{c_1} = 1 - w_1$ and the c_2 candidates get an expected payoff of $W_{c_2} = 1 - w_2$ at equilibrium.

The strategies that we will show are at SPE with these payoffs are given as automata [42] in Table 1 (these are based on strategies defined in [45]), which are parameterized by offers z_{ij} that agent i proposes to agent j. If the proposer i deviates, then both players move to a threat state where the new proposing agent j proposes a new split using additional parameters u_j. The goal is to choose these parameters for all possible pairs of firms and candidates so that, at equilibrium, we have $u_j = U_j$ for all agents j, i.e. they represent their outside options, and each offer z_{ij} is immediately accepted by the responding agent, resulting in the payoffs given in Theorem 1.

Fix agent i as an arbitrary agent in the market of some type (a firm or either candidate type). Then, let $\pi_i(Z, U)$ be the strategy of agent i where agent i plays according to Table 1, parameterized by $Z = \{z_{ij}\}$ and $U = \{u_i\}$. We will refer to this strategy as π_i for brevity.

Agent Actions	Base State	Threat State
Agent i Propose	$(z_{ij}, 1-z_{ij})$	$(1-\frac{u_j}{\delta}, \frac{u_j}{\delta})$
Agent i Accepts	-	Iff $y \geq \frac{u_i}{\delta}$
Agent j Propose	-	$(\frac{u_i}{\delta}, 1-\frac{u_i}{\delta})$
Agent j Accepts	Iff $y \geq 1-z_{ij}$	Iff $y \geq \frac{u_j}{\delta}$
Agent i Opts Out	Iff $1-y \leq z_{ij}$	Iff proposer and $y \geq \frac{u_j}{\delta}$
Agent j Opts Out	No	Iff proposer and $y \geq \frac{u_i}{\delta}$
Transitions	Go to Threat if agent i deviates in this match	Absorbing

Table 1: Strategies for matched agents i and j when agent i proposes first and agent j responds, parameterized by values z_{ij}, $u_i \in [0,1]$, e.g. $(z_{ij}, 1-z_{ij})$ is the split of the surplus agent i proposes to agent j. Note we use y as a generic variable representing an offer to the responder, so that the proposed split is $(1-y, y)$ regardless of whether an agent deviates from this strategy.

The range of payoffs in Theorem 1 is a direct result of the existence of a range of possible proposal values $z \in Z$. This range is a function of u_i and u_j, so long as each $u_i = U_i$, and satisfy the conditions given in Proposition 2 (again adapted from [45] for our setting). This range of z values arises from several properties of the bargaining game we imported from [45]. The proposer has an advantage and can ask for all of the surplus except for the responder's outside option which defines the upper end of their z range. Both agents have the opportunity to opt out and get their outside option at each stage of negotiation, so neither agent can take advantage of the other by forcing them to incur an extra discount factor δ. Thus, the responder can make a credible threat to reject any offer less than their equilibrium share since they can opt out if the proposer does not accept their counteroffer. To keep the proposer from opting out after the first round, the responder must offer them at least what their outside option would be with no time discount in the next time step. Thus, the least the proposer is ever willing to accept is related to the most that the responder could get in the next round and this defines the lower bound of their z range. See [45] for a more complete discussion of the intuition for how multiple equilibria arise out of these types of strategies.

Proposition 2 *Given that each agent i plays π_i as described above, the strategy profile $\pi = \{\pi_i\}$ is at SPE whenever, for all agents i, $u_i = U_i$ and, for all agents j that i can bargain with, $z_{ij} \in [1 - \delta(1 - \frac{u_i}{\delta}), 1 - u_j]$, $u_i \leq \delta^2(1 - \frac{u_j}{\delta})$, and $u_j \leq \delta^2(1 - \frac{u_i}{\delta})$.*

We start with a proof outline that reduces this proposition that the market is at SPE to showing that each possible bargaining match between two agents is at SPE given fixed expected outside options U_i under π. We complete the proof of Proposition 2 in Appendix A.

Proof (outline).

In the strategy π, each agent's strategy is described by an automoton, and as such, to show that π is an SPE, it suffices to show that are no beneficial "one-shot" deviations [42], i.e. no single action an agent may take to improve their expected payoff for every possible state the automota may be in, fixing all other actions. Moreover, by assumption that the market is stationary (any agent leaving the market is replaced with an agent of the same type), the expected outside option for any agent i is stationary as well, implying U_i is well-defined.

In π, each agent starts in the same state at the beginning of bargaining across all times that they are matched, and actions and payoffs depend only on the other agent they are bargaining with. Thus the optimal action at any given time, fixing all other actions, depends only on the restriction of the game to a single bargaining match with the same, fixed expected outside options U_i.

We can now return to proving Theorem 1 by using Proposition 2.

Proof. Consider an arbitrary w_1, w_2 that satisfy conditions (1)-(3) in the theorem statement and suppose $\tau \leq \frac{\delta^2}{1+\delta}$. Let $p \in [0, 1]$ be the probability of a firm

matching with a c_1 candidate. Then, we will construct strategies for firms f, c_1 candidates, and c_2 candidates that are in SPE with the desired expected payoffs.

Each agent will use a strategy given in Table 1, parameterized by offers. It suffices to consider strategies that are only dependent on the three types of agents – firms, c_1 candidates, and c_2 candidates – so that z_{ij}, u_i, and u_j only depends on the types of i and j. As such, we will refer to the parameters as z_{fc_k}, z_{c_kf}, u_f, and u_{c_k}, for $k \in \{1, 2\}$.

First, we set u_f and u_{c_k} so that they will be equal to the expected outside option (discussed in Section 3) of the firms and candidates respectively at equilibrium. Let

$$w_f = pw_1 + (1-p)w_2$$

Note that $w_1, w_2 \in [0, 1]$ and so $w_f \in [0, 1]$. Then let

$$\begin{aligned} u_f &= \tau w_f, \\ u_{c_1} &= \tau(1 - w_1), \\ \text{and } u_{c_2} &= \tau(1 - w_2). \end{aligned}$$

It follows immediately from these choices, the constraints on w_1 and w_2, and the assumption that $\tau \leq \frac{\delta^2}{1+\delta}$ that we have $u_f \leq \delta^2(1 - \frac{u_{c_1}}{\delta})$, $u_{c_1} \leq \delta^2(1 - \frac{u_f}{\delta})$, $u_f \leq \delta^2(1 - \frac{u_{c_2}}{\delta})$, and $u_{c_2} \leq \delta^2(1 - \frac{u_f}{\delta})$ as needed for Proposition 2.

We now need to set z_{fc_k} and z_{c_kf}, i.e. Z. In particular, we need $z_{fc_k} \in \left[1 - \delta\left(1 - \frac{u_f}{\delta}\right), 1 - u_{c_k}\right]$ and $z_{c_kf} \in \left[1 - \delta\left(1 - \frac{u_{c_k}}{\delta}\right), 1 - u_f\right]$ to be able to apply Proposition 2. But we also need

$$1 - w_k = \frac{1}{2}z_{c_kf} + \frac{1}{2}(1 - z_{fc_k}),$$

because we need $1-w_k$ to be the expected payoff for c_k candidates: From Table 1, observe that whatever split (z_{ij}, 1-z_{ij}) is proposed first, it is always accepted. And recall that an agent proposes first with probability $\frac{1}{2}$ and responds first with probability $\frac{1}{2}$, so the expected payoff for candidate c_k is exactly $\frac{1}{2}z_{c_kf} + \frac{1}{2}(1 - z_{fc_k})$.

Moreover, recall that a firm matches with a c_1 candidate with probability p and with a c_2 candidate with probability $1-p$, so the expression for the expected payoff to a firm is

$$p\left(\frac{1}{2}z_{fc_1} + \frac{1}{2}(1 - z_{c_1f})\right) + (1-p)\left(\frac{1}{2}z_{fc_2} + \frac{1}{2}(1 - z_{c_2f})\right) = pw_1 + (1-p)w_2 = w_f.$$

To satisfy the expected payoff expressions, it suffices to choose an arbitrary $z_{fc_k} \in \left[1 - \delta\left(1 - \frac{u_f}{\delta}\right), 1 - u_{c_k}\right]$ and set

$$z_{c_kf} = 1 + z_{fc_k} - 2w_k$$

To satisfy $z_{c_kf} \in \left[1 - \delta\left(1 - \frac{u_{c_k}}{\delta}\right), 1 - u_f\right]$, the following bounds on z_{fc_k} results from the setting above

$$2w_k - \delta\left(1 - \frac{u_{c_k}}{\delta}\right) \leq z_{fc_k} \leq 2w_k - u_f.$$

Since we have $z_{fc_k} \in \left[1-\delta\left(1-\frac{u_f}{\delta}\right), 1-u_{c_k}\right]$, we require all of the following constraints to hold

$$2w_k - \delta\left(1-\frac{u_{c_k}}{\delta}\right) \leq 2w_k - u_f, \qquad (i)$$

$$2w_k - \delta\left(1-\frac{u_{c_k}}{\delta}\right) \leq 1 - u_{c_k}, \qquad (ii)$$

$$1 - \delta\left(1-\frac{u_f}{\delta}\right) \leq 2w_k - u_f, \qquad (iii)$$

$$1 - \delta\left(1-\frac{u_f}{\delta}\right) \leq 1 - u_{c_k}. \qquad (iv)$$

Solving (i)-(iv) for bounds on w_1 and w_2 leads to the following constraints

$$w_1 \geq w_2 - \frac{\delta-\tau}{\tau(1-p)}, \qquad (1)$$

$$w_2 \geq w_1 - \frac{\delta-\tau}{\tau p}, \qquad (2)$$

$$w_k \leq \frac{1}{2}\left(\frac{1+\delta-2\tau}{1-\tau}\right), \qquad (3)$$

$$w_1 \geq \frac{1}{2}\left(\frac{1-\delta+2\tau(1-p)w_2}{1-\tau p}\right), \qquad (4)$$

$$\text{and } w_2 \geq \frac{1}{2}\left(\frac{1-\delta+2\tau p w_1}{1-\tau(1-p)}\right). \qquad (5)$$

Notice that (1) and (2) are always satisfied since $w_k \leq 1$ and $\frac{\delta-\tau}{\tau(1-p)} \geq 1$ and $\frac{\delta-\tau}{\tau p} \geq 1$ when $\tau \leq \frac{\delta}{2}$ which is true when $\tau \leq \frac{\delta^2}{1+\delta}$. So, (1) and (2) are trivially satisfied with $w_k \geq 0$.

Therefore, we need w_1 and w_2 to satisfy $(3)-(5)$ so that $(i)-(iv)$ are satisfied. Since w_1 and w_2 are given to satisfy these constraints by the theorem statement, we can conclude $z_{fc_k} \in \left[1-\delta\left(1-\frac{u_f}{\delta}\right), 1-u_{c_k}\right]$ and $z_{c_k f} \in \left[1-\delta\left(1-\frac{u_{c_k}}{\delta}\right), 1-u_f\right]$.

As a consequence of the way we set the Z values, the expected payoffs to each agent at equilibrium will be such that all firms get an expected payoff of $W_f = pw_1 + (1-p)w_2$ and all c_k candidates get an expected payoff of $W_{c_k} = 1 - w_k$. As a result, the expected outside options for each agent will be exactly what we set the parameters as, that is, $u_f = U_f$ and $u_{c_k} = U_{c_k}$.

Now we have a set of strategies π for every agent with $z_{fc_k} \in [1-\delta\left(1-\frac{u_f}{\delta}\right), 1-u_{c_k}]$, $z_{c_k f} \in [1-\delta\left(1-\frac{u_{c_k}}{\delta}\right), 1-u_f]$, $u_f \leq \delta^2(1-\frac{u_{c_k}}{\delta})$ and $u_{c_k} \leq \delta^2(1-\frac{u_f}{\delta})$. We also have $u_i = U_i$ for all agents i and, therefore, we can apply Proposition 2 and say that π is in SPE where all firms get $pw_1 + (1-p)w_2$ in expectation and all c_k candidates get $1-w_k$ in expectation and this concludes the proof.

4.1 Discussion

From constraints (2) and (3) in the Theorem 1 statement, notice that the lower bound on w_1 depends on a fixed value of w_2 and vice versa. Suppose $w_2 > w_1$, these constraints imply that there is an additive gap between the lower bound on w_1 and the value of w_2. The largest such gap is $\frac{\delta-\tau}{1-\tau p}$ which is $\approx \delta$ as $\tau \to 0$ such that the gap can be greater than $\frac{1}{2}$ when δ is sufficiently larger than τ. This means that there are cases where one type of candidate gets more then $\frac{1}{2}$ of the surplus and the other gets less than $\frac{1}{2}$. Thus, by Theorem 1, there exist strategies that are in SPE where c_1 candidates are getting significantly more in expectation than c_2 candidates at equilibrium. Note that the gap size grows as the probability p of matching with c_1 candidates grows and that, conversely, the gap persists even when p is small such that increasing the representation of c_2 candidates alone does not prevent the gap in payoffs. Similarly, the largest additive gap between a fixed value of w_1 and the lower bound of w_2 is $\frac{\delta-\tau}{1-\tau(1-p)}$.

In this market, no agent has any advantage over the others in terms of information, power in the bargaining game, and claim to the surplus, at the start and yet, as we have shown, it is possible for agents to choose strategies that are in SPE where one type of candidate receives a greater split of the surplus than the other. As such, this market is susceptible to discrimination without any initial asymmetric advantage among any of the agents. Although fairness metrics based on distributional inequality would detect this instance of discrimination, their normative assumptions would say that equalizing resources, like the split of the surplus, again would solve future discrimination. However, our model shows that the same type of discrimination could occur again even after temporarily equalizing resources. As such, traditional fairness metrics would be insufficient to address the discrimination in this model.

In our market, c_1 candidates were able to credibly threaten to reject a larger value than c_2 candidates. The ability for the candidates (and firms) to set their strategies in this way is not resource based. We can turn to relational equality for a possible explanation: Perhaps the social relationships between the firms and both kinds of candidates are not the same. By the firms choosing a strategy where they take more of the surplus from c_2 candidates than c_1 candidates, the firms are showing their belief that c_2 candidates are not entitled to as much of the surplus as c_1 candidates for an arbitrary reason. Further, the candidates confirm this view through their choice of a strategy that accepts the firms' proposals. Crucially, the firms and candidates had the correct belief at equilibrium about the outside options of their opponents. So, *beliefs about your opponent's behavior and outside option* engendered the inequality we see at equilibrium.

Further, if it is possible for there to be a difference in the initial social relationships between the agents, then it is possible for these relationships to change after equilibrium like the resource allocation changed from initiation to the equilibrium state in our market. Therefore, there could be *additional discrimination* happening undetected in our model in terms of an altering of the quality of social relationships between, say, c_2 candidates and c_1 candidates. The asymmetry in the payoffs to the agents opens up the possibility for feedback loops to exac-

erbate the differential treatment they experience. We leave formally measuring social relationship quality and how it changes as future work.

Now consider another case, where $w_1 = w_2 = \frac{1}{2}(\frac{1+\delta-2\tau}{1-\tau})$ and note that this value is $> \frac{1}{2}$. Here, the firms' strategy does not depend on the candidate type and the candidates receive the same split of the surplus at the end. However, the firms then get more of the surplus than the candidates even though we assumed that the firms and candidates are equally entitled to the surplus. In this way, firms were able to acquire more power over the candidates at equilibrium which gives us insight into the social relationship between the two types of agents in this model. However, we do not say how the firms were able to acquire this power, and we leave to future work studying how agents choose their strategies in a way that gives them an advantage at equilibrium state. Different from above, traditional fairness metrics based on sensitive attributes *would not* detect discrimination in this case since there is no difference in resource allocation between candidate types. This indicates that it is necessary to look beyond sensitive attributes to uncover possible instances of discrimination in a model.

4.2 Conclusion and Future Work

In light of these results, it may be possible for an ML model to exacerbate the social relationship differences seen in the highlighted cases. This underscores the importance of expanding ML notions of fairness to include non-distributional notions of equality, like the quality of social relationships between all agents in a model. This work does not present a solution to this type of discrimination nor do we define all instances of relational equality violations. Rather we hope to bring attention to a blind spot in current discussions of measuring fairness in algorithms. We do this through the example of a discriminatory outcome that requires something more than a distributional view of equality, namely relational equality, to fully understand the source of discrimination.

Future work will focus on understanding how to detect instances of relational inequality of the kind that we have described and correct for it. To this end, we plan to demonstrate how agents might learn the strategies we have described over time using their *beliefs about the other agents' (and their own) outside options*. Further, we would like to be able to design measures that can detect relational inequality. Finally, we would like to provide a framework for designing ML algorithms that do not exhibit this type of discriminatory outcome.

References

1. Adam, G.A., Chang, C.H.K., Haibe-Kains, B., Goldenberg, A.: Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In: Machine Learning for Healthcare Conference. pp. 710–731. PMLR (2020)
2. Agranov, M., Cotton, C., Tergiman, C.: The use of history dependence in repeated bargaining (2018)

3. Anderson, E.S.: What is the point of equality? Ethics **109**(2), 287–337 (1999)
4. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of data and analytics, pp. 254–264. Auerbach Publications (2016)
5. Arrow, K.J.: The theory of discrimination. In: Discrimination in labor markets, pp. 1–33. Princeton University Press (2015)
6. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. Nips tutorial **1**, 2017 (2017)
7. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021)
8. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 514–524 (2020)
9. Birhane, A.: Algorithmic injustice: a relational ethics approach. Patterns **2**(2), 100205 (2021)
10. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery **21**, 277–292 (2010)
11. Cavounidis, C., Lang, K.: Discrimination and worker evaluation. Tech. rep., National Bureau of Economic Research (2015)
12. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810 (2018)
13. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women. In: Ethics of data and analytics, pp. 296–299. Auerbach Publications (2018)
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
15. Dworkin, R.M.: Equality, luck and hierarchy. Philosophy & Public Affairs **31**(2), 190–198 (2003)
16. Elford, G.: Survey article: Relational equality and distribution. Journal of Political Philosophy **25**(4) (2017)
17. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S.: Runaway feedback loops in predictive policing. In: Conference on fairness, accountability and transparency. pp. 160–171. PMLR (2018)
18. Eubanks, V.: Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press (2018)
19. Fang, H., Moro, A.: Theories of statistical discrimination and affirmative action: A survey. Handbook of social economics **1**, 133–200 (2011)
20. Fernandez, R., Glazer, J.: Striking for a bargain between two completely informed agents (1989)
21. Fish, B., Stark, L.: It's not fairness, and it's not fair: The failure of distributional equality and the promise of relational equality in complete-information hiring games. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '22, Association for Computing Machinery, New York, NY, USA (2022), https://doi.org/10.1145/3551624.3555296
22. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 329–338 (2019)
23. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)

24. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning–what is it good for? arXiv preprint arXiv:2004.00686 (2020)
25. Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society **22**(7), 900–915 (2019)
26. Hu, L., Chen, Y.: Fairness at equilibrium in the labor market. arXiv preprint arXiv:1707.01590 (2017)
27. Hutchinson, B., Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 49–58 (2019)
28. Hyde, C.E.: Multiple equilibria in bargaining models of decentralized trade. Economic Theory pp. 283–307 (1997)
29. Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 375–385 (2021)
30. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
31. Kasy, M., Abebe, R.: Fairness, equality, and power in algorithmic decision-mhickmaning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 576–586 (2021)
32. Kelley, B., Sisneros, L.: Broadband access and the digital divides. policy brief. Education Commission of the States (2020)
33. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in neural information processing systems **30** (2017)
34. Lax, D.A.: Commentary on understanding pennzoil v. texaco: Market expectations of bargaining inefficiency and potential roles for external parties in disputes between publicly traded companies. Va. L. Rev. **75**, 367 (1989)
35. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
36. Malik, N.: Does machine learning amplify pricing errors in housing market?: Economics of ml feedback loops. Economics of ML Feedback Loops (September 18, 2020) (2020)
37. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
38. Miller, C.C.: When algorithms discriminate. The New York Times **9**(7), 1 (2015)
39. Neuhäuser, L., Karimi, F., Bachmann, J., Strohmaier, M., Schaub, M.T.: Improving the visibility of minorities through network growth interventions. Communications Physics **6**(1), 108 (2023)
40. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
41. O'neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Crown (2017)
42. Osborne, M.J., Rubinstein, A.: Bargaining and markets. Academic press (1990)
43. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
44. Phelps, E.S.: The statistical theory of racism and sexism. The american economic review **62**(4), 659–661 (1972)
45. Ponsati, C., Sákovics, J.: Rubinstein bargaining with two-sided outside options. Economic Theory **11**, 667–672 (1998)

46. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Annals of internal medicine **169**(12), 866–872 (2018)
47. Roselli, D., Matthews, J., Talagala, N.: Managing bias in ai. In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 539–544 (2019)
48. Ruf, B., Detyniecki, M.: Towards the right kind of fairness in ai. arXiv preprint arXiv:2102.08453 (2021)
49. Schemmel, C.: Distributive and relational equality. Politics, philosophy & economics **11**(2), 123–148 (2012)
50. Shaked, A., et al.: Opting out: bazaars versus 'hi tech'markets. Investigaciones Economicas **18**(3), 421–432 (1994)
51. Tadelis, S.: Game theory; an introduction (2013)
52. Ulph, A., Ulph, D.: Labour markets, bargaining and innovation. European Economic Review **42**(3-5), 931–939 (1998)
53. Zhang, J., Bareinboim, E.: Fairness in decision-mhickmaing—the causal explanation formula. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

A Proposition 2 Proof

Recall Proposition 2.

Proposition 2 *Given that each agent i plays π_i as described above, the strategy profile $\pi = \{\pi_i\}$ is in an SPE whenever for all agents i $u_i = U_i$ and, for all agents j that i can bargain with, $z_{ij} \in [1 - \delta(1 - \frac{u_i}{\delta}), 1 - u_j]$, $u_i \leq \delta^2(1 - \frac{u_j}{\delta})$, and $u_j \leq \delta^2(1 - \frac{u_i}{\delta})$.*

Using the one-shot deviation principal we will prove that the strategies in Table 1 are in a SPE for any values of z_{ij} within the ranges stated in Proposition 2. Ponsati et al.[45] showed similar strategies to our Table 1 strategies constitute an SPE in a singular bargaining game.

The one-shot deviation principle holds if no agent can deviate from their strategy at a single node and receive a higher expected payoff. This is also referred to as being one-stage unimprovable and strategy profiles that are one-stage unimprovable at all decision points on and off the equilibrium path constitute an SPE [51].

Intuitively, by showing that no single deviation is beneficial to agent i at any node, their original strategy in Table 1 must be an SPE. Our game tree is infinite, but the types of nodes possible in our game are finite, and can be classified into two states – a base and threat state. We will enumerate against the type of deviations possible in each case and by each agent below, allowing us to cover all potential deviations. We also include the optimal strategy to demonstrate the optimal outcome and to prove that deviating does not yield better outcomes than those afforded by Table 1.

We take agent i to be the proposer first and agent j to be the responder first. Recall that y denotes a generic offer to the responder. Agent i can be either a candidate or firm, and agent j would be of the opposite type. Recall that the bargaining phase ends when one agent decides to opt out, sending both agents back to the matching pool to begin the matching and bargaining process anew or if they agree on a surplus split. Since the agent expects to obtain their outside options by opting out, for the purposes of this proof, both agents achieve their expected outside options in full immediately upon opting out.

A.1 Base State Deviations

In this paragraph, we will consider deviations when agent i is proposing in the base state. First, suppose agent i proposes according to Table 1 and offers $(z_{ij}, 1 - z_{ij})$ Here, agent j will accept the proposal and agent i will get a payoff of z_{ij}. Now suppose that agent i deviates and offers $(1 - y, y)$ where $y < 1 - z_{ij}$. Agent j will decline since $y < 1 - z_{ij}$. Since agent i deviated and $1 - y > z_{ij}$, agent i will not opt out and neither will agent j. Instead, they will transition to the threat state and agent j becomes the proposer in the next time step. Agent j will propose a split of $(\frac{u_i}{\delta}, 1 - \frac{u_i}{\delta})$. Since $y \geq \frac{u_i}{\delta}$, agent i will accept the offer, giving agent i a final payoff of $\delta \cdot \frac{u_i}{\delta} = u_i$. Since $z_{ij} \geq u_i$ by assumption, their deviation

did not improve their payoff. Now, let us suppose instead agent i deviates and proposes $(1-y, y)$ where $y > 1-z_{ij}$. According to agent j's strategy, they accept since $y \geq 1 - z_{ij}$. Thus, agent i's final payoff is $1 - y$, where $1 - y < z_{ij}$. Thus, their deviation did not improve their payoff. From both of these cases, agent i has no useful deviation when proposing in the base state.

Now, let us consider deviations made by agent j when they are responding to a proposal in the base state. First, suppose agent i proposes $(1 - y, y)$ where $y \geq 1 - z_{ij}$. If agent j does not deviate, then they would accept y and recieve it as their payoff. Suppose instead that agent j deviates and declines y. Then, agent i would opt out since $1-y \leq z_{ij}$. Thus, agent j receives an expected payoff of u_j. By assumption, $1 - z_{ij} \geq u_j$, thus $y \geq 1 - z_{ij} \geq u_j$ and agent j did not benefit from deviating in this partition. Next, suppose agent i proposes $(1-y, y)$ where $y < 1-z_{ij}$. If agent j does not deviate, then they will reject this offer and neither agent will opt out since $1 - y > z_{ij}$ and both agents transition to the threat state in the next time step. Agent j becomes the proposer and proposes $(\frac{u_i}{\delta}, 1 - \frac{u_i}{\delta})$ which agent i accepts and agent j gets a payoff of $\delta(1 - \frac{u_i}{\delta})$. Suppose instead that agent j deviates and accepts $y < 1 - z_{ij}$ as a payoff. Since we've assumed that $1 - \delta(1 - \frac{u_i}{\delta}) \leq z_{ij}$ from the range of possible z_{ij} values, we see that $y \leq \delta(1 - \frac{u_i}{\delta})$. As a result, agent j would not have improved their utility by deviating from their equilibrium strategy at this node. So, agent j has no useful deviations when they are responding to an offer in the base state.

Now we consider agent i opt out deviations in the base state. First, suppose that agent i opts out when $1-y \leq z_{ij}$. Thus, the game ends and agent i receives their outside option of u_i. Suppose instead agent i deviates and doesn't opt out when $1 - y \leq z_{ij}$. Agent j would also not opt out and become the proposer in the next time step and they will now offer $(\frac{u_i}{\delta}, 1 - \frac{u_i}{\delta})$. Agent i will accept this offer, and their final payoff will be $\delta\frac{u_i}{\delta} = u_i$. As a result, this was not a useful deviation. Next, suppose agent i does not opt out when $1-y > z_{ij}$. Then, agent j would also not opt out and would propose $(\frac{u_i}{\delta}, 1 - \frac{u_i}{\delta})$ in the next time step which agent i would accept such that agent i gets a payoff of $\delta\frac{u_i}{\delta} = u_i$. Suppose instead that agent i deviates by opting out when $1 - y > z_{ij}$. Agent i's final payoff is then u_i and this was not a useful deviation. As a result, agent i does not benefit from deviating from their opt out strategies in the base state.

We also must consider agent j deviating from their opt out strategy in the base state. Under our base state strategies, agent j never opts out. If agent j deviates and opts out, their payoff would be u_j. Suppose agent j had not deviated, then they would get to propose an offer of $\frac{u_i}{\delta}$, which agent i would accept. Then, agent j gets a payoff of $\delta(1 - \frac{u_i}{\delta})$. From the conditions of Proposition 2, we have $u_j \leq \delta^2(1 - \frac{u_i}{\delta}) \leq \delta(1 - \frac{u_i}{\delta})$, so agent j did not improve their payoff by opting out.

A.2 Threat State Deviations

Notice in Table 1 that the strategies for agent i and agent j are completely symmetric. So, in this section, it suffices to consider deviations for just agent i from their threat state strategies. The proofs for agent j turn out to be exactly

the same with the indices i and j swapped. Throughout these proofs, we will assume without loss of generality that we begin in time step $t > 1$ of bargaining.

In this paragraph, we will consider deviations from agent i's proposer strategy. First, suppose agent i proposes $(1 - \frac{u_j}{\delta}, \frac{u_j}{\delta})$ in accordance with the Table 1 strategies. Agent j will accept this offer and the final payoff for agent i will be $\delta^t(1 - \frac{u_j}{\delta})$. Suppose instead that agent i deviates by proposing $(1 - y, y)$ where $y < \frac{u_j}{\delta}$. Agent j will reject this offer since $y < \frac{u_j}{\delta}$. Then agent i will not opt out since $y < \frac{u_j}{\delta}$ and agent j will also not opt out since they are not the proposer. Bargaining then moves on to time step $t + 1$. Agent j will propose $(\frac{u_i}{\delta}, 1 - \frac{u_i}{\delta})$. Agent i will accept this offer since $y \geq \frac{u_i}{\delta}$. Thus, agent i's final payoff is $\delta^{t+1}\frac{u_i}{\delta} = \delta^t u_i$. Since we assumed that $u_i \leq \delta^2(1 - \frac{u_j}{\delta})$, then it follows that $\delta^t u_i \leq \delta^t(1 - \frac{u_j}{\delta})$ and agent i did not get a better payoff by deviating. Next, suppose agent i deviates by proposing $(1 - y, y)$ where $y > \frac{u_j}{\delta}$. According to the Table 1 strategies, agent j accepts since $y \geq \frac{u_j}{\delta}$. Then, agent i gets $\delta^t(1-y) < \delta^t(1 - \frac{u_j}{\delta})$. Therefore, deviating does not give agent i a better payoff when they are the proposer in the base state.

Let us now consider deviations from agent i's responder strategy in the threat state. First, suppose agent j has offered $y < \frac{u_i}{\delta}$. If agent i does not deviate, then they will reject this offer and not opt out since they are not the proposer. Agent j will also not opt out since $y < \frac{u_i}{\delta}$. Bargaining then moves on to time step $t+1$ where agent i is the proposer. Agent i would then propose $(1-\frac{u_j}{\delta}, \frac{u_j}{\delta})$ and agent j would accept this offer since $y \geq \frac{u_j}{\delta}$. As a result, agent i would receive a payoff of $\delta^{t+1}(1-\frac{u_j}{\delta})$. Suppose instead that agent i deviates and accepts $y < \frac{u_i}{\delta}$. Then, agent i gets a payoff of $\delta^t y < \delta^t \frac{u_i}{\delta}$. Recall the assumption that $\frac{u_i}{\delta} \leq \delta(1 - \frac{u_j}{\delta})$. From this, $\delta^t y < \delta^{t+1}(1-\frac{u_j}{\delta})$ and deviating does not give agent i a better payoff at this game node. Next, suppose agent j offers $y \geq \frac{u_i}{\delta}$. If agent i does not deviate, then they would accept and receive a payoff of $\delta^t y \geq \delta^t \frac{u_i}{\delta}$. Suppose instead that agent i deviates and rejects this offer. Then, agent j would opt out since they are the proposer and $y \geq \frac{u_i}{\delta}$. As a result, agent i gets a payoff of $\delta^t u_i$ which is not better than a payoff of $\delta^t y \geq \delta^{t-1} u_i$. Therefore, deviating does not give agent i a better payoff when deviating in the threat state as a responder.

We will now consider the cases where agent i deviates from their opt out strategy in the threat state. First, suppose agent i has just rejected an offer $y < \frac{u_i}{\delta}$. If agent i does not deviate then they will not opt out and agent j would also not opt out as the proposer since $y < \frac{u_i}{\delta}$. The bargaining moves on to time step $t+1$ with agent i as the proposer. Then, agent i proposes $y = \frac{u_j}{\delta}$ and agent j accepts and agent i gets $\delta^{t+1}(1 - \frac{u_j}{\delta})$ as a payoff. Suppose instead that agent i deviates and opts out. Then, agent i gets $\delta^t u_i$. Recall our assumption that $u_i \leq \delta^2(1 - \frac{u_j}{\delta})$. From this, $\delta^t u_i < \delta^{t+1}(1 - \frac{u_j}{\delta})$ and agent i is not better off by deviating at this node. Next, suppose agent i has just rejected an offer $y \geq \frac{u_i}{\delta}$. If agent i does not deviate, then they will not opt out, but agent j will opt out since they are the proposer and $y \geq \frac{u_i}{\delta}$. Then, agent i gets $\delta^t u_i$. Suppose instead that agent i deviates and opts out, then they will also get $\delta^t u_i$ and agent i is not better off by deviating at this node. Next, suppose agent j has just rejected an offer $y < \frac{u_j}{\delta}$. If agent i does not deviate and does not opt out, then agent j would

also not opt out since they are the responder. Bargaining then moves on to time step $t+1$ and agent j is the proposer. Agent j would propose $\frac{u_i}{\delta}$ and agent i would accept and get $\delta^{t+1}\frac{u_i}{\delta} = \delta^t u_i$. Suppose instead that agent i deviates and opts out as the proposer with $y < \frac{u_j}{\delta}$. Then, agent i payoff is $\delta^t u_i$ and agent i is not better off by deviating at this node. Finally, suppose agent j has just rejected some offer $y \geq \frac{u_j}{\delta}$. If agent i does not deviate and opts out after agent j rejects the offer $y \geq \frac{u_j}{\delta}$, then agent i again gets $\delta^t u_i$. Suppose instead that agent i deviates and does not opt out as proposer when $y \geq \frac{u_j}{\delta}$. Then, agent j is the responder, so they would not opt out as well and bargaining moves on to time step $t+1$. Agent j is now the proposer and offers $y = \frac{u_i}{\delta}$ which agent i accepts and agent i gets a payoff of $\delta^{t+1}(\frac{u_i}{\delta}) = \delta^t u_i$. Therefore, agent i does not get a better payoff by deviating at this node and thus agent i does not benefit from deviating from their opt out strategy in Table 1.

Thus, we have shown that no single deviation from the Table 1 strategies is beneficial to either agent in any state at any time step. Therefore, through the one-shot deviation principle, this set of strategies is in an SPE under the conditions of Proposition 2.