MSc Economics

Track: Behavioural Economics and Game Theory

Master Thesis

Reverse Ultimatum Game with Exogenous Risk of Breakdown

by

Kalpit Raj
15435792

First Supervisor: Prof. dr. C.M. (Matthijs) van Veelen
Second Supervisor: Prof. dr. J.H. (Joep) Sonnemans
Credit: 15 ECTS

UNIVERSITY OF AMSTERDAM
Economics & Business

Economics and Business Ethics Committee (EBEC)

Amsterdam Business School
Plantage Muidergracht 12
1018 TV Amsterdam
The Netherlands
T +31(0)20 525 76 51
www.abs.uva.nl

Date	**Our reference**	
16-07-2025	EB-18277	
Contact	**Telephone**	**E-mail**
Economics and Business Ethics Committee	+31 (0) 20 525 76 51	ebec@uva.nl

Subject

EBEC approval

Dear Kalpit Kalpit Raj,

The EBEC has evaluated your research proposal and has concluded that it is consistent with the UvA EB ethics regulations.

It has therefore decided to approve the research project. Your project and approval number is: EB-18277.

The information about your project and this approval can be retrieved at https://rms.uva.nl/browse/EB-18277.

Sincerely,

Prof. dr. V.S. Maas
Chair EBEC

Abstract

This research modifies a two player bargaining model inspired by the Reverse Ultimatum Game to examine scenarios in which negotiations face exogenous breakdown risk stemming from elements of the surrounding environment that lie entirely beyond the negotiators' control. A standard theoretical equilibrium analysis is conducted, and a solution is derived through backward induction. The model is then simulated using reinforcement learning agents who learn to play the game, reflecting human-like strategic behaviour. This approach allows for the observation of how strategies emerge and how payoffs converge across the population. The findings highlight the effectiveness of reinforcement learning in economic games where uncertainty is inherent and optimal strategies evolve dynamically. Results indicate that agents successfully converge to the theoretical solution, and the distribution of offers across the population shifts when breakdown risks are low, favouring the responder through higher payoffs. The study also examines the evolutionary dynamics of risk-seeking responders who strategically position themselves to demand higher offers, demonstrating that posing as a hard bargainer can compel proposers to increase their share, potentially leading to better payoffs.

Table of Contents

Introduction

Bargaining is ubiquitous. It is quite essential to society, and has been widely studied by different fields, in economic, business, and behavioural psychology (Backus and Tadelis, 2020; Reppert, 2020). The primary aim of bargaining is to reach a decision acceptable to two or more parties. It involves finding equilibrium points that balance mutual needs and opportunities. However, in most real-life situations, bargaining is not 'frictionless' and often incurs costs from haggling (Muthoo, 1999). One primary source of bargaining costs arises from the possibility that negotiations may break down due to exogenous and uncontrollable factors. Suck risks of breakdown arises from many external factors, which negotiators perceive as random. For example, an attractive outside option might emerge for one party, opportunities for cooperation may vanish, or a third party might intervene in the process (Muthoo, 1999). Thus, to better understand the relationship between negotiation process and participants, it is crucial to search for mathematical models and alternative decision making.

It is game theory that provides such formal analysis of conflict and cooperation. Its concepts offer a language to formulate, structure, analyse, and understand strategic interactions between agents pursuing their own goals (Turocy and von Stengel, 2001; Peleckis, 2015). More importantly, there exists a subset of negotiations characterized by round-based dynamics, in which each rejected proposal triggers progression to the following round. In such models, equilibrium outcomes depend crucially on who has the right to make a proposal at what point in time, and on how costly the delay in moving from one round to the next is. Building on this structure, the literature applies such game-theoretic models to diverse economic phenomena: fairness and other-regarding preferences via dictator games (Carpenter, 2003); deadline effects (Roth et al., 1988; Güth et al., 2005); reputation and repeated interaction (Kreps et al., 1982); asymmetric information (Valley et al., 1992); outside options and threat points (Binmore et al., 1986); communication and cheap talk (Valley et al., 1991); bargaining under uncertainty and ambiguity (Beattie et al., 1994); and social distance and identity (Hoffman et al., 1996).

Given the centrality of proposal power and delay costs, the deadline dimension is especially salient. Gneezy et al. (2003) advance this line of inquiry by introducing the Reverse Ultimatum Game (RUG), which models how time can be strategically leveraged to make deadlines binding. The simplest version of the reverse ultimatum game involves two players, A and B,

who must decide how to split N dollars/tokens. Player A initiates by proposing an offer $x \in [0, N]$ integer number of tokens to Player B. If Player B accepts the offer, the game ends with Player B receiving x tokens and Player A receiving $(N - x)$ tokens. However, if Player B rejects the offer, Player A may respond with a higher offer, which must be strictly greater by at least one token, ensuring both players continue to receive positive shares. This process continues until Player B accepts an offer, or Player A decides to end the negotiation, in which case both players receive nothing. The game also terminates if Player B rejects the highest feasible offer that Player A can make. In much of the literature, deadlines are considered tight, implying no possibility for extensions. This generates incentives to reach agreements in the "eleventh hour", that is at or very close to the deadline. (For empirical support, see Cramton and Tracy (1992) and experimental validation in Gneezy et al. (2003).

This bargaining framework provides a foundation for addressing the paucity of research on exogenous breakdown risk in multi-stage negotiations. In settings such as labour negotiations, political agreements, and resource allocation, bargaining often unfolds over multiple stages. For example, consider the Brexit negotiations between the EU and the UK, where each postponement increased the likelihood of an automatic "no-deal" outcome due to legislative and statutory deadlines imposed by the UK Parliament as an exogenous risk which neither party could fully control. These risks re-emerged at each stage, mirroring the uncertainty faced in multi-round bargaining environments. This situation can be described as a dynamic bargaining environment involving two parties facing a stationary, stochastically extendable deadline. Notably, experimental studies on the effect of exogenous breakdown in bargaining are limited. The existing work (Rami Zwick et al, 1992), (Weg et al, 1990) predominantly rely on the alternating, two-person, sequential bargaining framework of Rubinstein (1982) which inherently assumes shared bargaining power over infinite horizons.

Because negotiations are often costly and can vary widely, involving everything from simple take it or leave it offers to complex exchanges of concessions and counteroffers, this research focuses on asymmetric bargaining in which one party controls the offer sequence over a finite stage by stage process. Apart from the study by Zwick et al. (1992), much of the existing work is theoretical and lacks empirical analysis. Accordingly, this study contributes by providing a standard game-theoretic analysis of strategic bargaining models with exogenous breakdown risk, validated through Q-learning based agent simulations which has practical implications in e-commerce.

The research extends the reverse ultimatum game framework into a multi-stage version, in which after each rejected offer, the proposer faces a new decision: to either end the negotiation altogether, resulting in both players receiving nothing, or to make a strictly better subsequent offer (minimum increase of a single integer token), but only with a probability $(1 - p)$; with probability (p), the negotiation breaks down and both players receive nothing. In clear terms, the transition to the next bargaining stage is not deterministic. After each rejection, the game is subject to a breakdown, governed by an exogenously specified and stationary probability $p \in (0,1)$. This probability is not known to the agents and is independently drawn at each stage, introducing fundamental uncertainty about whether bargaining will persist. In the limiting case $p = 0$, the model reduces to the classical reverse ultimatum game with guaranteed continuation, whereas at $p = 1$ every rejection brings the negotiation to an immediate end.

As such, the game can be interpreted as having a stationary and stochastically determined continuation structure, where the possibility of proceeding to the next stage is governed by an exogenous risk of breakdown. This creates a negotiation environment where the horizon for agreement is uncertain and unfolds probabilistically after each rejection, rather than being bounded by a fixed deadline. The theoretical analysis undertaken in this research yields an equilibrium acceptance threshold, derived through backward induction, which delineates the minimum offer a responder would rationally accept. This, in turn, characterizes a set of acceptable offers above the threshold and identifies the proposer's optimal initial offer that is immediately accepted, resulting in agreement at the first stage and constituting the subgame perfect outcome of the game.

Building on the theoretical analysis, the thesis entails a systematic investigation of play by adaptive agents that form and adjust their strategies in response to experience. The work deploys artificial agents that explore to play the game through trial and error, which causes them to adjust their strategies in response to experience. This dual approach shows how artificial negotiators adapt under uncertainty and underscores the need for agent-based bargaining models where optimal decisions emerge from complex, multi-actor interactions. The general framework for such agents is reinforcement learning in the machine learning literature (Sutton and Barto, 1998). Through the application of reinforcement learning agents

in uncertain environments, the research provides insight into the optimal policy learned by reinforcement learning agents.

In addition to exploring how artificial agents behave and respond under uncertain conditions, this research aims to illuminate the role of reputation in bargaining. By examining a heterogeneous population with varied risk tolerances, the study investigates the evolutionary stability of different strategies. The findings show that it is evolutionarily viable for responders to adopt a tough stance by consistently rejecting offers below an individually established threshold, even when this threshold exceeds the theoretical equilibrium. This behaviour can prompt proposers to increase their offers, as evidenced by the reputation built through recent outcome histories, and ultimately allows responders to accumulate greater rewards over time, despite the inherent breakdown risk.

Thus, the aim of this thesis is to explore computational approaches for artificial agents to play the reverse ultimatum game with exogenous risk of breakdown, comparing the obtained results with the classical game theoretical prediction. In particular, the research seeks to answer the following questions: How do artificial agents perform in playing the classical single responder reverse ultimatum game? How does uncertainty regarding the risk of breakdown in a negotiation impacts the distribution of offers? Can risk-seeking behaviour act as a profitable evolutionary strategy in bargaining environments with uncertain continuation, and under what conditions does it persist?

The remainder of the paper is organized as follows. The subsequent section presents a comprehensive review of the existing literature on bargaining and artificial agents, establishing the foundation for this study. It is followed by the methods section, which outlines the research methodology used to construct the simulation framework and conduct the standard theoretical analysis, including key implementation details. The results section reports the findings from agent-based simulations and introduces an extended framework for examining the effects of varied risk profiles under the influence of reputational bargaining. Finally, the discussion and conclusion section highlights the main research findings, acknowledges the limitations of the study, and suggests directions for future research.

Literature Review

Bargaining with Breakdown

Two-party bargaining is a foundational problem in economics with wide-ranging practical relevance, including commercial negotiations, labour contracts, and corporate mergers (Nash, 1950; Harsanyi, 1956; Luis C. Dias & Rudolf Vetschera, 2022). A critical feature of many such negotiations is the possibility of exogenous breakdown where the negotiation process may be abruptly terminated by factors beyond the control of either party. For example, bargaining may collapse if a third party seizes a shared opportunity, or due to external shocks such as regulatory changes, political upheavals, natural disasters, or technological disruptions. These events impose uncertain, externally-driven constraints on the negotiation horizon. Motivated by these scenarios, the present research investigates bargaining dynamics under stochastic exogenous breakdown. It seeks to understand how the risk of an externally imposed collapse shapes the strategic behaviour of agents, particularly when concessions must be weighed against the possibility of irreversible negotiation failure.

Rami Zwick et al. (1992) studied the two-player sequential bargaining behaviour with exogenous breakdown by considering bargaining as a game in extensive form with alternating offers. It was a seminal effort to empirically test for risk of exogenous failure instead of cost of delay *(δ)*, which is a key determinant of agreement timing in Rubinstein (1982) model of bargaining. However, a major limitation of the work was the assumption of common knowledge of risk among the participants despite real bargaining involving incomplete information, including uncertainty about the level of risk. This thesis addresses the limitation of their work by moving beyond the common knowledge of exogenous breakdown assumption and introducing uncertainty about the breakdown probability. The data is collected by running multiple simulations through co-evolving learning agents of proposers and responders that engage in multi-stage bargaining game over multiple episodes in a well-mixed population.

Herrera et al. (2025) developed a theoretical model of political negotiation in which a ruinous outcome (e.g., U.S. default) may occur following rejection of a proposal, introducing an exogenous risk of breakdown. The model considered a two-player, simultaneous bargaining model with uncertainty using stationary equilibrium and exogenous proposals that materialises

into a ruinous outcome at each subsequent disagreement. The result showed that regardless of how ruinous a crisis may be for each player, brinkmanship enlarges the equilibrium scope for agreement. This research could be understood to build on its limitation by moving away from simultaneous decision of approval voting on exogenous offers to a strategic structure of offer-making. Additionally, unlike where brinkmanship expands the scope of agreement (gridlock) in equilibrium due to shared knowledge of the breakdown risk, this study reveals that risk-taking behaviour by responders can lead to coercive gains. Specifically, responders who portray themselves as "tough" by consistently rejecting low offers (despite the risk of breakdown) which puts the negotiation at the brink of a precipice, are able to earn higher rewards over time. This shows that brinkmanship is not just a theoretical lever to induce agreement but can function as a viable strategy through which agents gain bargaining power in uncertain, sequential negotiations.

Calabuig and Olcina (2000) studied a repeated wage bargaining environment for players to adopt commitment strategies to influence outcomes. In particular, they introduce "fighting unions" who consistently demand the maximum allowable wage in every stage game. The model incorporates incomplete information by assuming a positive probability that either party may be a commitment type, and demonstrates that normal types may find it optimal to imitate these tough strategies to secure better payoffs. However, their model adopts a simultaneous bargaining structure, in which both the firm and the union submit wage offers in each round, and disagreement deterministically results in a strike, yielding zero payoffs to both parties. While analytically tractable, this setup overlooks sequential and asymmetric structures of negotiation that frequently occur in real-world institutional bargaining, such as procurement, public contracts, or regulatory decisions, where one party retains exclusive control over proposing offers.

The present research advances this literature by introducing a sequential bargaining framework with single-sided proposal authority and embedding a stochastic mechanism for breakdown. In contrast to the deterministic strike outcome assumed in their setup, rejection in the current model triggers a stochastic risk of breakdown, reflecting the fact that not all strike actions or delay decisions materialize into failure. This probabilistic risk structure offers a novel extension: some rejections escalate into a breakdown, while others permit renegotiation, mimicking real-world negotiation environments with uncertain escalation. By simulating this environment using learning agents, the analysis demonstrates that responders who engage in

risky posturing (i.e., accepting only higher offers) can earn higher long-term payoffs and survive evolutionarily. This provides empirical support for the strategic use of commitment and offers a dynamic perspective on bargaining with exogenous breakdown, complementing and extending the theoretical insights of Calabuig and Olcina.

Schauer et al. (2023) distinguished three forms of uncertainty by identifying what aspect of a negotiation the uncertainty pertains to, referred to as the 'objects of uncertainty'. Amongst these types, “issue-based unpredictability” arises when the consequences of choosing certain options are at least in part unpredictable. Such uncertainty is inherent in the choice options, which leaves the negotiators with payoffs that are risky (specific probabilities for each choice options), ambiguous (a range of possible probabilities for each choice option), or even uncertain (no information about the probabilities for each choice option). Given the framework developed in this research, responders choice to reject an offer opens the door to an uncertain outcome (zero or strictly better offer). In light of their assessment, the author highlights the lack of research in the aforementioned domain. Thus, this thesis could be understood as an effort to add insight the existing literature by providing an equilibrium analysis and supporting it with simulation data.

Kreps, Milgrom, Roberts and Wilson (1982) establish that in repeated bargaining with imperfect information, a history of acceptances and rejections serves as a credible commitment device. Their model demonstrates that even a small chance of facing a responder who rejects low offers sustains higher initial proposals in equilibrium. Crucially, however, their framework assumes identical players with infinite horizons and no risk of breakdown. There is no mechanism for finite-horizon collapse risk, nor for private variation in acceptance behaviour beyond a common cutoff. Furthermore, Muthoo (1999) rigorously characterizes subgame perfect equilibria in sequential bargaining when negotiations may collapse at each stage. His work shows how the threat of exogenous breakdown raises the minimum equilibrium offer under risk neutrality. Yet his analysis presumes that all responders share a single, identical acceptance rule and does not explore how observable past behaviour might signal individual risk tolerance.

This thesis puts an effort to fills these gaps by endowing each responder with a private acceptance threshold revealed through its recent reputation history and by embedding both proposer and responder strategies in an evolutionary framework. Through simulated

generations, the model identifies which threshold–offer combinations yield the highest fitness under incomplete information, thereby extending Kreps et al.'s reputation logic and Muthoo's collapse analysis to finite-horizon environments with diverse risk attitudes.

Reinforcement Learning Algorithm

Software agents have been extensively used in ultimatum game research. Kimbrough, Wu, and Zhong (2002) employed finite automata to compare agent behaviour with the predictions of classical game theory. Earlier studies formalised bounded rationality by modelling decision makers as finite automata constrained by a limited set of internal states (Golbeck et al. 2005; Hillmann and Kuhn 2010). Uyttendaele, De Jong, and Tuyls (2008) applied agent based approaches to investigate social dilemmas through the ultimatum game, while Kimbrough, Wu, and Zhong (2001) demonstrated how genetic learning agents can manage complex systems such as the MIT Beer Game. These rule based agents operate through explicit IF THEN structures, where each observed bargaining state activates a predefined response. This design allows researchers to track exactly how strategies evolve in response to the history of offers and outcomes. Building on this foundation, the present study develops an evolutionary model in which proposers follow rule based policies to capture risk heterogeneity among bargainers. Each proposer evaluates a responder's recent reputation history and selects an offer according to fixed rules. When embedded in an evolutionary framework, this approach identifies which strategies produce the highest fitness across generations in negotiation environments with incomplete information.

Furthermore, the present thesis moves beyond these approaches by pursuing a more advanced modelling paradigm that mimics human-like cognition in uncertain contexts. At the core of this approach is Q-learning (Watkins, 1989), an algorithm designed for agents to learn optimal actions through interaction with environments whose rules may not be fully known in advance. Though individual Q-learning has been deployed in multiple settings, the more complex scenarios involve multi-agent Q-learning, which inculcates "players" that can also be referred to as "agents" in a game which resorts to Q-learning to choose their actions. Thus, an individual agent in multi-agent Q-learning perceives its environment as nonstationary because the behaviour of the other agents changes over time due to the learning of an optimal policy which makes it quite difficult to analyse. Kimbrough et al. (2002) and Vasileuski Kramskova et al.

(2023) dealt with Q-learning in a multi-agent setting in the ultimatum game to run simulation with encoded behavioural biases. However, it is still under researched. Thus, this work is an adds to the understanding of multi-agents reinforcement learning (MARL) (Christianos et al., 2024) in uncertainty.

Larsen et al. (2010) introduced a reinforcement learning algorithm that adapts standard Q-learning to handle state uncertainty by using posterior-weighted updates. In real-world decision-making tasks, agents often receive ambiguous cues about which state they occupy. Their algorithm models this ambiguity probabilistically and adjusts the Q-value update based on beliefs about the true state, ensuring more accurate credit assignment under noisy observations. Their framework effectively addresses environments where ambiguity stems not from outcomes, but from the mapping between sensory inputs and hidden states. However, the analysis was limited to environments where agents possess probabilistic knowledge about state transitions, and not settings involving strategic interaction or feedback from other agents. In contrast, the current research maintains full observability of the game state but introduces reward uncertainty. By applying standard Q-learning in this setting, the study shows that agents can still converge to near-optimal strategies even without posterior belief updates. This highlights the robustness of Q-learning under a different but practically significant form of environmental noise where outcomes are uncertain, but state perception is not.

Ez-zizi et al. (2023) further distinguished between expected vs. unexpected uncertainty and state vs. reward uncertainty in reinforcement learning. They demonstrated that agents; particularly humans; respond differently to stochastic variability (expected uncertainty) versus sudden, structural changes (unexpected uncertainty), and that reward-based noise is more tractable for learning than ambiguous state conditions. Their experiments focused on behavioural adaptation under uncertainty types in isolated tasks with limited inter-agent dynamics. The present study builds on these findings by embedding expected reward uncertainty into a strategic, multi-stage bargaining environment. Rather than simulating abstract uncertainty conditions, the thesis implements a Reverse Ultimatum Game (RUG) with a stochastic breakdown after each rejection, simulating real-world negotiation risks. Notably, the agents are not informed of the underlying probability of breakdown, yet still learn adaptive policies over time. This complements Ez-zizi et al.'s argument by showing that simple Q-learning mechanisms remain effective under uncertain but structured environments where uncertainty is embedded within adversarial interactions. It adds a novel game-theoretic

dimension to the empirical reinforcement learning literature, extending current understanding of how agents adapt to unpredictable outcomes in strategic settings.

Methodology

The following section highlights the approach utilised to study the equilibrium analysis of the game with an exogenous risk of breakdown and details the use of adaptive agents that form and adjust their strategies in response to experience.

Previous work on artificial agents are modelled as finite automata (Hopcroft and Ulman 1979; Wolfram 1994) to search for robust strategies in repeated games. This framework has been utilised to study a wide array of economic games. Binmore and Samuelson (1992), Sandholm and Crites (1995) and others used it to study the Iterated Prisoner's Dilemma (IPD). Dworman, Kimbrough, & Laing (1995) discovered high quality negotiation policies where the behaviour of the players were presented by finite automata. The current work is built on reinforcement learning (Sutton and Barto, 1998). Previous work on reinforcement learning and bargaining (Fang Zhonga , Steven O. Kimbroughb and D.J. Wu, 2002) showcased the ability of coevolving agents to play the Ultimatum game and confirm the findings with simulation results. Recently, Proaño Mora et all. (2023) modelled biased thinking in artificial agents in the ultimatum game. To the best of current knowledge, this thesis represents the first attempt to model the Reverse Ultimatum Game using Q-learning agents within a reinforcement learning framework.

Game Setup

The study adopts the reverse-ultimatum framework of Gneezy, Haruvy & Roth (2003) and embeds it in an extensive-form, finite-horizon, sequential bargaining game with an exogenous breakdown risk. Consider two players, a proposer (P) and a responder (R), bargaining over a "pie" of $N = 10$ discrete units. Time is discrete and an entire bargaining episode lasts until agreement or collapse. There is no discounting and the only stochastic element is the collapse risk attached to each rejection. In the initial stage P chooses an integer offer $x_t \in \{1, \dots, 9\}$. On observing x_t in any stage $t \in \{1,9\}$ the responder either accepts, ending the game with payoffs $(N - x_t, x_t)$, or rejects. Thus, the action space of responder being $a_R = \{Accepet, Reject\}$. A rejection is followed by two successive moves:

- Exogenous Break-down draw**:** With probability $p \in (0,1)$ the negotiation breaks down and both players receive zero; with probability $1-p$ the bargaining continues. It is stochastic in nature and is drawn independently in every stage.

- Proposer decision**:** If the negotiation continues, the proposer chooses between two actions: $a_p = \{Increase, Stop\}$. Choosing $Stop$ terminates the negotiation with zero payoff to both players. Choosing $Increase$ commits the proposer to a strictly higher offer in the next stage, i.e., $x_{t+1} \geq x_t + 1$ by a minimum increment of a single unit token $(+1)$.

Thus, the game ends and both players earn zero either due to a negotiation breakdown following a rejection, or because the responder rejects the highest possible offer of 9.

Sub-Game Perfect Equilibrium Analysis

The Sub-Game Perfect Equilibrium of this game is achieved through Backword Induction (Muthoo, 1999). For simplicity, let proposer be constrained to propose a strictly higher offer by only a single unit post rejection. Let τ denote the smallest offer that a risk neutral R weakly prefers to rejection. Such that accepting τ is still attractive while accepting $\tau + 1$ is not. Precisely, $Accept$ is better than $Reject$ when:

$$\tau \geq (1-p)(\tau+1) \Longrightarrow p(\tau+1) \geq 1 \qquad (1)$$

Solving (1) yields a threshold

$$\tau \geq (1-p)/p \qquad (2)$$

Given (1) and (2), at any stage t when offered $x_t \in \{1, \dots, 9\}$, the best-response of R:

$$\sigma_R(x_t) = \begin{cases} Reject, & x_t < \tau \\ Accept, & x_t \geq \tau \end{cases}$$

Similarly, the proposer's best response immediately following a rejection of x_t is to increment the offer by one unit, provided the current offer is strictly less than the maximum permissible value. Formally, the best-response strategy of the proposer is simply defined as:

$$\sigma_P(x_t) = \{Increase, \qquad x_t < 9$$

This reflects the fact that, at any stage, choosing to increase the offer always yields a higher payoff than halting the negotiation: $N - x_{t+1} > 0$. Through backward induction, knowing that the responder accepts offers which are greater than the threshold (τ) and reject otherwise. The payoff for proposer decreases in x for every $x \geq \tau$. Furthermore, proposing offers for $x < \tau$, it carries the possibility of materialising into a ruinous outcome by $(1-p)$ post rejection by the responder which makes the expected payoff even lower. Thus, proposer maximises its payoff by choosing exactly the smallest integer satisfying $x \geq \tau$ as the initial offer. Once, τ is realised, any further improvement would only result in reducing the proposer's payoff without altering the acceptance probability. The unique Sub-game Perfect Equilibrium is realised at $(\mathrm{x}^* = \tau, Accept)$.

Consequently, a more realistic view of the Reverse Ultimatum Game is modelled by enabling the proposer to offer concession offers not limiting to a single unit increase upon rejection. In other words, the proposer is permitted to adjust the offer by any strictly positive amount greater than the previous offer, allowing for flexible concessions that better reflect real-world bargaining dynamics where strategic jumps in offers may be made to avoid the risk of breakdown. Though, the responder does not know what the next proposed value would be at the rejection, the relaxation of the constraint does not result a change in the Sub-game Perfect Equilibria and it can be shown through backward induction

Given that a proposer can make a new, strictly higher offer $x_{t+1} \geq x_t + 1$. There is no upper bound on the concession, so the proposer can choose any $x_{t+1} \in \{x_t + 1, \dots, N-1\}$. Then, let x be the smallest offer that satisfies $x \geq (1-p)/p$ as the initial offer. Upon receiving offer x, responder compares the sure payoff from accepting x to the expected payoff from rejecting, which depends on the proposer's strategy over the continuation offers. Since the responder anticipates that upon rejecting an offer x, the proposer will respond with the minimum strictly higher offer of $x' = x + 1$. This is justified because offers greater than $x + 1$ guarantees

acceptance but reduces the proposer's payoff further, it is strictly suboptimal. Hence, the best response of the proposer after rejection is to offer x+1. Anticipating this, the responder finds it optimal to accept the initial offer x rather than reject which yields the inequality $x \geq (1-p)(x+1)$, simplifying to $p(x+1) \geq 1 \Longrightarrow \tau = 1 - p/p$ as in the constrained increase set-up. Hence, the unique subgame perfect equilibrium again features immediate agreement, where the proposer directly offers τ, and the responder accepts at stage one.

Simulation Framework

The simulation design for the bargaining given with exogenous breakdown is composed of minimum 100,000 rounds to ensure agents reach convergence. It is a well-mixed population with 20 agents (equal share of responders and proposers) that results in a sufficient database to identify pattern. The roles are allocated at the start of the simulation and remains same till the end with random interaction after each episode. The value to be split between the two players is 10 units (tokens), which will be the same every round until the end of the experiment. The distribution of 10 units will be in integer numbers to facilitate the extraction of database. Model parameters, ε, endowment and the number of episodes to play etc. can be referred in the input files (GitHub).

Constrained Increase Model

On the framework of reinforcement learning, the following section entails describing about the model and implementation in more detail. Since, Q-Learning is a model-free reinforcement learning with a Markov Decision Process (MDP), which utilises values called Q-values that influence the agents in finding the next best action, given its current position or state it is in (Watkins and Dayan, 1992). Thus, the quality of a particular action (a) in a given state (s):

$$Q(s, a)$$

These Q-values are shown and stored in a Q-table, having one row for each possible state and one column for each possible action, representing an AI agent's optimal policy (π) for acting in the current environment it is in. The Q-values form the value function that estimate the expected return for the agent in a given state. Subsequent section details on rewards set-up,

specific reinforcement learning algorithm that is designed, and finally the code design using Python environment.

Rules (State-Action Pairs): Episodes correspond to independent one-shot bargaining interactions played repeatedly across a well-mixed population. In each episode, a random proposer-responder pair is drawn to play the game. At the beginning of each episode, proposer makes a decision to optimally suggest an initial offer as every subsequent proposed offer is mechanically fixed by single unit increase. The reason to not include the $stop$ in the action space is motivated to simplify the search for optimal policy as the learnt value for stopping the game would be trivially zero. As a result, no parameter is required to denote state which reduces the action set to a one-dimensional lookup table which is how much to offer to responder. The value function is simple:

$$Q(a) \text{ where } a \in [1, N-1]$$

For responder, the action part is either to accept or reject, and the state part is defined to be an integer offer currently on the table by the respective proposer. Hence, it is a two-dimensional array indexed by the offer s:

$$s \in [1, N-1] \text{ and } a \in \{Accept, Reject\}$$

Rewards: Both players receive their respective rewards at the end of each episode. The rewards are updated using the first visit Monte-Carlo approach which form the state-action value function:

$$r_P = \begin{cases} 10 - \mathrm{x}^*, & if\ some\ offer\ \mathrm{x}^*\ is\ accepted \\ 0, & if\ the\ negotiation\ breaks\ down \end{cases}$$

$$r_R = \begin{cases} \mathrm{x}^*, & if\ some\ offer\ \mathrm{x}^*\ is\ accepted \\ 0, & if\ the\ negotiation\ breaks\ down \end{cases}$$

The same reward is traversed into every state–action pair visited by the respective agent during the episode. Since the Monte Carlo method requires the full episode to complete before updates

can be made, the Q-values are calculated as the average return from all previous visits to that state-action pair. Formally, the update rule is:

$$New\ Q(s,a) \leftarrow Q(s,a) \times \frac{(n-1)}{n} + \frac{r}{n}$$

where n denotes the number of times the agent has visited the pair (s, a) and r is the reward obtained at the end of the episode. This formulation is well-suited for environments with uncertain and stochastic rewards, as in the current setup. By updating Q-values using a sample average, early noisy estimates are gradually outweighed by accumulating observations. Over time, this enables convergence toward the true expected values, despite variability in outcomes and exploration.

Reinforcement Learning: Given it is a multi-stage game, there are multiple steps and decision points for each agent in one single episode. The decision-making at each point is according to the values of the suitable Q(s, a) and an ε-greedy policy. With probability 1-ε the agent chooses the optimal policy which is the available action that has the highest Q(s, a) value, and with probability ε the agent chooses randomly and uniformly among the other available actions. The pseudo code for simulation can be found in the Appendix.

Unconstrained Increase Model

Rules (State - Action Pairs): The ability of the proposer to offer higher concession value subject to a minimum increase of single unit (token). It alters the need for the proposer to not only make an active decision at the beginning but also how much to increase ("jump") after each subsequent rejection. Since an agent's policy is constrained on the state (Sutton & Barto, 2018), the proposer's policy must condition on the last rejected offer and the action as the next offer. Furthermore, the decision to make an initial offer does not bore any state dependencies and is limited to action space $a \in [1, N-1]$. The state post each rejection, $s \in [1, N-2]$ and $a \in [s+1, N-1]$

Rewards: The rewards structure remains the same as illustrated in Constrained Increase Model.

Reinforcement Learning: The policy of such reinforcement learning agents remains the same to maximise the rewards earned in the bargaining and minimise the cost of potential breakdown by reducing the occurrence of rejections. Since, the setup of the game enables a final reward to propagate back into each state-action pair. The model utilises implicit path-dependent reasoning (Deng et all, 2023) which penalises the proposer to start below the immediately accepted offer (τ). This approach is inherently more efficient as it does not require to add an additional dimension regarding at which round the decision has taken place. The pseudo code for the same can be referred in the Appendix.

Evolutionary Model of Bargaining with Reputation

The model is constructed with an aim to collect simulation data on bargaining with varied risk tolerance of negotiators. It is a common feature in bargaining to often see the past behaviour of the counterparty such as in wage-labour negotiations (for e.g. Firm and union) (Lerh et al., 2018). Thus, a model is built to showcase, how communication employed through a reputation building exercise between proposers and responders can allow for alternate strategies other than sub-game perfect equilibrium prediction.

In-light of the standard theoretical analysis of the bargaining game *(Unconstrained Increase Model).* An equilibrium threshold is achieved where responder accepts an offer when $Accept \geq E[Reject]$. Under the assumption of complete information, a rational proposer, aware of this condition, makes an initial offer that just meets the responder's acceptance threshold, thereby minimizing the cost of agreement. However, a key limitation of this framework is the lack of heterogeneity in risk preferences across the population. To address this, the model requires a mechanism that allows individual players to reveal their degree of risk tolerance. Reputation serves this role, capturing a more realistic feature of bargaining scenarios where negotiating parties may exhibit diverse risk attitudes. In such settings, a subset of responders may demonstrate greater risk tolerance by committing to a bargaining position in hope of earning higher future rewards as it can be communicated to the proposer. These individuals are often referred to as "tough" responders or hard bargainers (Fanning and Wolitzky, 2020).

Reputation Mechanism

A reputation mechanism governs communication between the two parties, with responders revealing their preferences through a record of past acceptances and rejections. Proposers examine this history before making each offer. Responders who accept only offers above a certain cutoff display this behaviour in their reputation table. In the model, each responder is represented as a hard-wired agent with an acceptance threshold. An agent accepts any offer at or above its private threshold and rejects all others, thereby embedding its risk tolerance directly in its decisions. The responder population is divided into types, each corresponding to either of the following acceptance cutoffs:

$$T_k = \{i : \tau_i = k\}, \qquad k = \{1,2, \dots ,9\}$$

where τ_i is responder i's acceptance threshold. The goal of a proposer is to learn to make an offer that is immediately accepted by analysing the table of past outcomes of a respective responder. The history that each proposer can observe is limited to the six most recent outcomes. This restriction is intended to reflect real-world conditions, where records of counterparty behaviour are often imperfect, fragmented, or intentionally withheld.

Henceforth, proposers are defined as rule-based agents that use a deterministic policy derived from each responder's recent six outcomes rather than updating through reinforcement learning. The policy-function takes input of a six-row history table where each row consists of an integer offer previously made and the corresponding outcome ("Accepted" or "Rejected"). This table at time t is denoted as:

$$H_t = [(x_{t-6}, o_{t-6}), (x_{t-5}, o_{t-5}), \dots , (x_{t-1}, o_{t-1})],$$

where each $o_i \in \{A, R\}$ and $x_t \in \{1, .. ,9\}$. The policy $\pi(H_t)$ first inspects whether all six outcomes are "A" or all are "R." If every row ends in "A," it sets the next offer x_{t+1} equal to one unit below the smallest x_t in which $o_i = A$, ensuring maximal surplus extraction while still winning acceptance. If every row ends in "R," it sets x_{t+1} to one unit above the highest offer with $o_i = R$, seeking to overcome the responder's demonstrated toughness. Furthermore, in a mixed history scenario the proposer looks back over past rounds to identify the highest

offer that was rejected and the lowest offer that was accepted. The proposer then makes its next offer one unit above the highest rejected amount unless that would exceed the lowest accepted amount, in which case it offers exactly the lowest accepted amount. Figure-1 depicts the flowchart of the policy function.

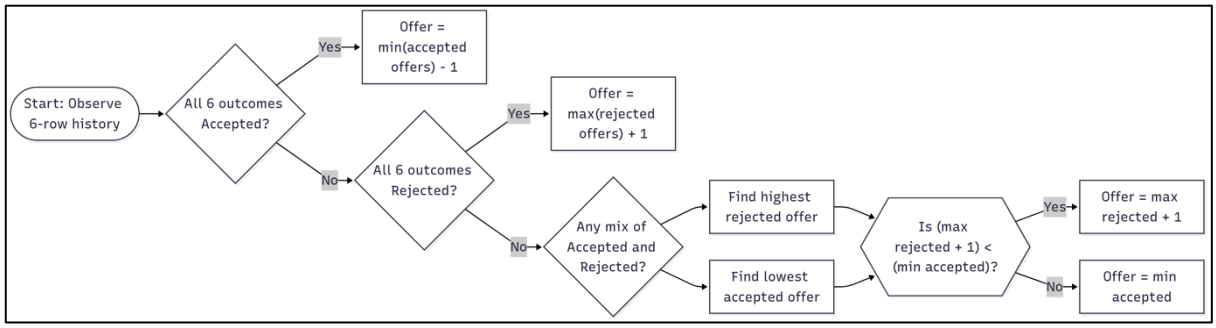

Figure-1: Flowchart representing the deterministic policy function used by proposers, mapping observed six-period responder history to the subsequent offer decision in the bargaining game.

This dynamic response undertaken by the proposers aims to minimize the risk of breakdown by ensuring immediate acceptance rather than running into the risk of materialising a ruinous outcome. Although this approach enables proposers to secure a guaranteed reward, it can significantly influence the earnings of responders. It may be beneficial for responders to maintain a tough stance and accept lower rewards in the early interactions, as building a reputation can eventually result in higher future rewards. Importantly, the results from the model can provide insights into the evolutionary stability of the strategy in which agreements are reached at the established equilibrium threshold as predicted by standard analysis, and whether this strategy remains robust against mutant strategies that involve offering and accepting amounts above the equilibria.

Evolutionary Selection

Since, the aim of this model is to see evolutionary emergence of "tough" responders who signal their toughness through a visible history of outcomes accessible to the proposer before each offer. A selective pressure is applied exclusively to the responder population at the completion of each generation, defined as a series of interactions. At each generation, a responder's fitness is measured by their average reward, calculated as the total accepted payoff divided by the number of interactions completed. These fitness scores are normalized across the responder population to determine selection probabilities for replication. In every generation, a fixed

fraction (α) of responders are selected uniformly at random for replacement. Each replaced agent is probabilistically assigned a new type, with the probability of becoming type T_k proportional to mean reward R_k achieved by that type in the current generation, according to:

$$P_k = \frac{R_k}{\sum_{j \in K} R_j}$$

where:

- P_k is the probability that a replaced agent becomes a type T_k,
- R_k is the average reward earned by responders of type T_k in the current generation,
- K is the set of all responder types

This formulation ensures in enforcing an evolutionary pressure towards reward-maximizing strategy. To maintain population heterogeneity and promote exploration, a mutation mechanism operates at each generation for both proposer and responder populations. Among responders, mutation results in a shorter history window and random changes to acceptance thresholds, creating agents who display a range of risk preferences.

Proposers, on the other hand, do not undergo an evolutionary selection. The decision is motivated by first, as the primary aim of the model is to examine the evolutionary stability of responder strategies, particularly whether rejecting offers at the predicted equilibrium threshold remains effective when alternative behaviours emerge. Second, the optimal strategy for proposers is simply to make an offer that is likely to be accepted right away, which minimizes the risk of negotiation breakdown. Nonetheless, to maintain variety within the proposer group, a fraction of proposers are assigned as mutants in every generation. These proposers ignore responder history and start with random offers ranging from [1,9], which contributes to a more dynamic bargaining environment. All algorithmic details for these processes are available in the Appendix.

Results

The artificial agents play a series of reverse ultimatum game, first the repeated multi-stage game with constrained increase, second against unconstrained increase version.

Reverse ultimatum game with Incremental increase

Figure-2 shows the average Q-table for all proposers and responders in the population at the end of 1-Million episodes. At a breakdown risk of 0.18, the sub-game perfect equilibrium is depicted as: $\tau = 1 - 0.18/0.18 \Rightarrow \tau = 4.55$.

Average Proposer Q-Table

	offer	Q-Value
0	1.00	3.95
1	2.00	3.98
2	3.00	4.35
3	4.00	4.73
4	5.00	4.91
5	6.00	3.90
6	7.00	2.88
7	8.00	1.90
8	9.00	0.92

Left Figure-2(a): Proposer Q-Table with highest value learnt for an offer of 5.

Average Responder Q-Table

	Accept	Reject
1	1.00	2.13
2	2.00	2.68
3	3.00	3.30
4	4.00	4.08
5	5.00	4.92
6	6.00	5.69
7	7.00	6.09
8	8.00	6.05
9	9.00	0.00

Right Figure-2(b): Responder Q-table showing rejection valued higher below offer 5, with acceptance preferred from 5 onwards.

Hence, responders are expected to reject offers below 5 and accept offers at 5 and above. Similarly, proposers are expected to propose 5 as the initial offer. As depicted, the Q-Tables in Figure-1 for respective players confirm with the theoretical equilibria. Figure-3 shows the evolution of initial offers made by the proposer along with offers accepted by the responder through histogram for each episode. For the first 10000 episodes, agents randomly chose offers which is primarily determined by high exploration rate. As number of episodes progress along with decaying exploration rate, proposers learn to offer 5 as the initial offer which immediately gets accepted.

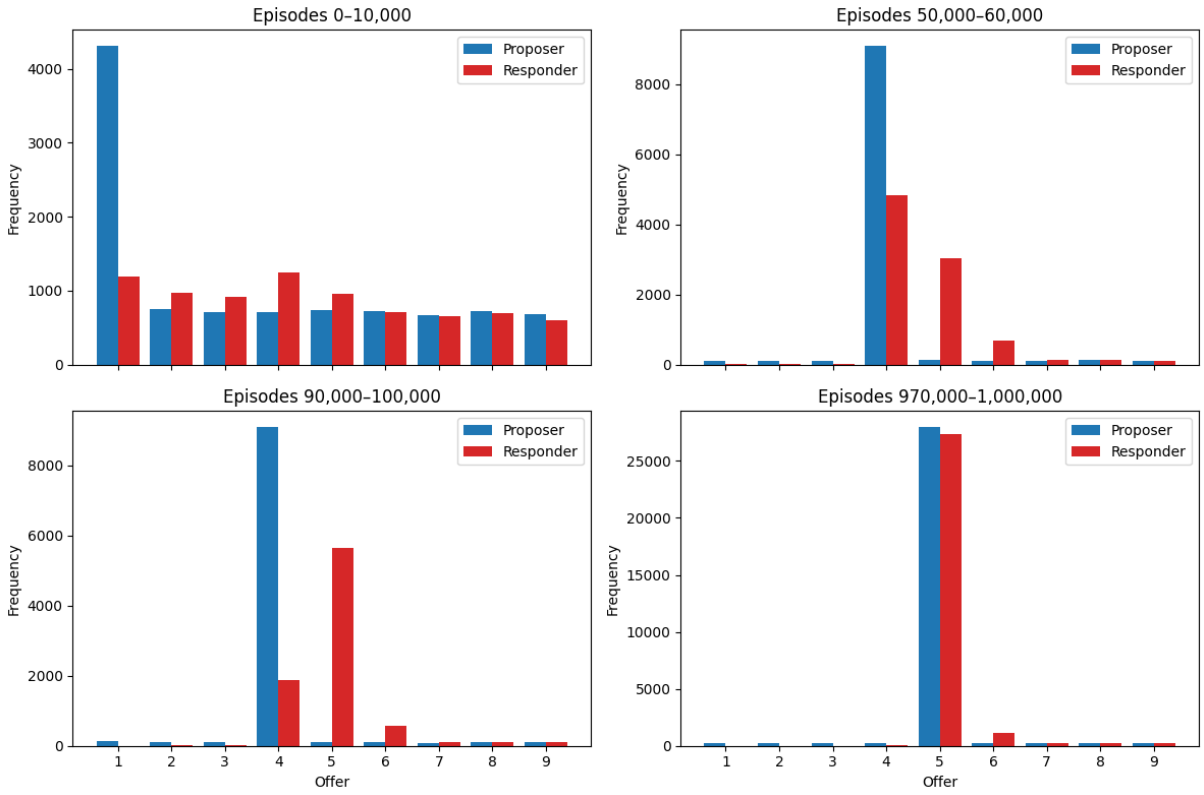

Figure-3: Initial offers by proposers and accepted offers by responders across selected episode intervals.

Figure-4 illustrates the converges of payoffs over 1M episodes. It depicts convergence of the payoffs by each player to the perfect sub-game equilibrium of (5,5) where initial proposed offer of 5 gets immediately accepted by the responders in the first stage.

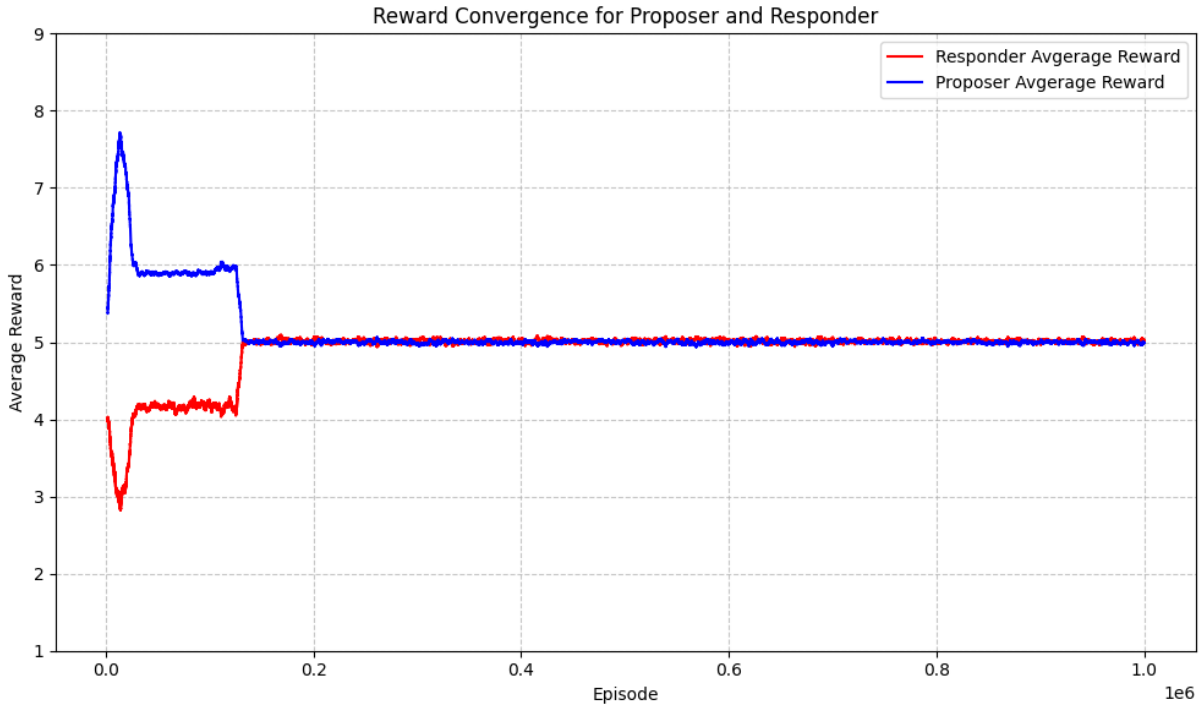

Figure-4: Rolling average (2,000 episodes) of proposer and responder rewards over 1 million rounds.

Reverse ultimatum game with Unconstrained Increase

Figure-5 shows the average Q-table for all proposers and responders at the end of 100000 episodes. Though proposers are allowed to make a subsequent offer post each rejection higher than a single unit increase than the last offer made but it does not change the equilibrium at a breakdown risk of 0.18. Thus, proposers realise the minimum acceptable amount by responders.

Average Proposer Q-Table

	offer	Q-Value
0	1.00	4.51
1	2.00	4.47
2	3.00	4.43
3	4.00	4.65
4	5.00	4.97
5	6.00	3.87
6	7.00	2.81
7	8.00	1.83
8	9.00	0.87

Left Figure-5 (a): Proposer Q-Table with highest value learnt for an offer of 5.

Average Responder Q-Table

	Accept	Reject
1	1.00	2.73
2	2.00	3.13
3	3.00	3.60
4	4.00	4.05
5	5.00	4.88
6	6.00	5.44
7	7.00	6.01
8	8.00	5.20
9	9.00	0.00

Right Figure-5(b): Responder Q-table showing rejection valued higher below offer 5, with acceptance preferred from 5 onwards.

Proposers learn to open the bargaining with an initial offer of 5 which immediately gets accepted (Figure-6).

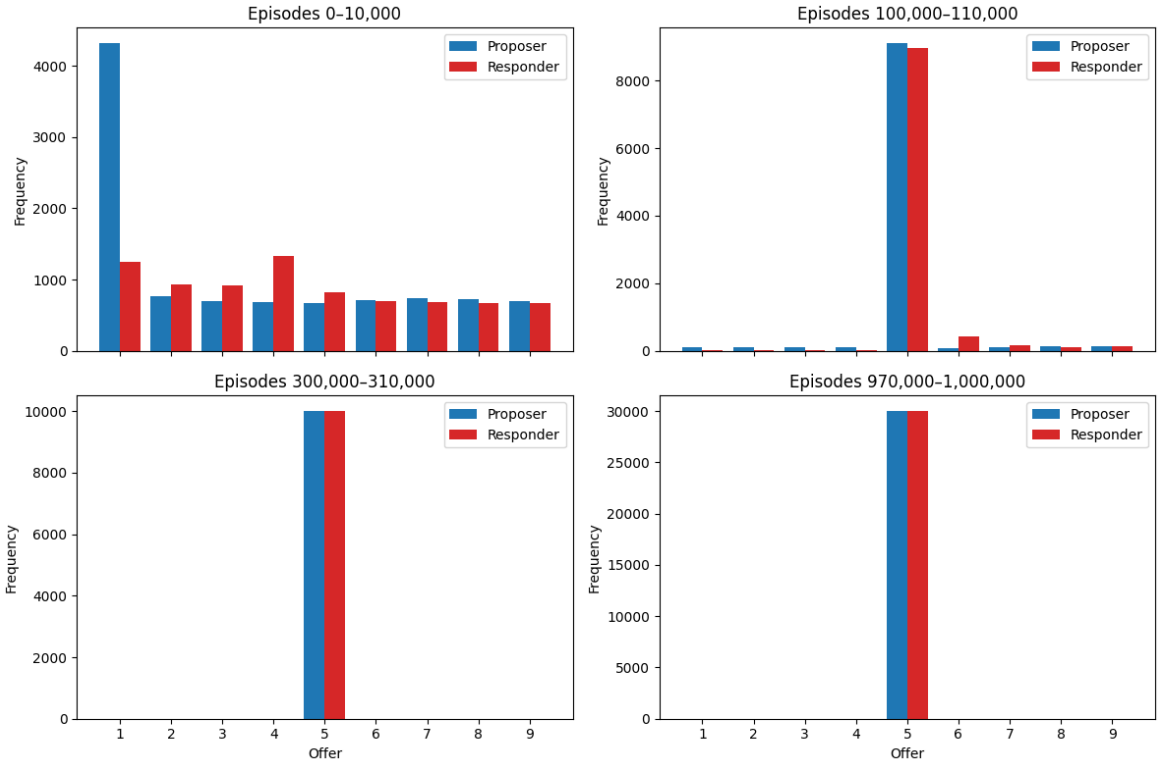

Figure-6: Initial offers by proposers and accepted offers by responders across selected episode intervals.

The average payoffs of responder and proposers also converge to the subgame perfect equilibrium (Figure-7) for 1M episodes run.

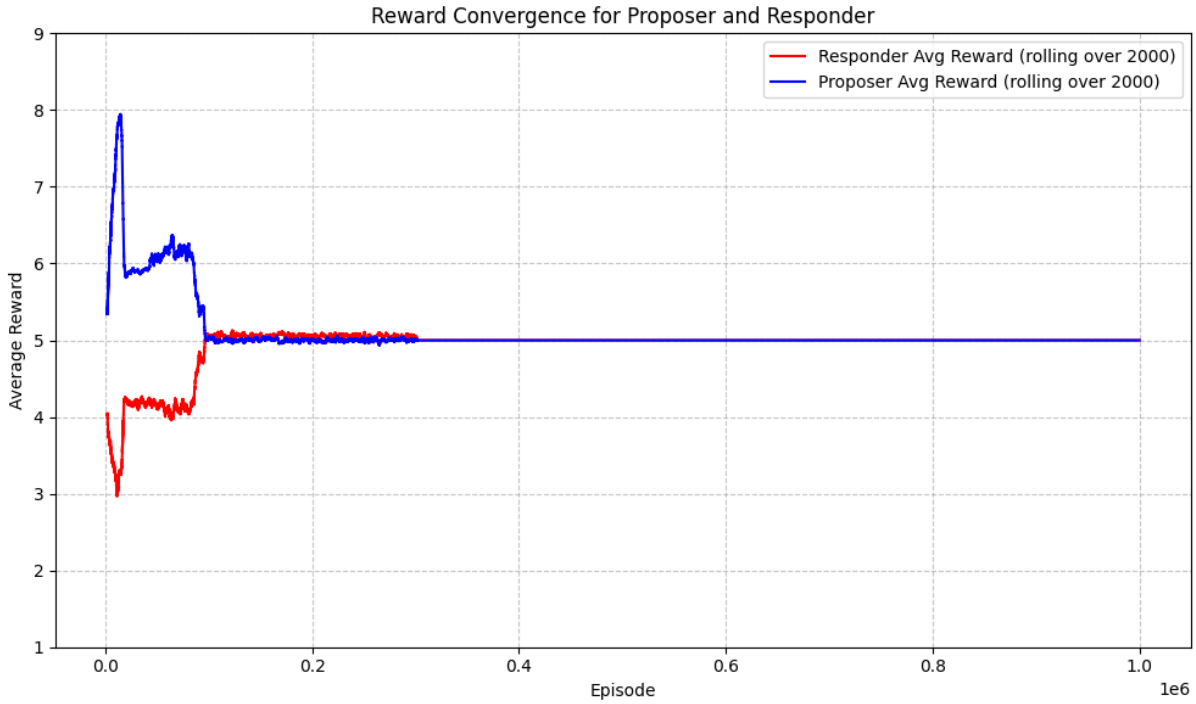

Figure-7: Rolling average (2,000 episodes) of proposer and responder rewards over 1 million rounds

The results show the ability of co-evolving agents to learn the optimal policy under reward uncertainty, where stable equilibrium strategy emerges. Responders, through Q-learning, accurately estimate the value of rejection and eventually adopt a fixed acceptance threshold, τ, below which all offers are rejected. Once this threshold stabilizes, proposers also simultaneously learn to begin bargaining with τ, knowing that any less favourable offer would be rejected. Thus, the interaction dynamics naturally converge to a rational outcome where proposers open with τ *(subgame perfect equilibrium)*, reflecting both players' adaptation to long-run expected payoffs.

Classical Reverse Ultimatum Game with No Deadline

The results provide insights into the ability of these reinforcement learning agents to learn to play the traditional reverse ultimatum game without deadlines with a single responder. Given that the risk of breakdown is set to zero and the proposer is allowed to make an offer as long as the offer is strictly higher than the minimum increment (1 token) post rejection by the

responder. The subgame perfect equilibria of a classical reverse ultimatum game predicts one token to the proposer and the remainder (N − 1 tokens) to the responder.

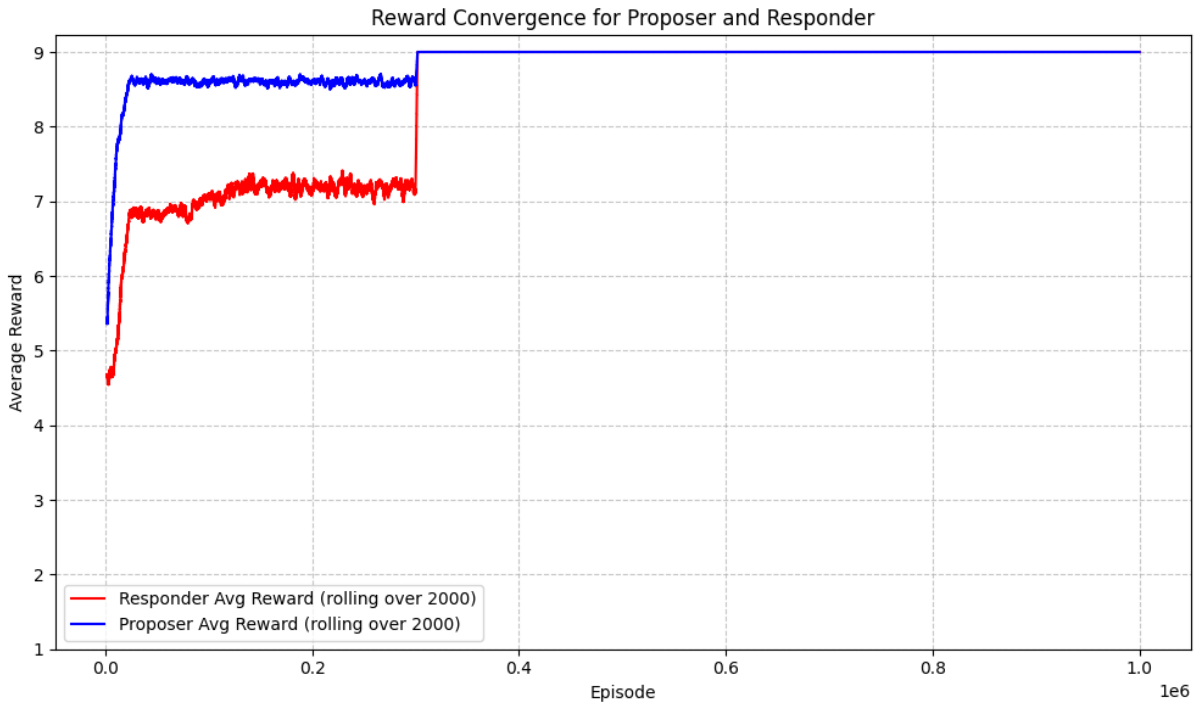

Figure-8: Rolling average (2,000 episodes) of proposer and responder rewards over 1 million rounds.

The above Figure-8 shows the average rewards for proposers and responders over multiple episodes. At the end of 300000 episodes when the exploration rate is set to zero i.e. agents always chose the optimal policy, the convergence parallels to the subgame perfect equilibria of (1,9). Figure-9 shows the average learnt Q-table for all proposers and responder in the population.

Average Proposer Q-Table

	offer	Q-Value
0	1.00	1.52
1	2.00	1.41
2	3.00	1.31
3	4.00	1.24
4	5.00	1.17
5	6.00	1.12
6	7.00	1.09
7	8.00	1.06
8	9.00	0.99

Left Figure-9 (a): Proposer Q-Table for offers.

Average Responder Q-Table

	Accept	Reject
1	1.00	8.53
2	2.00	8.62
3	3.00	8.70
4	4.00	8.76
5	5.00	8.81
6	6.00	8.84
7	7.00	8.86
8	8.00	8.92
9	9.00	0.00

Right Figure-9(b): Responder Q-table for actions associated with each offered split.

Responders have successfully achieved to learn to reject all the offers leading to 9 and accept always. Consequently, proposers Q-table also confirms the story with an average expected

payoffs for all offers averaging to 1. The spike in early offers is due to the exploration by the responder to play a sub-optimal policy (accepting offers below 9) in the early rounds which caused proposer to receive larger payoffs than 1. Though, decaying exploration rate causes q-value to converge to 1 but due to sample average updating, the new payoffs get updated slowly due to lower weights to smoothen early noise.

Distribution of Offers with Varying Risk of Breakdown

The simulation can provide insight into the question: How does the varying of breakdown risk affects distribution of offers in a bargaining? Figure-10 shows the distribution of average payoffs for agents at varying risk of breakdown.

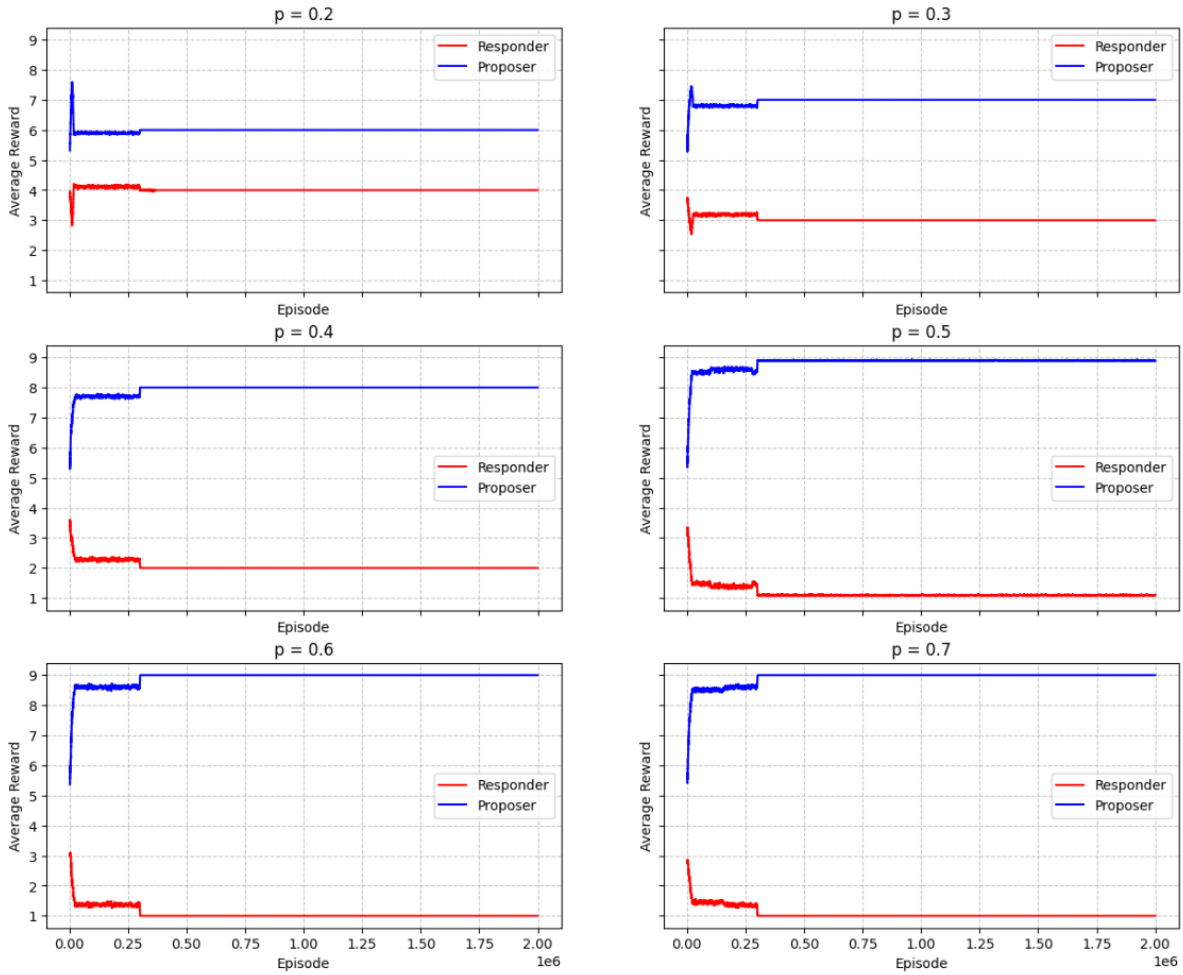

Figure-10: Average rewards for proposers and responders with varying exogenous risk of breakdown.

For higher risk of breakdown in bargaining, the equilibrium converge towards low offers made by the proposer as the associated cost of rejecting is high for responders. However, lower risk

of breakdown favours responder with higher average payoffs earned. As depicted, the vertical difference between the average payoffs for proposer and responder increases with breakdown probability. Furthermore, the work by Herrera et al. (2025) on bargaining reveals that proposition on equilibrium reaches early with increasing breakdown chance (h). The work here, confirms with the finding that a high breakdown risk of greater than or equal to 0.5, it is always better for responders to accept immediately which in turn allows proposer to offer the split (9,1) and convergence occurs soon enough (within first 25000 interactions) that matches with the theoretical SPE:

$$\tau = \frac{1 - 0.5}{0.5} \Rightarrow \tau = 1 \geq 1 \; \forall p \geq 0.5$$

Evolutionary Model of Bargaining with Reputation

The observed results for this section builds on the standard theoretical analysis of the multi-stage bargaining model with the exogenous risk of breakdown without reputation. The chosen probability of breakdown inherited in the environment is at $p^* = 0.18$. Henceforth, the achieved equilibrium threshold, as predicted by the standard theoretical analysis defined through backward induction is at $\tau^* = 5$ in the environment (For how this is achieved, please refer to the section: SPE Analysis). Here, proposer without any form of communication offers the minimum acceptable amount $x_1^* = 5$ that immediately gets accepted at period 1.

Building on this analysis, the initial responder population is composed of agents with heterogeneous risk profiles. Seventy per cent of responders are risk neutral with acceptance threshold $\tau^* = 5$, while the remaining thirty per cent is divided equally among three "tough" types. We define responder types $\mathrm{T_k}$ for $k \in \{5,6,7,8\}$, where each agent in T_k accepts only offers of at least k tokens. The type proportions are: $P(T5) = 0.7, P(T6) = P(T7) = P(T8) = 0.1$. On the proposer side, agents apply the rule based policy to the most recent six outcomes of each responder (see methodology).

Each generation consist of 600 interactions and the simulation is run over 2000 generations to gather sufficient evidence. The chosen proportion of agents marked randomly for replacement is considered $\alpha = 30\%$ at each generation. Through the probabilistically replacement, a new

responder is born to have its type by probabilistically duplicating with the probability that a new responder takes on a specific type $T_k \in \{T5, T6, T7, T8\}$ is proportional to the average performance (reward) of the respective type in the current generation. Formally, the selection probability is defined as:

$$P_k = \frac{R_k}{\sum_{j=5}^{8} R_j}$$

where:

- P_k is the probability that a replaced agent becomes a type T_k,
- R_k is the average reward earned by responders of type T_k in the current generation,
- $\sum_{j=5}^{8} R_j$ is the sum of average rewards across all responder types.

The mutation is fixed with a small probability at (5%) for both proposers and responders at each generation. Figure-11 displays the observable reputation histories for a sample of 25 responders at generation 25. Each row represents a responder's most recent interactions, where the entries capture the offer made and the corresponding response. These histories inform proposer strategy

Responder	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Responder 33	(6, 'R')	(1, 'A')	(9, 'R')	(7, 'A')	(6, 'A')	(9, 'A')
Responder 80	(4, 'A')	(5, 'A')	(9, 'A')	(9, 'R')	(4, 'R')	(8, 'R')
Responder 72	(9, 'A')	(4, 'R')	(1, 'A')	(7, 'R')	(2, 'R')	(8, 'A')
Responder 55	(8, 'R')	(9, 'A')	(8, 'R')	(1, 'R')	(5, 'R')	(2, 'R')
Responder 73	(4, 'A')	(9, 'A')	(6, 'A')	(1, 'A')	(8, 'A')	(2, 'R')
Responder 8	(9, 'A')	(8, 'A')	(1, 'R')	(5, 'A')	(3, 'R')	(1, 'R')
Responder 62	(5, 'R')	(6, 'A')	(5, 'R')	(7, 'R')	(8, 'A')	(5, 'A')
Responder 85	(1, 'A')	(5, 'A')	(1, 'A')	(9, 'A')	(8, 'A')	(6, 'R')
Responder 0	(7, 'R')	(9, 'A')	(5, 'A')	(7, 'A')	(9, 'A')	(2, 'R')
Responder 23	(8, 'A')	(9, 'A')	(3, 'A')	(8, 'A')	(4, 'R')	(6, 'R')
Responder 50	(5, 'R')	(1, 'A')	(3, 'R')	(5, 'A')	(5, 'A')	(1, 'A')
Responder 92	(1, 'A')	(2, 'R')	(6, 'R')	(2, 'R')	(8, 'A')	(2, 'A')
Responder 98	(1, 'A')	(7, 'R')	(9, 'R')	(5, 'R')	(7, 'R')	(3, 'A')
Responder 6	(9, 'A')	(3, 'R')	(3, 'R')			
Responder 19	(2, 'A')	(3, 'A')	(6, 'R')	(2, 'R')	(2, 'R')	(2, 'R')
Responder 70	(6, 'R')	(4, 'R')	(6, 'A')	(7, 'A')	(5, 'A')	(4, 'A')
Responder 39	(4, 'R')	(1, 'A')	(3, 'R')	(2, 'R')	(1, 'R')	(3, 'R')
Responder 41	(7, 'R')	(5, 'R')	(7, 'A')	(2, 'R')	(2, 'R')	(7, 'R')
Responder 66	(7, 'A')	(1, 'R')	(9, 'A')	(2, 'A')	(5, 'R')	(8, 'R')
Responder 84	(8, 'R')	(1, 'A')	(3, 'A')			
Responder 1	(2, 'R')	(4, 'R')	(8, 'R')	(7, 'R')	(9, 'R')	(8, 'A')
Responder 31	(8, 'R')	(3, 'R')	(8, 'R')	(8, 'R')	(4, 'R')	(5, 'R')
Responder 35	(6, 'R')	(3, 'A')	(1, 'A')	(1, 'R')	(9, 'A')	(1, 'A')
Responder 38	(1, 'A')	(8, 'R')	(5, 'A')	(5, 'A')	(6, 'A')	(8, 'A')
Responder 77	(9, 'A')	(8, 'R')	(6, 'A')	(5, 'R')	(8, 'R')	(1, 'A')

Figure-11: Reputation tables for 25 responders in generation 25 showing their six most recent offer-outcome pairs. Shorter histories indicate new or mutated responders.

Figure-12 depicts the evolution of responder acceptance thresholds over 2000 generations. Responders initially requiring a minimum offer of 5 comprised only 10 % of the population, yet by generation 500 those demanding a minimum of 8 exceed 80 %. The system then undergoes transient fluctuations: around generation 750, responders with a threshold of 7 briefly rise in prevalence, but they are quickly outcompeted by higher risk-tolerant responders of accepting offers 8 and above, who consistently secure higher payoffs. Hence, responders with the highest acceptance threshold prove evolutionarily stable despite occasional perturbations, they sustain a majority (over 50 %) at each generation

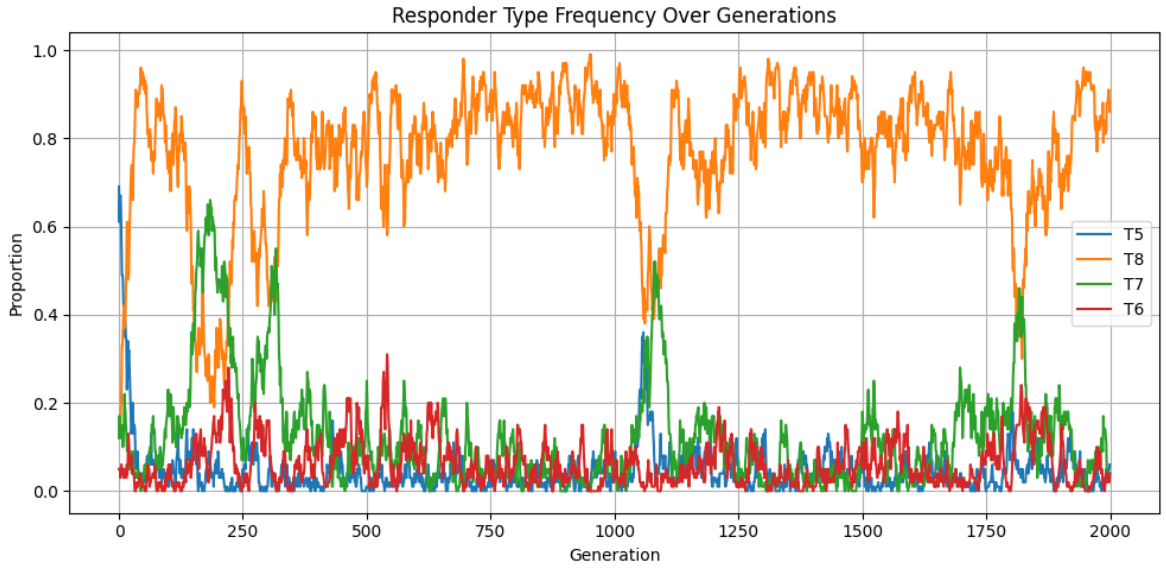

Figure-12: Shows the proportion of responders over varied acceptance threshold of offers over 2000 generations

Discussion & Limitations

Sutton and Barto (2018) introduced reinforcement learning as a computational model of how actors (or agents) can learn to make optimal decisions through repeated experiences. The results from the thesis shows that artificial agents using reinforcement learning have been found to be capable of playing the reverse ultimatum game efficiently and effectively. It further extends to showcase the capability of Q-learning to learn the optimal policy which is fundamental to reinforcement learning algorithm in the domain of uncertain environments with inconsistent rewards. The work adds to the existing literature on Q-learning algorithm in economic games of importance such as repeated public goods game (Chenna Reddy Cotla, 2015); impact of group size on cooperation (Kazuhiro Tamura and Satoru Morita, 2024), ultimatum game with behavioural biases (Añasco Flores, 2023); (Zhong et al, 2002), "stateless" agents where actions are not conditioned on memory in prisoners dilemma (Arthur Dolgopolov, 2024), stag-hunt game (Bilancini and Boncinelli, 2020).

Given the role of reinforcement learning in game theory, its recent development for analysing varied situation through agent run simulation in Traffic management and control (vehicular and network), e-commerce application, cyber security. There is an existing research gap of creating machine learning algorithms that can replicate and represent more human-like surroundings in game theory settings which this work aids to fill the gap by contributing to labour negotiations. This research also adds to the Multi-Agent Reinforcement Learning (Ning and Xie, 2024) where multiple agents either receive insufficient or excessive information from their environment and simultaneously adapt to the behaviour of others in a non-stationary environment. Since, here the policy of each agent evolves over time due to implicit modelling of exogenous breakdown probability, it results into non-stationary environment in the context of sequential bargaining unlike standard matrix games or cooperative coordination settings.

The ability of such agents to learn the optimal policy and replicate the finding with the theoretical prediction illustrates the successful implementation of the reward structure. This is achieved within a Monte Carlo framework, which, given sufficient observations, is able to approximate the probability distribution of possible outcomes in an environment characterized by uncertainty. The results from this research suggest that, in settings where the continuation of interaction is uncertain due to probabilistic breakdown, the Monte Carlo method enables agents to learn in real time. This is particularly applicable in environments like personalized

online platforms, strategic negotiation games, and decision-support systems, where repeated episodes can be simulated or observed. The results also compounds with the existing literature of using Monte-Carlo algorithm to investigate approaching for treating uncertainty and creating risk-sensitive decision maker (Jorge, 2023). To further complement the ability of reinforcement learning, the results show that when placed in an environment, the agents defined under the framework can eventually converge to play sub-game perfect equilibrium. This can be shown through success of replicating the equilibrium for a classical reverse ultimatum game. This model can be utilised by studies to add more nuances and extrapolate simulation data through these agents.

The literature on standard game theoretical analyses of strategic bargaining models (e.g., inter alia, Muthoo 1999) are based on the concept of subgame perfect equilibria. Such analyses prescribe a strategy for each party such that agreement is achieved at time zero (or approximately). One party starts the negotiation by making the equilibrium offer and this offer is immediately accepted by the other party. This research draws a similar approach to find a subgame perfect equilibria of a strategic multi-stage bargaining game where breakdown looms over which does not evolve endogenously and is determined by exogenous forces which lies outside the control of parties engaged in bargaining. The theoretical analysis reveals the emergence of an acceptance threshold that structures the equilibrium outcome of the game. This threshold corresponds to the minimum offer that is guaranteed to be accepted by the responder, thus marking the lowest concession a proposer can make without risking breakdown. In equilibrium, the proposer offers precisely this threshold value, which is immediately accepted, leading to agreement without delay. Despite the possibility of multiple rounds, the game's structure implies that rational agents prefer to avoid sequential concessions due to the risk of breakdown at each rejection.

Importantly, the position of this threshold is sensitive to the exogenous risk of breakdown. As the probability of breakdown increases, the equilibrium threshold shifts downward. This dynamic enhances the strategic advantage of the proposer (or first mover), who can secure a greater share of the surplus by exploiting the responder's aversion to breakdown. Conversely, when the risk of breakdown is low, responders are more inclined to reject low offers, anticipating better terms in future rounds. This mirrors the findings in Herrera et al. (2025), where lower breakdown probabilities induce greater brinkmanship and delay, as negotiators seek to improve their payoff by leveraging the low urgency of collapse. Moreover, the analysis

underscores that the possibility of incremental concessions does not alter the equilibrium outcome. Even in settings where proposers can improve their offers gradually after rejection, the dominant strategy remains to offer the equilibrium threshold at the outset.

Additionally, this study examined whether it is beneficial for responders to reject offers at the equilibrium threshold when there is indirect communication through observable history. The simulation introduced a subset of risk-seeking responders who consistently reject offers that would be accepted under standard equilibrium analysis. Despite exposing themselves to a higher probability of breakdown, these responders established a credible commitment to demand better offers. Over time, proposers adapt to this behaviour by offering more generous shares to avoid rejection and the associated losses. Although offering less would yield a higher immediate payoff, it carries the risk of breakdown, which results in a strictly worse outcome for the proposer. It remains preferable to secure a smaller payoff than to receive nothing. This dynamic creates a reputational feedback loop where acting tough becomes advantageous. Even though these responders initially form only a small portion of the population, their strategy yields higher average payoff, leading to an increase in their share across generations. The outcome shows that when proposer agents have access to responder history, reputation serves as an implicit form of communication. In such environments, rejecting at the threshold is not necessarily irrational. Instead, it can be a strategically sound approach that reshapes proposer expectations and shifts the equilibrium in favour of the responder. The persistence and dominance of this strategy under evolutionary pressures suggest that toughness, when consistently signalled and observed, can be rewarded even in high-risk settings.

Limitation

The assumption that the breakdown probability remains exogenous and constant may be relaxed in future work by allowing this risk to evolve endogenously or vary over time. While this simplification is analytically tractable, it restricts the model's realism in capturing how risk evolves in real-world negotiations. In political or institutional bargaining such as legislative budget negotiations (Agranov and Tergiman, 2014) or international treaty talks (Fearon, 1998) the risk of breakdown is often shaped dynamically by prior actions, shifting public constraints, or institutional deadlines. Future work could extend the model by allowing the probability of breakdown to evolve endogenously as a function of negotiation history or behavioural patterns,

enabling richer modelling of perceived brinkmanship or institutional escalation, see Wilko Bolt and Alexander F. Tieman (2006) for reference.

Another limitation of this work inculcates the slow convergence of Q-values in uncertain environment as the agents require sufficiently long observations (1-Million or more) to form value functions for state-action pair. Future research could employ other methods such as Moment Matching Offline Model-Based Policy Optimization (MOMBO) for faster convergence (Kandemir et al., 2025). It could also shed light on identifying the right balance between exploration and exploitation so agents can safely exploit good strategies without becoming trapped in suboptimal policies. Comparative work could evaluate techniques such as intrinsic reward schemes, decaying exploration rates to clarify trade-offs among speed of learning, safety, and robustness.

Additionally, the model assumes fixed population structures and dyadic interactions between proposers and responders. It does not account for the possibility that strategic behaviour may be influenced by networked interactions or reputational spillovers across multiple negotiations. As demonstrated in Calabuig and Olcina (2000), the presence of repeated interactions and belief-based commitment types can significantly alter the bargaining outcomes. Future work could embed the current framework within a network or spatial setting to examine whether clusters of "tough" or "weak" responders emerge and persist based on their structural positions and local interaction patterns.

As the simulation tracks the evolutionary persistence of tough responder types and their payoff consequences, it does not disentangle the cognitive mechanisms through which agents process stochastic outcomes and update strategies under uncertainty. Recent findings in behavioural reinforcement learning (Larsen et al., 2010; Ez-zizi et al., 2023) emphasize the role of ambiguity aversion, misperception of stochasticity, and non-Bayesian learning in complex environments. Incorporating such psychological noise or heterogeneous updating rules could provide deeper insights into why agents deviate from optimality in practice and how such deviations influence the macro-level equilibrium dynamics observed in bargaining systems.

Conclusion

This study investigated a reverse ultimatum game that incorporates a risk of breakdown at each stage when no concession is reached because of external factors. The standard theoretical analysis, solved by backward induction, revealed a subgame perfect equilibrium characterised by a threshold offer that the responder accepts immediately. Guided by this result, the research employed Q learning agents that imitate human reasoning in uncertain settings to generate simulation data and examine the strategies adopted by such players. The agents converged to the subgame perfect equilibrium, confirming the theoretical prediction. Their play in the classical reverse ultimatum game also generates benchmarks that can be used to compare human behaviour with fully rational behaviour and to evaluate possible experimental variants of the game. Understanding this multi stage bargaining environment with exogenous breakdown has practical value beyond its scientific interest. In electronic commerce, virtual markets frequently require autonomous agents to negotiate while facing a constant possibility of negotiation failure. The findings here offer initial guidance for the design and management of artificial agents that must bargain over divisions of surplus in such contexts.

The research then extended the analysis to reputational bargaining by introducing heterogeneous risk profiles in the responder population. Each responder was assigned an individual acceptance threshold, recorded through a publicly observable history of past outcomes. Simulation results showed that responders who adopt a tough stance by rejecting low offers can induce higher offers from proposers and earn superior long-run payoff. In the evolutionary model, responders that committed to the highest share built the strongest reputation and eventually dominated mutant populations with lower risk tolerance. These outcomes highlight the strategic importance of reputation when negotiations can break down.

Overall, the study contributes to the empirical literature at the intersection of game theory, experimental economics, behavioural economics, and artificial intelligence. By combining theoretical analysis with agent based simulations, it provides new evidence on how learning and reputation interact in bargaining under uncertainty and offers a foundation for further work on autonomous negotiation in economic environments.

References

Agranov, M., & Tergiman, C. (2014). Communication in multilateral bargaining. Journal of Public Economics, 118, 75–85. https://doi.org/10.1016/j.jpubeco.2014.06.006

Akgül, A., Haußmann, M., & Kandemir, M. (2025). Deterministic uncertainty propagation for improved model-based offline reinforcement learning. arXiv. http://arxiv.org/abs/2406.04088

Amato, C. (2025). An initial introduction to cooperative multi-agent reinforcement learning. arXiv. http://arxiv.org/abs/2405.06161

Añasco, J., Naranjo Navas, B. J., Proaño Mora, P. A., & Vasileuski Kramskova, M. A. (2023). Simulation of ultimatum game with artificial intelligence and biases. ACI Avances en Ciencias e Ingenierías, 15(1). https://doi.org/10.18272/aci.v15i1.2304

Backus, M., Blake, T., Larsen, B. J., & Tadelis, S. (2020). Sequential bargaining in the field: Evidence from millions of online bargaining interactions. Quarterly Journal of Economics, 135(3), 1319–1361.

Beattie, J., Baron, J., Hershey, J. C., & Spranca, M. D. (1994). Psychological determinants of decision attitude. Journal of Behavioral Decision Making, 7, 129–144.

Binmore, K., & Samuelson, L. (1992). Evolutionary stability in repeated games played by finite automata. Journal of Economic Theory, 57(2), 278–305.

Bolt, W., & Tieman, A. F. (2006). On myopic equilibria in dynamic games with endogenous discounting. (Journal unknown).

Calabuig, V., & Olcina, G. (2000). Commitment and strikes in wage bargaining. Labour Economics, 7(3), 349–372.

Cázares, J. I. G., & Mijatović, A. (2022). Monte Carlo algorithm for the extrema of tempered stable processes. https://doi.org/10.1017/apr.2023.1

Carpenter, J. (2003). Altruistic behavior in a representative dictator experiment. Journal of Economic Behavior & Organization, 51(1), 27–45.

Cotla, C. R. (2015). Learning in repeated public goods games: A meta-analysis. SSRN. https://doi.org/10.2139/ssrn.3241779

De Jong, S., Uyttendaele, S., & Tuyls, K. (2008). Learning to reach agreement in a continuous ultimatum game. Journal of Artificial Intelligence Research, 33, 551–574.

Dias, L. C., & Vetschera, R. (2022). Two-party bargaining processes based on subjective expectations: A model and a simulation study. Group Decision and Negotiation, 31(4), 843–869.

Dolgopolov, A. (2024). Reinforcement learning in a prisoner's dilemma. Games and Economic Behavior, 144, 84–103. https://doi.org/10.1016/j.geb.2024.01.004

Dworman, G., Kimbrough, S. O., & Laing, J. D. (1995). On automated discovery of models using genetic programming: Bargaining in a three-agent coalitions game. Journal of Management Information Systems, 12(3), 97–125. https://doi.org/10.1080/07421222.1995.11518093

Ennio Bilancini, & Boncinelli, L. (2020). The evolution of conventions under condition-dependent mistakes. Economic Theory, 69(2), 497–521.

Fang Zhong, Kimbrough, S. O., & Wu, D. J. (2002). Cooperative agent systems: Artificial agents play the ultimatum game. Group Decision and Negotiation, 11(6), 433–447.

Fearon, J. D. (1998). Bargaining, enforcement, and international cooperation. International Organization, 52(2), 269–305. http://www.jstor.org/stable/2601276

Friedman, D., & Sunder, S. (1994). Experimental methods: A primer for economists. Cambridge University Press.

Golbeck, Ryan. Finite Automata to Represent Bounded Rationality.

Herrera, H., Macé, A., & Núñez, M. (2025). Political brinkmanship and compromise. International Economic Review. https://doi.org/10.1111/iere.12760

Harsanyi, J. C. (1956). Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks', and Nash's theories. Econometrica, 24(2), 144–157. https://doi.org/10.2307/1905748

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. American Economic Review, 86(3), 653–660.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2001). Introduction to automata theory, languages, and computation. ACM SIGACT News, 32(1), 60–65.

Kimbrough, S. O., & Wu, D. J. (2001). Designing artificial agents for e-business: An OR/MS approach (Working paper). The Wharton School, University of Pennsylvania.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Reputation and imperfect information. Journal of Economic Theory, 27(2), 253–279.

Kühn, Clemens and Hillmann, Katja, Rule-Based Modeling of Labor Market Dynamics (August 27, 2010). Available at SSRN: https://ssrn.com/abstract=1666891 or http://dx.doi.org/10.2139/ssrn.1666891

Larsen, T., Leslie, D., Collins, E., & Bogacz, R. (2010). Posterior weighted reinforcement learning with state uncertainty. Neural Computation, 22(5), 1149–1179.

Lehr, A., Vyrastekova, J., Akkerman, A., & Torenvlied, R. (2018). Horizontal and vertical spillovers in wage bargaining: A theoretical framework and experimental evidence.

Rationality and Society, 30(1), 3-53. https://doi.org/10.1177/1043463117754079 (Original work published 2018)

Muthoo, A. (1999). Bargaining theory with applications. Cambridge University Press.

Nash, J. (1951). Non-cooperative games. Annals of Mathematics, 54(2), 286–295. https://doi.org/10.2307/1969529

Ning, Z., & Xie, L. (2024). A survey on multi-agent reinforcement learning and its application. Journal of Automation and Intelligence, 3(2), 73–91. https://doi.org/10.1016/j.jai.2024.02.009

Peleckis, K. (2015). The use of game theory for making rational decisions in business negotiations: A conceptual model. Entrepreneurial Business and Economics Review, 3(4), 105–121.

Rapoport, A., Weg, E., & Felsenthal, D. S. (1990). Effects of fixed costs in two-person sequential bargaining. Theory and Decision, 28(1), 47–71. https://doi.org/10.1007/BF00139238

Robot Learning Editorial Board (1999). Robot learning (Connell & Mahadevan, Eds.). Robotica, 17(2), 229–235. https://doi.org/10.1017/S0263574799271172

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. Econometrica, 50(1), 97–109. https://doi.org/10.2307/1912531

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems, 37(1–2), 147–166.

Schauer, M., Majer, J. M., & Trötschel, R. (2023). Nine degrees of uncertainty in negotiations. Negotiation Journal, 39(2), 207–228. https://doi.org/10.1111/nejo.12426

Tamura, K., & Morita, S. (2024). Analysing public goods games using reinforcement learning: Effect of increasing group size on cooperation. Royal Society Open Science, 11(12), 241195. https://doi.org/10.1098/rsos.241195

Turocy, T. L., & von Stengel, B. (2001). Game theory. In Encyclopedia of Information Systems (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-39940-9

Valley, K. L., Peterson, S., & Wilcox, N. T. (1992). Asymmetric information in bargaining: Experimental evidence. Journal of Economic Behavior & Organization, 17(3), 415–434.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8, 279–292.

Wolfram, S. (1994). Cellular automata and complexity. Addison-Wesley.

Zwick, R., Rapoport, A., & Howard, J. C. (1992). Two-person sequential bargaining behavior with exogenous breakdown. Theory and Decision, 32(3), 241–268. https://doi.org/10.1007/BF00134151

Appendix

The following section entails the information regarding code for simulations along generating the results.

Repository of Source File

This research utilised open source software application, Anaconda. It includes various packages including Jupyter-Notebook for runtime in Python environment. The code for replicating the results of this research can be accessed to the respective GitHub repository: https://github.com/JonSnow016/Thesis.git

Use of Libraries

The libraries in Python that were utilised to construct the simulation include:

```
# Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from tabulate import tabulate
import random
import csv
import os
```

Figure-13: List of Libraries Utilised

Constrained Increases Model

The following consists of the pseudo-code for the logical flow of the reinforcement learning:

Initialize Q(s, a) = 0 for both agents
Repeat (for each episode)
Proposer: chooses an initial offer a using ε-greedy policy
Responder: observes state (s=a) and chooses action a using ε-greedy policy
If Responder accepts:
  Episode ends. Proposer and Responder gets reward r and r'
If Responder rejects:
  With probability p, breakdown is triggered and both receive reward 0
    Update value functions with reward 0
  With probability 1 − p, Proposer increments offer to s+1
    Responder observes s' and chooses action a'
    Repeat the above steps until:
      Offer is accepted with terminal payoffs as above
      Breakdown occurs
      Maximum offer N−1 is reached where rejection leads to breakdown
At end of episode, update Q(s, a) values for all state-action pairs visited during episode for both agents in their respective Q-table

$Q(s, a) \leftarrow Q(s, a) \times (n-1)/n + r/n$
(where n is the total number that the state action pair has been visited)
Until k episodes have been played

Unconstrained Increase Model

The following presents the pseudo-code for reinforcement learning agents in the Reverse Ultimatum Game, where proposers are allowed to make a new offer after each rejection by the responder without being constrained by a fixed increment between offers

Initialize Q(s, a) = 0 for both agents
Repeat (for each episode)
Proposer: chooses an initial offer a using ε-greedy policy
Responder: observes state (s=a) and chooses action a using ε-greedy policy
If Responder accepts:
  Episode ends. Proposer and Responder gets reward r and r'
If Responder rejects:
  With probability p, breakdown is triggered and both receive reward 0
    Update value functions with reward 0
  With probability 1 − p, Proposer observes s and increments offer to a ∈ [s+1, N-1]
    Responder observes its state (s'=a) and chooses action a'
    Repeat the above steps until:
      Offer is accepted with terminal payoffs as above
      Breakdown occurs
      Maximum offer N−1 is reached where rejection leads to breakdown
At end of episode, update Q(s, a) values for all state-action pairs visited during episode for both agents in their respective Q-table
$Q(s, a) \leftarrow Q(s, a) \times (n-1)/n + r/n$
(where n is the total number that the state action pair has been visited)
Until k episodes have been played

Evolutionary Model of Bargaining Reputation

The following outlines the simulation flow used to observe the evolutionary dynamics of a subpopulation of responders with varying acceptance thresholds in a bargaining environment subject to an exogenous risk of breakdown.

Initialize responder with varying threshold based on population ratio
Assign each responder a history list and initial history window length
Assign each proposer either as normal (uses history) or mutant (fixed offer)
Repeat for each episode
If proposer is normal:
    Extract responder's recent history (last k offers and outcomes)
    Apply rule-based policy:
      If only acceptances: propose just below lowest accepted offer
      If only rejections: propose just above highest rejected offer
      If both: propose in-between or above max rejection
  If proposer is mutant:
    Start with fixed offer; if rejected, increase by 1 in next round

Engage in the actual bargaining
    Responder accepts if offer ≥ threshold
        Else, with probability p: negotiation ends (breakdown)
        Else: proposer increase offer by 1
    Update responder's history and reward from the interaction
    Keep an interaction count for each responder
When interaction count for each responder equals to x interactions: (Single generation)
    Compute average reward for each responder
    Normalize rewards into selection probabilities
    Select a fraction of low-performing responders for replacement
    Replicate high performers probabilistically
    Apply mutation with small probability to each population side
Until k generations are played