EVOLUTIONARY DYNAMICS OF ARTIFICIAL AGENTS: EXPLORATION AND LEARNING IN GAMES

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

in

Mathematics

by Brian Mintz

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

May 8, 2025

Examining Committee:

(chair) Feng Fu

Scott Pauls

Natalia Komarova

Dimitrios Giannakis

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

The natural world abounds with examples of complex behavior in humans and many other species. Evolutionary game theory is a powerful mathematical framework to understand the origins of many such behaviors like cooperation. Since these behaviors are often selected against initially, understanding why they are so widespread has been a longstanding question. Rather than assuming agents' rationality, like in traditional game theory, this approach studies the mutation and selection of strategies themselves. However most behavior is neither perfectly rational nor entirely determined by genetics. This dissertation works to bridge the gap between these two perspectives by analyzing models where individuals follow a variety of approaches to learning their strategy. One series of models characterizes the stability of equilibria in the notoriously intractable issue of exploration vs exploitation, consider constant, frequency, and time dependent selection. We extend this to groups of agents who simultaneously learn from their surroundings, advancing theory for a Machine Learning method combining Reinforcement Learning with Genetic Algorithms. We then examine agents that consider general types of norms, finding examples the can and cannot promote cooperation. Lastly we find the counter-intuitive effect that introducing trivial topics can completely change whether a population will polarize or reach consensus, based on an opinion dynamics model where agents partially adopt

the behavior of those around them. Taken together, these results improve our understanding of why particular behaviors have spread so successfully. This goal of understanding and optimizing behavior has applications to many fields, from biology, computer science, and economics, to psychology, ecology, and sociology. Several open questions remain, making evolutionary game theory a promising area for future research with broad scientific impacts. As we see rapid shifts in society brought on by advances in artificial intelligence and changes to the political and literal climate, it becomes increasingly important to understand how to foster cooperation within and between communities.

Preface

This work would not have been possible without the continued support of many individuals. I would like to begin by thanking my advisor Feng Fu for their rare combination of great kindness, thoughtful insights into the world of academia, and of course terrific research advice. I also consider myself extremely lucky to have the parents I do, their sympathy and efforts to support me have made all the difference in my life. Thank you.

I would also like to express my appreciation for the many friends I have made at Dartmouth over these last five years, including my cohort and those from earlier and later years. This period of my life has been a tremendous time of growth, and it has been truly wonderful having you all alongside this journey. Lastly, I'd be remiss not to mention Jeff and Katie at the jewelry studio and Greg and Pete at the wood shop. These were the most pleasant surprise and led to many enjoyable hours of crafting. Thank you for expanding my perspective on ways to be creative.

Contents

Chapter 1

Introduction

Section 1.1

Overview and Structure

This dissertation uses approaches from Evolutionary Game Theory to understand a variety of behaviors, most notably cooperation and learning. By leveraging stochastic and deterministic models of social and evolutionary systems, it provides a rigorous theory for the origin of these behaviors. These investigations have applications to numerous fields. Psychologists concern themselves with understanding the workings of the human mind. Sociologists and economists study how this individual complexity coalesces to form dynamics within and between groups, with the latter often assuming a form of rationality of individuals. In contrast, biologists work to understand the underlying mechanisms of living organisms, and ecologists research the complex systems created by the myriad interactions we see in nature. In contrast to these descriptive approaches, Computer Scientists care about a prescriptive recommendation to optimally solve problems. Connecting all of these is the shared goal of understanding

	Year	Journal
Chapter 2	2022	MDPI mathematics
Chapter 3	2023	MDPI mathematics
Chapter 4	2025	Chaos
Chapter 5	2024	Proceedings of the Royal Society A

Table 1.1: Publication listing for the chapters of this dissertation.

how interactions at small scales combine to form emergent properties in systems.

The next three chapters address the issue of exploration versus exploitation with models of increasing complexity. Chapter 2 considers several different models of mutation, and finds examples in each where mutation rates can counterintuitively evolve upwards despite a constant fitness landscape. Chapter 3 extends this analysis to frequency and time dependent selection, finding a transition between an attracting and repelling equilibria as the period of environmental change decreases. This relatively simple form of exploration is then extended using a common model of learning in chapter 4, where different cases of a social dilemma lead to positive and negative selection on an exploration parameter. The next chapter builds on this investigation by considering an alternative mechanism by which agents could make decisions in social dilemmas. It compares four norms, finding cases which can promote the emergence of cooperation and two that can not. Chapter 5 focuses on the social learning aspect of decision making, analyzing multidimensional opinion dynamics with heterogeneous weights. Varying types of symmetry in the equilibria of this model explain orders of magnitude longer absorption times and whether the population with polarize or reach consensus. Finally, chapter 7 provides some general future directions for this research, beyond the specific suggestions in each chapter. As these chapters have been adapted from publications, summarized in Table 1.2, they are mostly self -ontained so can be read independently.

Section 1.2

A Brief History of Evolutionary Game Theory

Theodosius Dobzhansky famously wrote "Nothing in Biology Makes Sense Except in the Light of Evolution." Indeed, the theory of mutation and natural selection has been shown to have great explanatory power in understanding behavior. The core idea of an Evolutionarily Stable Strategy, a refinement of the Nash Equilibrium, was introduced in 1973 article "The Logic of Animal Conflict" by Smith and Price[214]. Subsequent research applied this game theoretic approach to other topics in biology including sex allocation, contest behavior and signaling, and most notably, cooperation. Recently the 50th anniversary of this seminal work occurred, leading to a number of retrospectives of the field[99, 135, 196, 234].

In contrast to traditional game theory, which assumes a rationality, or extreme intelligence, for agents to maximize payoffs and reason about the intentions of other player, this approach assumes far simpler decision making, either through genetically inherited strategies or imitation of more successful strategies. This is extended to learning agents in chapter 4, allowing them to display more complicated behavior. Beyond the intuitive appeal of this evolutionary framework, one key strength is that it provides clear models for how strategies change over time, including replicator and adaptive dynamics (described in the next subsection). This gives an understanding of why certain strategies occur beyond traditional equilibrium concepts, which only explain why they persist.

One key limitation of many advancements in evolutionary game theory is an apparent arbitrariness in the choice of strategies considered or model of evolution[234].

This can bias results, causing a lack of robustness. For example, the Moran birth-death and death-birth processes are related models for evolution that, despite their similarity, can give substantially different predictions due to the global versus local nature of competition[118, 265]. This is especially important as the success of a strategy is highly frequency dependent, so considering different collections of strategies could lead to starkly differing conclusions. The next chapter addresses this aspect by considering a range of models for mutation, showing a consistent counterintuitive effect among all of them. Chapter 5 is similarly agnostic about norms, allowing for a comparison of four well-known examples as part of a far broader class.

Section 1.3

Mathematical Approaches

The work in this document comprises a range of stochastic and deterministic approaches to modeling the evolutionary systems it studies. Often, limits are taken in a stochastic model to obtain a deterministic model.

One of the main concepts introduced by Evolutionary Game Theory is that of evolutionary stability. An evolutionarily stable strategy (ESS) is one that has a selective advantage when all other strategies are rare. Intuitively, this means that if it takes over a population, it cannot be displaced. Formally, if $P(x, y)$ gives the payoff received by strategy x when interacting with strategy y, then x is a (strict) ESS if for all strategies y either

(a) $P(x, x) > P(y, x)$, or

(b) $P(x, x) = P(y, x)$ but $P(x, y) > P(y, y)$

, where the second condition accounts for strategies that perform equally well against x, but does worse among themselves. A weak ESS makes this last inequality weak instead of strict. This notion can be viewed as a refinement of the concept of a Nash Equilibrium (NE), which describes a situation where no player can improve their payoff by unilaterally changing their strategy, the first condition of an ESS. Because of this added condition, any strict (symmetric) NE is a strict ESS, and any weak ESS is a weak (symmetric) NE. Since a weak ESS still requires the first condition, it is more general than a weak NE. As the proportion ε of players following strategy y goes to zero, $P(x, y)$ dominates the differences in average payoff

$$(1 - \varepsilon)(P(x, x) - P(y, x)) + \varepsilon(P(x, y) - P(y, y))$$

between strategies x and y. Accordingly, in the case of only two strategies, x is an ESS precisely if the strategy distribution concentrated entirely in x is asymptotically stable. Thus the replicator equations can be used to detect evolutionarily stable strategies. The stability of this equilibrium implies any infrequent novel types will become extinct. Evolutionarily stable strategies can be seen as a global optimum. In contrast, adaptive dynamics, described below, finds local optima, as it only considers local mutations. Nonetheless, this is a necessary condition for global optimality.

Many models of evolution can be described by stochastic process. Often the transition probability depends only on the current state, making these Markovian. Such processes can be described by their transition matrix A, assuming they have a finite number of states. This can be used to update the probability distribution $\vec{x}_t$ at

time t through matrix multiplication,

$$x^t = A^t x_0 \tag{1.1}$$

In particular, all entries of A will be non-negative real numbers, and the all row-sums of A will be one, corresponding to the ith row being a probability distribution over the possible states i can transition to. The Perron-Frobenius theorem states that A has a maximum eigenvalue of one with a unique associate eigenvector, known as the stable distribution, which gives the long time limit of the probability of the process being in any given state. Intuitively, we can view this process as a weighted random walk on the state-space of the system. If the matrix has absorbing states, those which only transition to themselves, we can calculate the "fundamental matrix" as

$$(I - Q)^{-1} = I + Q + Q^2 + Q^3 + ...$$

where Q is the sub-matrix of transitions between non-absorbing states. The i, j-th entry of this gives the average number of times the process reaches state j before absorption starting from state i, so in particular the sum along this row gives the expected absorption time. This equations comes from the fact that absorption occurs after some finite number of steps in non-absorbing states, given by multiplication by Q. Starting from a state i corresponds to having the probability distribution entirely concentrated on that state with $x_i = 1$, so the probability of being absorbed into state j after t steps is given by the i, j-th entry of A^t.

For example, the Moran birth-death process considers a population of n individuals of two types, one individual is selected for birth proportional to their fitness, then

another is selected for death uniformly at random, ensuring a constant population size. Following [177], if we let x_i be the probability that a type with initially i individuals takes over a finite population of size N, then conditioning on a single step of this process gives the recurrence equation

$$x_i = \beta_i x_{i-1} + (1 - \alpha_i - \beta_i)x_i + \alpha_i x_{i+1} \tag{1.2}$$

where α_i and β_i are the probability of transitioning to having a state with one less or one more of this type. Rearranging this equation using the auxiliary variables $y_i = x_i - x_{i-1}$ and $\gamma_i = \beta_i/\alpha_i$ one obtains

$$y_{i+1} = \gamma_i y_i \tag{1.3}$$

Since $y_1 = x_1 - x_0 = x_1 - 0 = x_1$ one has $y_i = x_1 \prod_{j=1}^{i} \gamma_j$. Substituting this into the telescopic series $\sum_{j=1}^{N-1} y_j = x_N - x_0 = 1 - 0 = 1$ implies $x_1 \left(1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k \sum\right) = 1$ and therefore

$$x_i = \frac{1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \gamma_k}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \gamma_k} \tag{1.4}$$

since $x_i = x_1 + \sum_{j=1}^{i-1} y_j = x_1 \left(1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \gamma_k\right)$ using the same telescoping series. In particular if the mutant type is r has fitness r while the rest of the population has fitness one, then probability of fixation, being absorbed into the all mutant state starting from the state with one mutant, is

$$\rho = \frac{1 - \frac{1}{r}}{1 - \frac{1}{r^N}} \tag{1.5}$$

If a low-mutation limit is assumed, such that the average fixation time, where a

new mutant either becomes extinct or takes over the population, is less than the average arrival time of a new mutant, then one may accurately treat the population as monomorphic. In this case, one can compute the fixation probability of a mutant given the current population trait to obtain a Markov process on monomorphic population states, where the transition rates are given by fixation probabilities and the stable distribution gives the proportion of time each trait is present in the population.

One significant limitation of this approach is that the possible states in these Markov chains grow rapidly with population size or the number of types. Specifically, this is given by a combinatorial object, the number of compositions of the population size n into the number of possible types t, given as $\binom{n-1}{t-1}$ by Catalan. If there are just two types this is $n+1$, states are given by the number of type one, which can be any integer between zero and n. However already for $n = 3$ the number of compositions of a population with ten individuals is thirty six, which explodes to 210 for four types. These large matrices require software assisted computation of their eigenvalues and eigenvectors. Indeed, these cannot be found analytically for even small matrices, the insolubility of the quintic implies there is no general formula for the eigenvalues of even a five by five matrix. Due to the computational nature of this problem, it is often important to consider the numerical stability of the solver used.

By taking a large population limit, one can convert these stochastic dynamics to a system of ordinary differential equations known as the Replicator equations. Assuming there are finitely many types, we can use a vector $\vec{x}$ to give the population distribution, where x_i is the frequency of type i. These lie on the n dimensional simplex, given by $\sum_i x_i = 1$. If we define "fitness" as the growth rates of each type,

depending on the population compositions, then we have

$$\frac{d}{dt}x_i = x_i(f_i(\vec{x}) - \bar{f}) \tag{1.6}$$

where $f_i(\vec{x})$ is the fitness of type i given the population distribution $\vec{x}$ and $\bar{f} = \sum_i x_i f_i(\vec{x})$ is the average fitness in the population, introduced so the population size stays constant (indeed it makes $\sum_i \frac{d}{dt}x_i = 0$). These and govern the dynamics between types in a population. In two dimensions, a single variable x can be used for the proportion of type one, and dynamics essentially depend on the fitness of type one $f(x)$ as a function of x. When $f(x) = 0$, an equilibrium occurs, and the sign of $f'(x)$ determines it's stability. With three types, population distributions lie on the triangle $x_1 + x_2 + x_3 = 1$. Far more behaviors can occur, including limit cycles. The fact that the fitness is frequency dependent introduces significant non-linearity into this dynamical system, allowing for a variety of behaviors and restricting the types of analysis that may be performed. These equations can be put into matrix form as

$$\frac{d}{dt}\vec{x} = (F(\vec{x}) - \bar{f}I)\vec{x} \tag{1.7}$$

where $F(x)$ is a diagonal matrix with entries $F_{ii}(x) = f_i(\vec{x}) - \bar{f}$. Mutation can also be incorporated into this model by adding a mutation matrix M giving the transition rates between types yielding

$$\frac{d}{dt}\vec{x} = (QF(\vec{x}) - \bar{f}I)\vec{x} \tag{1.8}$$

If there is a natural ordering to these issues, there is a natural extension to an unac-

Table 1	A	B
A	20	0
B	17	1

Table 2	A	B
A	1	28
B	2	30

$\Rightarrow$

	Table 1	Table 2
A is an ESS	$N > 12$	$N < 22$
B is an ESS	$N < 53$	$N > 17$

Table 1.2: In finite populations, whether a strategy is an ESS or not can depend on the population size.

countably infinite number of types indexed by the unit interval. Here the system of ordinary differential equations become a single partial differential equation governing the distribution of types $p(x, t)$ as a function of time.

One weakness of this approach is that it completely removes stochastic effects that can be critical to evolution. In particular, for small populations randomness is a powerful force. Neglecting it can lead to quite different conclusions[181]. For example, if fitness is determined by the following tables, then it is possible for the population size to change whether a type is an ESS, modified appropriately for this finite context [177]. However it often only takes a moderate population size to minimize the role of randomness, for example $N = 50$. Since the vast majority of populations are above this size, this approach usually gives a good model for the population dynamics between types.

The last main tool used in this work is Adaptive Dynamics. This approach makes two assumptions to model evolution. The first is a sufficiently low mutation rate so that only two types need to be considered at once, and the population will be mostly monomorphic (in practice this needs only be an approximation). This simplifies the analysis, allowing one to define an invasion fitness $f_x(y)$, describing the likelihood a mutant trait x can invade a resident trait y. The second assumption is that mutations are local, allowing for the gradient to determine the approximate evolutionary

trajectory

$$f_x(y) \approx f_x(x) + (y - x)\frac{\partial}{\partial y} f_x(y)|_{y=x} \tag{1.9}$$

Since $f_x(x) = 0$, the sign of $\frac{\partial}{\partial y} f_x(y)|_{y=x}$ determines whether larger or small values of the trait can invade. Thus, it is used as a proxy for the rate of change $D(x)$ of the trait overall. Equilibria will then occur when $D(x) = 0$ with the sign of $D'(x)$ indicating their stability. A special case occurs when $D'(x) = 0$ but $D''(x) < 0$, indicating the population is at a local fitness minima, allow for mutants on both sides to invade, causing branching, also known as sympatric speciation. In multiple variables, the gradient gives the direction of greatest increase, so the most likely evolutionary outcome. Intuitively, this process is gradient ascent where the hill changes as a function of one's position on it.

Many extensions have been made to these basic models. For example, instead of a well-mixed population, several studies examine evolution on a network to capture spatial effects[224, 143]. Even a simple model with just cooperation and defection can produce chaotic behavior when spatial effects are considered, leading to beautiful fractals starting from symmetric initial conditions in a regular lattice [179]. Studies have also examined the effect of spatial structure on the emergence and spread of mutations, notably cooperative behavior [96, 203]. In general certain network properties are known to promote evolution, for example the star and funnel shapes where a small number of important nodes have the potential to replace many [177]. Other extensions include coevolutionary models, including ecological effects, or feedback loops between population dynamics and structure.

Across all these models, similar questions can be asked about the dependence of the outcomes on model parameters. For example, how does the equilibrium strategy

vary with mutation levels, or when do game parameters allow for the evolution of cooperation? These can be studied using bifurcation analysis to identify how properties of the equilibria depend on model parameters. In particular, often the linearization of a system is used around equilibria to evaluate their stability.

In the following work, the replicator-mutator equation is investigated using adaptive dynamics in chapters 2 and 3. Chapter 4 uses adaptive dynamics to analyze a combined learning and evolutionary process. Chapter 5 investigates a Markov process in a finite population, studying its stable distributions and various limits, as well as a large population limit of discrete time difference equations. Chapter 6 studies a generalized process where the transition probabilities $A(\vec{x})$ depend on the state of the system $\vec{x}$ and a large population system of nonlinear differential equations.

Section 1.4

Social Dilemmas and the Evolution of Cooperation

Cooperation broadly refers to any behavior that benefits another individual. Despite such actions almost universally coming at a cost to oneself, we see widespread examples of this in nature, from micro-organisms such as viruses and bacteria to macro-organisms like fish, meerkats, and vampire bats [66]. Indeed this effect is theorized to be the origin of multicellular life itself, from the endosymbiosis of mitochondria in a cell to colonies of cells cohering into an organism. Beyond its inherent scientific merit on the origins of complex life, understanding the reasons for why cooperation is so widespread has critical implications for a wide variety of important applications, from promoting peaceful international relations to uniting to solve clime change.

A variety of fields including philosophy, psychology, and sociology have tried to explain why cooperation occurs. Complementing these theoretical and experimental approaches, there have recently been several researching seeking to mathematically formalize such theories and apply them to various models. Most notable is the Prisoner's Dilemma where two individuals each simultaneously choose between cooperation and defection, and receive payoffs based on the pair of actions chosen, summarized by the table

	C	D
C	R	S
D	T	P

where $S < P < R < T$ are real numbers giving the payoff to the row-player. Lesser known is a weaker dilemma, the Hawk-Dove or Snowdrift games, where $P < S < R < T$. This is an anti-coordination game, where it is better to choose the opposite strategy of the other player. For interactions with more than just two individuals, the Public Goods Game is commonly used, encapsulating dilemmas like the tragedy of the commons. Individuals in a group of size N each have the choice of how much to contribute to central resource, which is then scaled by some function and uniformly divided. Letting $R(x)$ be the function determining the value of the common resource for a total contribution of x, and c_i the contribution of individual i, the payoff of individual i is

$$p_i(\vec{c}) = \frac{1}{N} R\left(\sum_{i=1}^{n} c_i\right) - c_i$$

In particular, most studies consider $R(x)$ linear and c_i only being zero or one for simplicity, though generalizations with nonlinear scaling functions and continuous contribution levels have also been investigated. In large groups there is an incentive

not to contribute, as this will make little difference for a large cost.

In this dissertation, we see social dilemmas in chapters 3, 4, and 5, where we study the exploration versus exploitation, the effects of learning alongside evolution for the alignment of artificial intelligence, and influence of various norms.

Section 1.5

Exploration vs. Exploitation

Everyone is routinely faced with recurring decisions. Where should one go for dinner? What song should one listen to next? Which novel drugs should we invest more research and development funds into? By facing the same questions repeatedly, we can estimate the best one. However, if we only choose the best known option, there is no possibility to discover new and better options. These are the fundamentally incompatible priorities of exploitation and exploration. Choosing either in isolation is clearly not optimal, and finding the best solution has been a longstanding question in computer science[48].

A canonical framework for thinking through this problem is the Multi-Armed bandit, where one repeatedly chooses from multiple slot machines, or "one armed bandits", with unknown payouts. Notorious for its intractability, mathematician Peter Whittle said it was suggested this problem be introduced to the Axis powers as "the ultimate instrument of intellectual sabotage,"[65] an exact solution was eventually found by Richard Bellman while working for the Rand Corporation[21]. However actually computing this was infeasible, so the problem remained effectively unsolved. The next breakthrough occurred with John Gittins in the 1970's, who introduced an index measuring the amount of money one would prefer to forgo an option that

had k successes out of n trials, given a geometric discounting of payoffs[87]. While mathematically optimal, there are serious doubts about whether such discounting is realistic, so later work focused on the notion of regret, a comparison of the payoff received with the best possible sequence of choices. Lai and Robbins discovered several algorithms that achieved the best possible performance, known as "Upper confidence bound algorithms," where a confidence bound is computed for each option, then the option with the largest upper bound is chosen[129]. Beyond guiding our daily lives, this theoretical work has significant impacts on important decisions made in research. In the early days of the internet, Google and Amazon carried out extensive A/B testing, routing users to different versions of their site, to determine which performed the best. Since then methods for testing have become more sophisticated and applied to diverse domains such as political campaigns and cutting edge medical treatments.

These approaches are focused on a prescriptive solution to the problem, rather than a descriptive analysis of how individuals actually solve it. The next three chapters address how natural selection leads us to solve this dilemma. We first focus on the simplest environment: constant selection. While intuitively exploration can only be detrimental here, as once the optimum is found, further exploration is always negative, we find cases where it can be beneficial in several models. Specifically, we think of exploration as mutation in a trait space, commenting on the question of the evolution of evolvability itself. This is followed by an extension where fitness varies with time or space, unlike in the multi-armed bandit problem where the rewards for a trait are fixed. Then chapter 4 considers a more sophisticated model of exploration based on reinforcement learning, a powerful artificial intelligence technique based on psychological theories of learning. Here the agents learn from each other, creating an

dynamic environment.

Chapter 2

The Point of No Return: Evolution of Excess Mutation Rate is Possible Even for Simple Mutation Models

Under constant selection, each trait has a fixed fitness, and small mutation rates allow populations to efficiently exploit the optimal trait. Therefore, it is reasonable to expect that mutation rates will evolve downwards. However, we find that this need not be the case, examining several models of mutation. While upwards evolution of the mutation rate has been found with frequency- or time-dependent fitness, we demonstrate its possibility in a much simpler context. This work uses adaptive dynamics to study the evolution of the mutation rate, and the replicator–mutator equation to model trait evolution. Our approach differs from previous studies by considering a wide variety of methods to represent mutation. We use a finite string

approach inspired by genetics as well as a model of local mutation on a discretization of the unit intervals, handling mutation beyond the endpoints in three ways. The main contribution of this work is a demonstration that the evolution of the mutation rate can be significantly more complicated than what is usually expected in relatively simple models.

Section 2.1

Introduction

Evolution occurs in a population through a repeated process of mutation, which introduces new traits, and selection, where traits leading to a higher reproduction rate outcompete less fit traits. Since mutation is fundamental to the process of evolution, determinants of changes in mutation rates are an important question to address. While mutation allows a species to adapt, it also likely causes harm through occasional deleterious mutations. These opposing forces make it nontrivial to determine the optimal levels of mutation, let alone how they will evolve. Prior studies have investigated this question in diverse areas, from phytoplankton to primates, [125, 233, 148, 46]. Theoretical treatments of mutation rate evolution allow for general insights to be obtained that apply to broad biological contexts. Insightful results have been found, including how the mutation rate can evolve up or down depending on the mechanism of selection [6, 141], among others. In this paper, we build on these prior works and consider a wide variety of mutation models that show counterintuitive effects on favoring excess mutation rate.

Often, mutation is thought of as the result of defects in the reproduction process. This can take place in a genetic context, for example, errors in DNA or RNA

transcription, or a cultural one, such as imperfect language acquisition or offspring behaving differently than their parents. Crucially, the mutation rate is itself subject to evolution because of the existence of error-correcting genes, in the genetic context, and the degree to which a society will change its belief and practices is itself a part of culture. As well as acting between generations, mutation can be thought of within a generation as, albeit random, exploration, for example, by varying hunting strategies or organization structures. The other mechanism, selection, can be constant or change in some deterministic manner. For example, the fitness of a trait may depend on the proportion of other traits in a population, that is, selection is frequency-dependent [6, 141, 184, 49]. Alternatively, selection may be time-dependent, perhaps a trait was initially viable but now is not [250, 174, 140, 172, 78]. This wide variety of contexts can all be encoded in the mathematical framework of evolutionary dynamics. By analyzing this, we can discover interesting effects and provide insights into the paths taken by evolution.

One of the most basic models of evolution considers a finite number of traits, each with a constant fitness f_i, effectively their rate of reproduction. A dependence on time or population composition can be easily incorporated by making the fitness of trait i a function of these, but doing so greatly complicates the subsequent analysis. It is also helpful to assume an infinitely large, well-mixed population, as this allows the dynamics to be described by a system of ODEs, the *replicator equation* $\frac{d}{dt}x_i = x_i(f_i - \phi)$ for each trait i with proportion x_i in the population, and $\phi \equiv \sum_i x_i f_i$ the average fitness (introduced by the normalization $\sum_i x_i = 1$). Under this model, a population adapts to a fitness landscape, with the most fit trait becoming dominant [176, 72, 262, 104, 231]. This equation encodes selection, and to account for mutation, we can

introduce a matrix $Q(u)$, whose ijth entry is the transition probability from trait j to trait i given a parameter u for mutation rate (while $Q(u)$ technically gives the degree, not rate, of mutation, these are interchangeable given a unit time step). Now the dynamics are given by the *replicator–mutation equation*

$$\frac{d}{dt}\vec{x} = (Q(u)F - \phi I)\vec{x} \tag{2.1}$$

where $\vec{x} = (x_1, x_2, \ldots, x_n)$, F has ii^{th} entry the fitness of trait i, and I is the $n \times n$ identity matrix [40, 219, 122, 106, 51, 5, 186, 4, 3, 112]. While this forms a complete description of the evolution of traits under a constant mutation rate, it does not specify how the mutation rate itself evolves. To do this, it helps to assume a separation of timescales between the evolution of a trait and the mutation rate [7]. This makes it sufficient to analyze how an invading mutation rate will perform in a population at equilibrium, formalized by *adaptive dynamics*.

Adaptive dynamics is a technique to predict evolutionary outcomes under small mutations [67, 83, 240, 84, 159, 161, 62, 100]. In this framework, mutants appear infrequently enough that competition between invaders and residents, who have reached equilibrium, occurs before subsequent mutants arise. The likelihood of the fixation of invading mutants with one-dimensional trait y into a homogeneous population with trait x is given by the *invasion fitness* $f_x(y)$, whose exact form depends on the particular application. Since mutations are small enough, the linear approximation $f_x(y) \approx f_x(x) + (y - x)\frac{\partial}{\partial y} f_x(y)|_{y=x}$ is accurate. Since $f_x(x) = 0$, this means invasion can only occur if $y - x$ is the same sign as $\frac{\partial}{\partial y} f_x(y)|_{y=x}$. Thus, up to a constant, the

rate of change of the trait is given by

$$\frac{d}{dt}x = D(x) := \frac{\partial}{\partial y} f_x(y)\Big|_{y=x} \tag{2.2}$$

This will have equilibria when $D(x) = 0$, with stability determined by the sign of $\frac{d}{dx}D(x)$.

Combining these two techniques, [7] determines that the appropriate invasion fitness of a small invading population with mutation rate u' and distribution x' into a resident population with mutation rate u and distribution $\tilde{x}$ is

$$f_{u,\tilde{x}}(u', x') = \lambda_{\max}(u', \tilde{x}) - \phi(\tilde{x}) \tag{2.3}$$

where $\lambda_{\max}(u', \tilde{x})$ is the maximum eigenvalue of $Q(u')F(\tilde{x})$, and $\phi(\tilde{x})$ is the average fitness of the distribution $\tilde{x}$, and $\tilde{x}$ is an equilibrium distribution of Equation (3.1). It is important to note that a positive invasion fitness does not necessarily imply that the invaders will replace the resident population, though this is a likely proposition.

With this mathematical framework, we are able to ask some profound questions about the evolution of mutation rate. What is the best mutation rate in a given setting? Can a population evolve toward a non-optimal, but stable, mutation rate? What are the effects of various design choices on the outcome of the system? Previous work has shown interesting results for time- and frequency-dependent fitness, but is this necessary for such behavior? Intuitively, one might expect low mutation rates to be preferred, as they cause a population to deviate less from an optimum. However, they can also lead to a population being stuck at a local, but not global, maximum. This work investigates these questions, most notably finding instances of nontrivial

mutation rate evolution in several plausible models of mutation. Surprisingly, we demonstrate that these effects can occur, even in the simple context of constant selection.

In what follows, we present details for the three models of mutation considered in this work in Section 2.2. In Section 2.3, we present and discuss specific results obtained for these models, respectively. We conclude our work with a brief discussion of potential extensions for future work in Section 2.4.

Section 2.2

Methods and Models

Whereas most models assume traits are unrelated, this work uses various notions of locality to model mutation, which should be more common to 'close' than 'far' traits. We represent mutation in three ways, depicted in Figure 2.1.

One option is motivated by genetics, considering traits a collection of genes, each with some finite number of alleles that mutate independently. For simplicity, one may take all genes to have the same number of alleles, and mutation to be symmetric [220, 208]. Then, this model becomes finite strings on a finite alphabet, with letters independently changing with some probability u to one of the other letters uniformly at random. This is depicted in the first part of Figure 2.1, in the case of binary strings of length two. Here, we see the relative mutation probabilities from a given trait. One benefit of this approach is the closed form entries of $Q(u)$, namely entry ij is $(u/(k-1))^d(1-u)^{n-d}$, where n is the length of the string, k is the size of the alphabet, and d is the number of positions that have different values in the strings i and j.

Another approach is to consider traits as representing some bounded one-dimensional

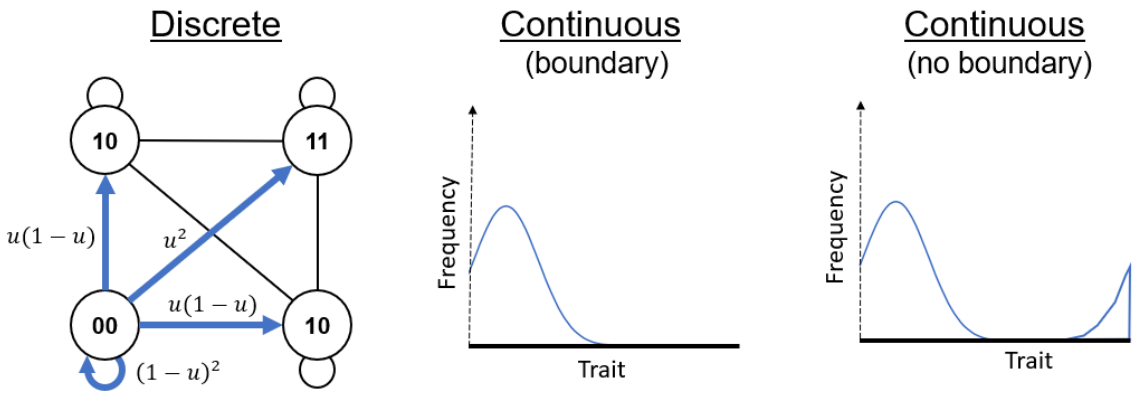


Figure 2.1: Diagrams of three models for mutation. First is finite strings on a finite alphabet with independent, uniform mutation on each letter. The arrows represent the possible mutations from 00 with their corresponding probabilities, with the remaining lines being possible mutations from other initial traits. The next two show local mutation on continuous traits. Mutations outside the boundary is truncated in the middle plot, and wraps around in the final plot. This removes the boundaries in the last model, making the trait space a circle. The local mutation is shown as a normal distribution, but it could be any curve, for example, an indicator function of width u.

quantity, such as an organism's weight or the length of an appendage. This could also be any behavior that can sensibly be assigned a number, such as a preference for cooperation. In this setting, it is natural to represent mutation as a local process since one should expect mutants not to differ too much from their parents. We formalize this idea by using a spread kernel that determines how offspring deviate from their parents. For example, this might be the normal distribution centered around the parents trait with standard deviation u, or a uniform probability to any trait within $u/2$ of the parent's trait. This is shown in the middle of Figure 2.1, where the curve represents the trait distribution of offspring of a given individual. Since the theoretical framework is based on a finite collection of traits, it is necessary for the trait space to be bounded. Consequently, mutation outside of these bounds must be

handled in some way. Two reasonable approaches are to either truncate out of bound values, or accumulate them and add the sum to the lowest possible trait. These often result in similar effects, so truncation is considered for simplicity, though the results for accumulation are also given. This is displayed in Figure 2.1, where the portion of the curve outside the interval is omitted. Alternatively, we also consider making the boundary periodic, resulting in a circular trait space. This is depicted in the last part of Figure 2.1; note how the portion of the normal curve beyond the boundary is translated through the other endpoint.

The model used will get encoded in the matrix $Q(u)$. The results of [7] can then be applied once one specifies the fitness function. In general, this is quite difficult, mainly due to determining the limiting distribution $\tilde{x}$. At equilibrium, the right-hand side of Equation (3.1) will be zero, yielding $Q(u)F(\tilde{x},t) = \phi\tilde{x}$, that is, $\tilde{x}$ is an eigenvector of $Q(u)F(\tilde{x},t)$. Since the entries of F themselves depend on the unknown $\tilde{x}$, and a time variable t, this is challenging to solve, so it is more effective to numerically solve the system of differential equations given by Equation (3.1), for example, by the forward Euler method. However, if fitness does not depend on the trait distribution or time, as in constant selection, then $F(\vec{x},t)$ is simply a constant matrix F, making $\tilde{x}$ the eigenvector corresponding to the maximum eigenvalue of $Q(u)F$. This may be solved explicitly, yielding a faster and more accurate solution for $\tilde{x}$. Further, this eigenvalue is the growth of the entire population, which is the same as the average rate of reproduction ϕ. Thus, the invading mutation rate u' will have positive invasion fitness for a resident u, when $\lambda_{\max}(Q(u')F) > \lambda_{\max}(Q(u)F)$. That is, the mutation rate will evolve in the direction of higher values of $\lambda_{\max}(Q(u)F)$, so the local maxima/minima of this curve will be stable/unstable equilibria. The matrix $Q(u)F$ and its eigenvalues

are calculated using Matlab, the code for which is available at the Github repository `https://github.com/bmDart/Evolution-of-Mutation-Rate`.

Section 2.3

Results

For each model of mutation considered, we identify conditions leading to regions where increases in mutation rate are favored, though the optimal mutation rates are those closest to zero.

In the simplest nontrivial case of finite strings on a finite alphabet, binary strings of length two, one often sees a non-monotonic curve. This was observed in [220] using binary strings of length three and a related fitness function. The authors explain that the large fitness for mutation rates near one result from an approximate cycle between states with mostly ones and states with mostly zeros. Therefore, if those states are fitter, higher mutation rates are selected for. A similar effect is seen for longer strings and larger alphabets, but is less pronounced. This is likely because the mutation is less likely to create such cycles. Interestingly, this effect disappears if mutation is instead allowed to result in any letter, including the starting letter (that is, when there is a $(u/k)^d(1-u+u/k)^{n-d}$ probability of transitioning between strings with d different characters, where n is the length of the string and k is the size of the alphabet). Figure 2.2 shows that for a certain fitness, mutation rates above 0.5 are favored: specifically, the fitness highest at the trait 00 and uniform elsewhere, panel (a). The maximum eigenvalue of $Q(u)F$ is plotted in (b), where one sees an increase near one, leading to upward evolution of mutation rate. One sees that the local minimum occurs around three quarters, so higher mutation rates will only evolve if

the initial value is significantly large, reflecting a highly inaccurate replication process. Since the goal of organisms is to replicate themselves, usually with an accuracy far better than half, such a scenario is unlikely, despite being mathematically possible. Lastly, we see the limiting distributions in panel (c), which confirm the explanation given above; for large mutation rates, the population is balanced between the fittest trait 00, and its opposite 11.

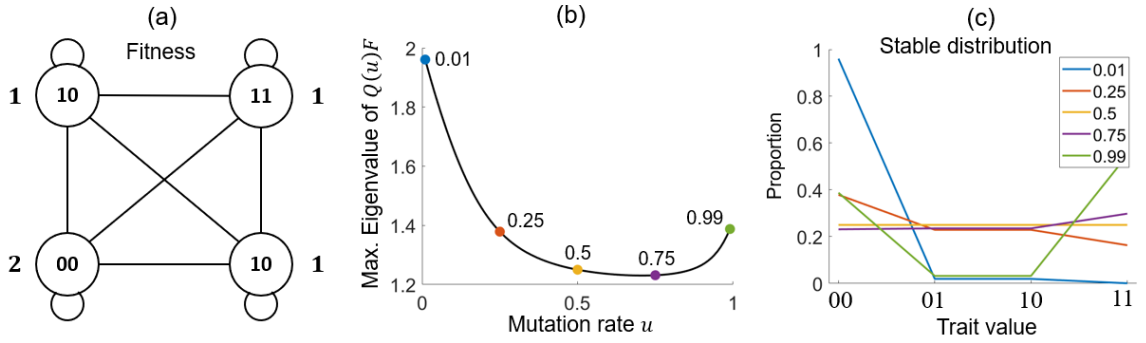


Figure 2.2: For the given fitness function (**a**), we see that mutation is eventually favored, and the maximum eigenvalues increase with mutation rate (**b**). Panel (**c**) shows that the stable distributions corresponding with the points of the same color in (**b**) flatten with increasing mutation, then become bimodal.

Next, when out of bound mutants are accumulated, the endpoints contain most of the population for large mutation rates. Therefore if the endpoints have high fitness, it makes sense that large mutation rates can be selected for. This is seen, for example, with the fitness function $f(x) = \cos(2\pi x)$, which is shown in Figure 2.3. More surprising is that this still holds when mutation is truncated at the boundary, though to a lesser extent. This is unexpected, as mutation near the boundaries is more toward the middle, so there will not be the same clustering at the endpoints. However, since traits in the center are mutated to more often than traits at the boundary, the distribution will be peaked in the center. This effect diminishes with a

larger mutation rate, so if traits at the boundary are more fit, the mutation rate can increase. An example of this is given in Figure 2.3. Like Figure 2.2, three panels show the fitness function, maximum eigenvalue as a function of u, and stable distributions. Here, we see a local minimum, and therefore an evolutionary unstable, mutation rate around 0.4. As explained, panel (c) shows that increasing the mutation rate from 0.5 to 1 flattens the distribution, causing more of the population to be at the optimum trait zero or one. A similar curve can be observed even in a very coarse discretization. Indeed, one only needs to use the points 0, 0.5, and 1, allowing one to explicitly solve for the maximum eigenvalue as a function of u for a simple spread kernel, for example, where offspring mutate uniformly within $u/2$ of their parents. To see if this effect could be replicated without the boundary effect, we considered creating a virtual boundary by sharply decreasing the fitness around a shrunk version of the fitness curve. However, initial attempts led to maximum eigenvalues that monotonically decreased with the mutation rate.

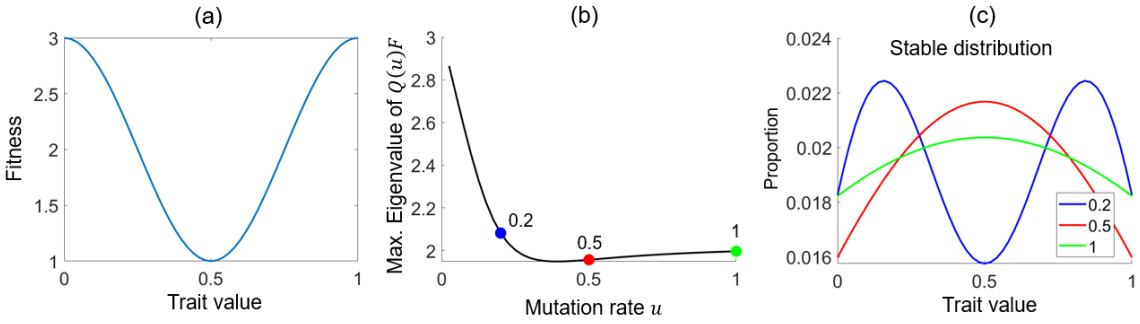


Figure 2.3: When fitness is maximal near the boundaries, for example $f(x) = \cos(2\pi x)$, shown in (**a**), accumulation near the endpoints can lead to increase fitness with mutation rate, seen by the increase in (**b**). In (**c**), we see that for intermediate mutation rates, the distribution has a peak in the center. This decreases with a larger mutation rate, increasing the overall fitness. These results are derived in the truncation model, though this effect is more pronounced when accumulation is used.

Even when there are no boundary effects, achieved by using a circular trait

space, we still found fitness landscapes leading to interesting effects in the evolution of mutation rates, shown in Figure 2.4 in the same manner as figures/no-return-fig 2.2 and 2.3. Specifically, this occurred for highly periodic fitness functions, such as $f(x) = \cos(10\pi x)$, and the spread kernel, where offspring mutate uniformly within an interval of length u around their parent's strategy. Panel (b) shows this setup leads to unstable equilibria, local minima of the maximum eigenvalue plot, but also stable equilibrium not seen in the earlier models. We suspect this occurs because of a wraparound effect, that is, for $u > 1$, the mutants wrap around the boundary, leading to a cluster at the diametrically opposite trait. As u increases further, the offspring once again become more prevalent near the parent's trait. Since one can think of the fitness on the circle as a periodic fitness on the infinite line, the same effect can be achieved for smaller mutation rates by rescaling the fitness function, that is, increasing its frequency. This is significant, as it demonstrates this effect can occur without unnaturally wrapping around the trait space. Further, by increasing the frequency even more, one should be able to create a local maximum in $\lambda_{\max} Q(u)F$ arbitrarily close to $u = 0$. This means that stable and unstable equilibria may be reached regardless of initial mutation rate, given an appropriate fitness function. Interestingly, this effect does not occur with the normal spread kernel, as the small tails prevent a similar wraparound phenomenon from occurring. It is also surprising that the stable distribution shown is centered around the boundaries, though some peak is to be expected given the low mutation rate. Lastly, we found that the fitness does not need to be strictly periodic, but can become constant near the boundary, and this effect will still occur, albeit to a lesser extent. Not only is this more realistic, but it also leads to a bounded set of traits in the population.

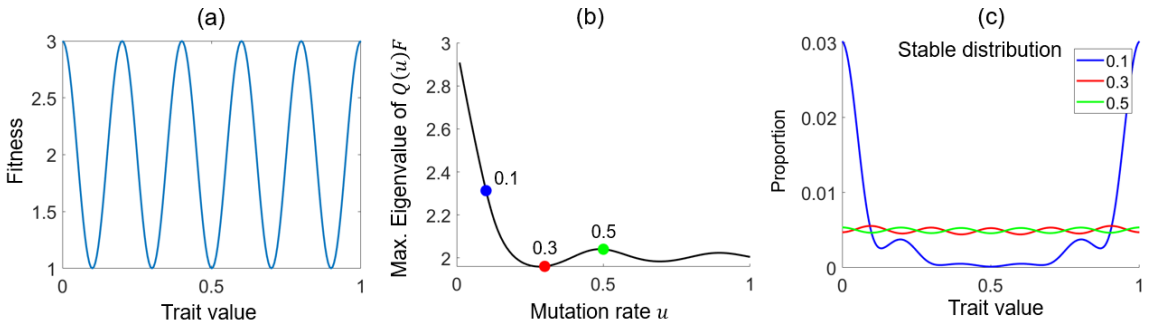


Figure 2.4: A periodic fitness function $f(x) = \cos(10\pi x)$, shown in (**a**), can lead to oscillations in the maximum eigenvalue, (**b**), creating both stable and unstable equilibria. The stable distributions, (**c**), are mostly periodic, though they peak around the boundary for low mutation rate.

Lastly, we found local stable and unstable equilibria for low mutation rates using a simple function, $f(x) = \sin(2\pi x)$, using the circular trait space and a normal spread function, depicted in Figure 2.5. One thing to note is that the mutation rate is low enough that only a few entries per row of $Q(u)$ are nonzero, that is, the spread kernel is not well-approximated. This is not a problem, as mutants are still accumulated to the closest trait. However, this might make the spread kernel effectively biased, which is consistent with the stronger effect seen in coarser discretizations. Interestingly, this effect does not appear in either interval model of mutation. In addition, a different curve is found for a rotated fitness function, which should not be the case, due to the symmetry of the circular model. This may also be an effect of the discretization, as even if the whole fitness is rotated, the particular values may not be, e.g., if the rotation is not a multiple of the spacing. Taken together, these results suggest the importance of specifying the mutation model considered and an intricate trade-off in the discretization of continuous traits. These issues are promising for further investigation, especially with more complicated fitness functions.

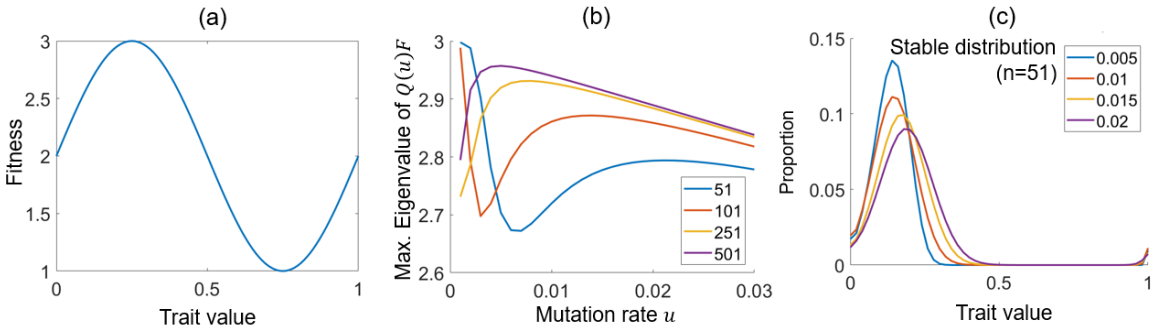


Figure 2.5: A simple fitness function $f(x) = \sin(2\pi x)$, shown in (**a**), yields a local optimum near zero. However this moves as the discretization becomes finer, shown as different lines in (**b**), as opposed to the corresponding plots in the previous figures/no-return-fig, where only one discretization is used (since these do not change with the number of traits used). Panel (**c**) shows how the stable distributions fail to capture the optimal fitness value, possibly from a bias due to the coarse discretization.

Section 2.4

Conclusions

In this work, we examined how mutation rate would evolve in various circumstances. To do this, we employed adaptive dynamics and the replicator–mutator equation, following the work of [7]. Specifically, we computed the maximum eigenvector of the matrix $Q(u)F$ for various values of u to determine the path of evolution. In several models of mutation, we found fitness functions that lead to nontrivial evolution of mutation rate.

First we represented traits as finite strings on a finite alphabet. This is inspired by the possibility of multiple alleles of a single gene, so it could be a good model of reality in simple cases of inheritance. Additionally, the closed form of $Q(u)F$ makes the computations simpler, though it is still unfeasible to determine the maximum eigenvalue for arbitrary F. Then, we examined local models of mutation, which

treats mutation as an incremental change. The challenge in this model is how to most realistically handle the out of bounds mutation, made necessary by the finite nature of the theoretical framework. Accumulation represents a failed attempt to mutate beyond the boundaries, resulting in a trait value at the boundary, whereas truncation effectively ignores such mutations. The issue of mutation beyond the boundaries can be avoided by making the trait space circular, though a circle may be a less natural way of representing traits. In all of these models, we found fitness functions that led to regions where mutation rate will evolve upwards, contrary to intuition. In most cases, this occurred for relatively large mutation rates. When interpreted in rates with appropriate time units, these numerical values found in our cases are realistic, as they represent a normalization for time scale, and therefore could lead to quite small changes between generations.

While this work examined the consequences of a constant fitness function, the method used is easily applicable to other spread kernels and trait topologies or even frequency-dependent game theory models for a range of social and biological applications [40, 219, 253]. For example, one could consider concave spread kernels, local mutation in higher dimensions, or mutation on a graph, as in [79]. The same framework can also be applied when fitness varies with frequency or time. For example, by encoding frequency-dependent selection using a two player symmetric game played in a well-mixed population, we found no selection or selection for lower mutation rates, and observed sensitivity to initial conditions and game parameters. Using a fitness with a single peak that oscillated at various frequencies, we found a sharp transition between selection for low and high mutation, and complicated cycles in the population. In this work, we show that despite simple models, the evolutionary dynamics

of mutation rate cannot be easily predicted.

Chapter 3

Social Learning and the Exploration-Exploitation Tradeoff

Cultures around the world show varying levels of conservatism. While maintaining traditional ideas prevents wrong ones from being embraced, it also slows or prevents adaptation to new times. Without exploration there can be no improvement, but often this effort is wasted as it fails to produce better results, making it better to exploit the best known option. This tension is known as the exploration/exploitation issue, and it occurs at the individual and group levels, whenever decisions are made. As such, it is has been investigated across many disciplines. We extend previous work by approximating a continuum of traits under local exploration, employing the method of adaptive dynamics, and studying multiple fitness functions. In this work, we ask how nature would solve the exploration/exploitation issue, by allowing natural selection to operate on an exploration parameter in a variety of contexts, thinking of exploration as mutation in a trait space with a varying fitness function. Specifically, we study how exploration rates evolve by applying adaptive dynamics

to the replicator-mutator equation, under two types of fitness functions. For the first, payoffs are accrued from playing a two-player, two-action symmetric game, we consider representatives of all games in this class, including the Prisoner's Dilemma, Hawk-Dove, and Stag Hunt games, finding exploration rates often evolve downwards, but can also undergo neutral selection as well depending on the games parameters or initial conditions. Second, we study time dependent fitness with a function having a single oscillating peak. By increasing the period, we see a jump in the optimal exploration rate, which then decreases towards zero as the frequency of environmental change increases. These results establish several possible evolutionary scenarios for exploration rates, providing insight into many applications, including why we can see such diversity in rates of cultural change.

Section 3.1

Introduction

In any learning process, individuals leverage past information along with the opinions of others to decide their best action. Broadly speaking, one can either continue using a strategy that has worked, or try a new approach. While exploration is necessary to discover better strategies, it often results in wasted effort, so it is usually better to exploit the best known strategy. These opposing approaches are very general, applying any time a decision must be made. As such, this concept is relevant across scales, both at the individual such as animal or cells, and group levels, in a wide range of areas from biology to economics [23]. Much work has gone into studying this issue from a variety of perspectives.

One can think of mutation as exploration in the space of genomes. Since all

lifeforms replicate their genetic information, the study of mutation rates has been a longstanding area in biology with significant implications. One theory, the drift-barrier hypothesis, posits that natural selection favors arbitrarily small mutation rates, and is evidenced by relative measures of mutation [149]. Other studies have investigated the mechanisms for viral RNA repair mechanisms, which allow for the mutation rate to evolve up or down depending on which errors get corrected, and phenotypic switching in bacteria, finding that recombination reduces or even eliminates stable non-zero switching rates [68, 142]. There has also been considerable theoretical work on mutation rates in sexually reproducing organisms, finding higher mutation rates can be selected for or against depending on model specifics, such as the type of fitness, whether individuals are haploid or diploid, reproduce asexually or sexually, and if so with or without recombination [35, 197]. Beyond the level of individual cells, decision making in humans has been studied in the exploration/exploitation framework, including its neuroscientific underpinnings [169, 131]. Additionally, this approach has been employed in several areas of ecology, including foraging and analyzing host-parasite or predator-prey systems [73, 150, 166]. Through simulation and analysis, these studies found that exploration rates generally decrease to or stabilize around zero, though factors like limited lifespans or recombination can make exploration less valuable.

Computer scientists have also investigated the balance between exploration and exploitation through evolutionary algorithms, which feature a mutation parameter [57, 71]. This value is critical to the success of the algorithm, however few general techniques guide its tuning. For example, particle swarm optimization is a technique that uses a collection of agents to discover optimal values in a complex space [126].

One approach, known as simulated annealing, decreases the exploration rate over time to concentrate the population around the global optimum. Yet another technique called reinforcement learning has individuals track the performance of a set of possible actions over time to determine the optimal choice [260]. In this framework, one makes an explicit policy for whether new actions are chosen to update these values, exploration, or the current best value is used, exploitation. This area has seen increasing interest from its application to artificial intelligence.

Lastly, there is a significant history of studying exploration/exploitation in economics [8]. Applications include theories of firm's flexibility and understanding product development and innovation [156, 89, 85]. By analyzing economic data, these studies characterize the optimal balance between exploiting present capabilities and exploring new ones. Additionally, March's seminal model of mutual learning in organization, where an individuals and the firm learn from each other dynamically, has been extensively studied and generalized over the last few decades in management science [154, 26, 133, 192]. These dynamics tend to lead to low rates of exploration, which is often beneficial in the short term but detrimental in the long term.

While previous studies have applied a variety of tools from different disciplines, few have asked how the exploration rate changes over time. Our study builds on recent work that applies the technique of adaptive dynamics to answer this, determining the evolution of exploration rate [7, 198]. We extend previous research by approximating a continuous trait space, considering a local model of exploration, and investigating several realistic fitness functions. Specifically, we investigate the evolutionary forces on exploration rates in dynamic environments. Earlier work has considered a finite set of traits, with fitness a function of the current environment that cycles through a

finite set of possibilities, finding the optimal exploration rate was near zero, and zero was a local optima [173, 22]. In contrast, our work investigates local exploration and considers a variety of contexts to determine fitness, broadly divided into two classes. The first uses a feedback mechanism between the strategies in a population and the fitness of those strategies. Specifically, we encode this as the average payoff of an individual when interacting with other players uniformly at random in a population playing a two-player two-action symmetric game. This approach is grounded in the tradition of evolutionary game theory. The other scenario we consider in this work is explicitly representing the fitness of each strategy by a time dependent function. In particular, we consider the fitness landscape with to have a single peak of some width, and whose location oscillates in time in some regular manner. Earlier work has found that in the absence of recombination, the optimal mutation rate maximizes the geometric mean of the fitness of a population [109]. Other studies have also investigated the population dynamics where a game governs fitness as above, but where the game changes over time [212, 247]. This case represents the fact that few environments are static in time, and often undergo periodic changes. For example, if we think of traits as preferred nesting sites in space, then the changing fitness could apply to the study of migration or dispersal. We will interpret some of our results through the concept of an Evolutionarily Stable Strategy (ESS), introduced by Maynard Smith and Price to describe traits that were stable under evolutionary change, which has been an influential idea throughout biology [103, 230]. By using a combination of analysis and simulation, we determine the evolutionary trajectory of the exploration rate in a variety of realistic yet understudied contexts.

Section 3.2

Methods

We think of the set of actions an individual could take as a bounded continuous set, specifically real numbers in the unit interval, and the best action as a trait. This may seem restrictive, but up to linear transformations it can capture any bounded trait one can reasonably assign a number to, for example an organism's height or weight. By putting traits in a space, we can ensure exploration is local, with an exploration kernel to describe the probability distribution of an individual's trait in the near future given its current trait and some exploration rate u. In this work we consider a normal distribution with variation equal to u. Specifically, the model we will use for population dynamics is the replicator mutator equation:

$$\frac{d}{dt}\vec{x} = (Q(u)F(\vec{x},t) - \phi I)\vec{x} \tag{3.1}$$

where $\vec{x}$ is the trait distribution, $Q(u)$ gives the probability of exploration from one trait to another based on a exploration rate parameter u, $F(\vec{x},t)$ is a diagonal matrix with iith entry the fitness of trait i given the population distribution $\vec{x}$ and time t, and ϕ is the average fitness in the population (introduced as in the classical replicator equation to ensure the vector $\vec{x}$ sums to one). Essentially this equation makes individuals reproduce according to their fitness, for example by being imitated through social learning, and explore by the matrix $Q(u)$, depending on their exploration rate u. This framework is built on vectors, so the trait space is necessarily finite. Consequently, exploration cannot occur outside of the boundary, which we handle by truncating these values. Alternatively, one could accumulate them at the endpoints,

or shift them to the other endpoint, making the trait space circular [165].

Equation (3.1) gives the short term population dynamics, and to represent the long term dynamics on exploration rate, we use the approach of adaptive dynamics. In this method, one considers the invasion fitness $f_x(y)$ of a mutant with trait y in a population of individuals all with trait x, defined as their reproduction rate. It assumes mutations on the exploration rate is rare, so it suffices to consider monomorphic populations, as one trait will fixate before the next mutant arises. Further, if we assume these mutations are small, we can use a linear approximation $f_x(y) \approx f_x(x) + (y - x)\partial_y f_x(y)|_{y=x}$. If the term $\partial_y f_x(y)|_{y=x}$ is positive, $y - x$ must be as well for the trait to fixate, so the trait will evolve upwards. The same thing happens if this term is negative, so we can think of it as the rate of change of our trait. This framework was connected to the replicator-mutator equation in [7], which determined $f_x(y)$ was the difference between the maximum eigenvalue of $Q(u)F(\tilde{x}, t^*)$ and the current average fitness, where $\tilde{x}$ is the stable distribution reached by the replicator-mutator equation under the exploration rate x, and t^* is the time the mutant emerges. Using this, we can investigate the evolution of exploration rate if we specify the fitness function $F(\vec{x}, t)$ and use the model of exploration specified above to define $Q(u)$. The code that implements this approach is available in the Github repository `https://github.com/bmDart/exploration-rate-evolution`.

To make fitness frequency-dependent, we consider a population playing a game. Each individual will then receive fitness that is the average payoff received over all possible interactions. Specifically, we will consider two-player, two-action, symmetric games, as these have a large degree of richness in their behavior. In these games, two players interact by each choosing one of two actions, A or B, then receive a payoff

dependent on the pair of actions chosen. There are four pairs, so one write can the payoffs to a player in the matrix

	A	B
A	a	b
B	c	d

where the rows correspond to a player's choice of A or B, and the columns indicate the other player's choice of action. This class is called symmetric, as both players use the same matrix to determine their payoffs. It includes well known examples like the Prisoner's Dilemma (PD), if $b < d < a < c$, where individuals always do better by choosing the second action, even though the best outcome is both players choosing the first. Also included is the less intense version called the Hawk-Dove (HD), also called the Snowdrift, game, if $d < b < a < c$. In this game, the optimal action is the opposite of the other player's action, making this an anti-coordination game. Another game this framework encompasses is known as the Stag-Hunt (SH) game, where $b < d < c < a$. Here the optimal action is the same choice the other player makes, so this is a coordination game.

Strategies in these games can be complex, but if the game only consists of one round, and players have no information about each other, any strategy can be completely described by a probability distribution over the actions, a mixed strategy. Since there are only two possible actions, any strategy is a single number x, the probability of choosing the first action. Then the average payoff to a player with strategy y interacting with a player of strategy x is

$$R(y, x) = ayx + by(1 - x) + c(1 - y)x + d(1 - y)(1 - x)$$

Since this is linear in x, the average payoff of an y player interacting with a population with mean strategy $\bar{x}$ is just $R(y, \bar{x})$. Since this is also linear in y, so the average payoff over this population is $R(\bar{x}, \bar{x})$. This function, $R(x, x)$, can give some insight, so we will refer to it as the population fitness of a strategy x. Interestingly, this can look differently within a class of games, for example $(a, b, c, d) = (2, 0, 2.5, 1.5)$ and $(2, 0, 5, 1.5)$ are two prisoner's dilemmas, yet the first makes $R(x, x)$ concave up while the other is concave down. We will see that increasing exploration rate often leads to a more spread out stable distribution, which moves the average strategy closer to 0.5, and see the effect this will have on a population's fitness. We can also apply Adaptive Dynamics, thinking of $R(y, x)$ as $f_x(y)$ the fitness of an invading strategy y into a resident population of all x-players. Here we see the strategy should evolve according to

$$\partial_y R(y, x)|_{y=x} = ax + b(1 - x) - cx - d(1 - x)$$

This is a line connecting $b - d$ at $x = 0$ to $a - c$ at $x = 1$, so there are essentially four cases depending on the relative signs of these terms, as shown in Figure 3.1 with representative games, and arrows indicating the dynamics of the invading strategies. Since the diagonal cases are essentially mirrors, we consider just the three games mentioned above.

Lastly, we consider time-dependent fitness functions where the fitness landscape has a single peak, of some width, whose location oscillates at some frequency. In particular, we take the fitness at time t to be the normal distribution with variance 0.1 and mean $(1 + \sin(\omega t))/2$, as this oscillates between the endpoints zero and one with period ω. We investigate the effect of changing this period and also the variance of this distribution.

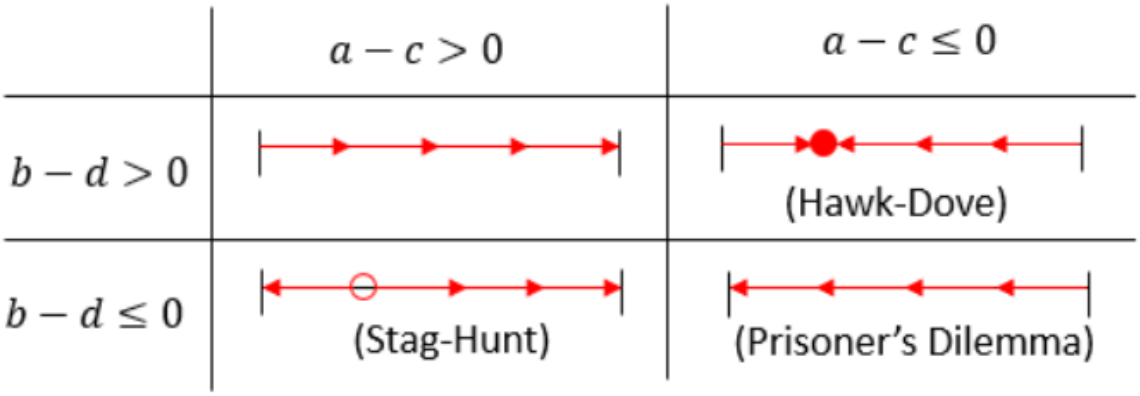


Figure 3.1: The four possible cases for the evolution of strategies in two-player two-action symmetric games. Depending on the entries of the payoff matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, either one endpoint will be attracting, or there will be an interior equilibrium that is either stable or unstable. The corresponding cases are labeled with their archetypal game.

Section 3.3

Results

First, we used the payoff matrix

$$\begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}$$

for the Prisoner's Dilemma, finding that the replicator-mutator equation stabilized at the distributions given in Figure 3.2. These show that lower exploration rates more closely exploit the optimal strategy of defection, as expected. However, those populations have lower fitness, as higher rates of choosing the second action, defection, are worse for the population overall, since the population fitness $R(x, x)$ is increasing for this game. Despite higher exploration rates leading to a population with greater fitness, we see the invasion fitness is only positive for lower exploration rates, so it can only evolve downward. This is a dilemma, as it is better to have a large exploration

rate and flat trait distribution, but this will be selected against.

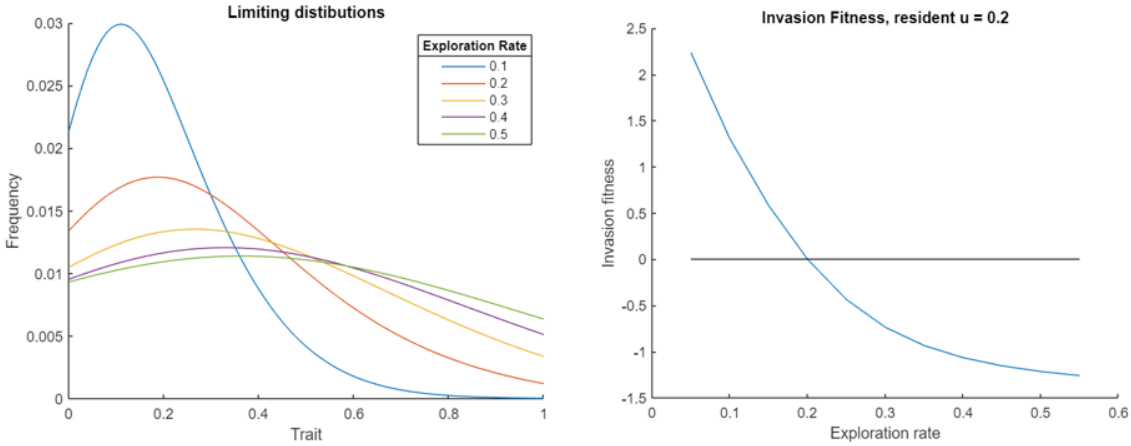


Figure 3.2: On the left we see the stable distributions of the replicator mutator-equation in the PD game for various exploration rates. The right plot shows the invasion fitness as a function of the invading exploration rate, for a resident value of 0.3. Since this is only positive to the left of the resident value, exploration rates can only evolve downwards. This is representative of all resident values.

The next game we considered is the Hawk-Dove game, with payoff matrix

$$\begin{bmatrix} 1-c & 2 \\ 0 & 1 \end{bmatrix}$$

where c is a parameter representing the cost of competing over a contested resource of value one. In this case, we see populations approach the equalizer strategy $cH + (1-c)D$, which makes all strategies have the same fitness, so there is no selection. Consequently, there is no selection on exploration rate, so it will be subject to neutral selection. This is consistent with the results of [7], which found multiple mutation rates could coexist in this game. Like in the previous game, different stable distributions are reached for different exploration rates. Here we see increasingly uni-

form distributions as the exploration rate increases, in Figure 3.3, which is expected, as this represents larger exploration. Surprisingly, we see a dependence on the game parameter c. When this is not 0.5, the population does reach the equalizer strategy c, as seen by plotting the average strategy over time, also in Figure 3.3. This results in downward selection on exploration rate. Despite all Hawk-Dove games have the same strategy dynamics from the perspective of a single player, this population model demonstrates different effects depending on a parameter's value. Interestingly, despite exploration rates evolving downwards for $c \neq 0$, the population's fitness can either increase or decrease with exploration rate depending on if c is above or below one half, as seen by considering the population fitness $R(x, x)$. Thus as in the PD, it is possible the exploration rate will evolve towards value that are worse for the population overall.

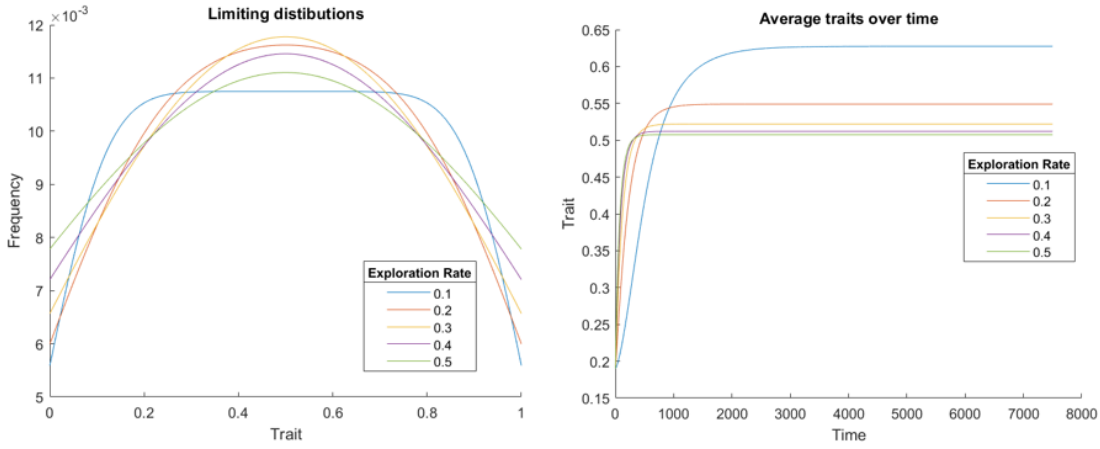


Figure 3.3: In the Hawk-Dove Game, we see flatter stable distributions, the first plot, for large exploration rates for $c = 0.5$. However, for $c \neq 0.5$, the average strategy does not reach the equalizer strategy c, shown in the second plot.

The final game we considered was the Stag-Hunt, with payoff matrix

$$\begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix}$$

In this case, the population reaches unimodal distributions as in the Prisoner's dilemma, with individuals favoring one option more than the other. This is because the optimal action is to choose the same action as the other player, so the population becomes increasingly concentrated towards whichever pure strategy the initial mean was closer to. As such, the population will evolve away from the unstable equilibrium of 0.5, as in the single player dynamics. Depending on whether the initial mean is above or below 0.5, the population fitness is either increasing or decreasing with exploration rate, since this moves the mean strategy closer to a half, which is good if the population is concentrated around one but bad if it is concentrated around zero. Despite this, exploration rates can only evolve downwards in both case, so in this game too, exploration rates can evolve to less desirable levels. However in this case, selection becomes neutral for sufficiently large initial exploration rates, since the population becomes centered around 0.5. Thus, exploration rates that start large will drift up and down, but eventually become caught around zero.

The other type of fitness we considered in this work had fitness an explicit time-dependent function, with no dependence on the distribution of strategies in the population. Specifically, we took $f(x,t) = \exp(-(x - (1 + \sin(\omega t))/2)^2)$ where ω is a parameter for the period of the oscillations. Here, one may also use the replicator-mutator equation to simulate the population dynamics, but now populations need not reach stable distributions. For example, a periodic fitness will lead to periodic

changes in the population. Nonetheless, we can adapt the results of the model by considering a time averaged fitness. Since the fitness function does not depend on the frequency of strategies, an invading subpopulation with a novel exploration rate will grow independently of the resident population. Thus, the exploration rate leading to a higher average fitness will eventually fixate. Here one must use the geometric mean of fitness, as populations grow geometrically. This is because fitness is essentially a reproduction rate, which are multiplied, not added, together to aggregate over time periods, as is done in the geometric mean. Indeed, the order of the geometric and arithmetic mean might swap between two sets, for example $\{50, 50\}$ and $\{100, 1\}$.

In Figure 3.4, we plot the time averaged fitness of each exploration rate, using fitness functions of various periods. For small periods, we see fitness is maximized around zero, decreasing with larger exploration rates until it reaches a local minimum then starts to increase. This means that for rapidly changing environments, it is best to have minimal exploration rate. However, if it starts above this minimum, exploration rates will increase arbitrarily high. This suggests some rates result in the population lagging behind the optimal strategy, to the extent that a uniform distribution is more effective. We see the opposite curve for sufficiently large periods, where environmental change is slow. Here, there is a local maximum at some nonzero exploration rate, indicating an intermediate level of exploration is optimal. Interestingly, as the period changes, the optimal exploration rate makes a jump from zero to an intermediate value. The exact period where this occurs and value the optimal rate jumps to would depend on specifics of the model, namely the type of curve defining the fitness. Further, optimal rate decreases with increasing period, that is, slower changing environments. This makes sense, as a sufficiently slow chang-

ing environment is effectively stable, for which arbitrarily small exploration rates are usually optimal. Comparable effects in the evolution of exploration rate are observed when the normal distribution has wider variance, indicating the generality of these results. Theoretically, one could also compute the time averaged fitness of the limiting exploration rates. When exploration rate is zero, the population will likely be entirely at the strategy that maximizes the time averaged fitness function, and when it is infinite, the strategy distribution will be uniform, so the time averaged fitness will simply be the average value of the function (which is constant in time, so equals its time average).

Section 3.4

Discussion

In this work, we investigated the evolution of exploration rate under variable selection, employing adaptive dynamics and the replicator-mutator equation. For frequency dependent fitness encoded by two-player/two-action, symmetric games, we found that exploration rates often evolve downward, but neutral selection is also possible. This means in most cases the exploration rate of zero (approaching from above) constitutes an ESS. Despite this, it is possible in all games we considered for larger exploration rates to be more beneficial to the population. The precise form of the exploration rate's evolution can also depend on the game's parameters or the initial conditions, as in the HD and SH games respectively. This suggests that while the cases we studied are representative of the possible dynamics in this class of games, further richness could be observed in future study. However, we conjecture that this class of games is incapable of selecting for large exploration rates, as opposed to the more com-

plicated class of two-player/three-action symmetric games, where it was found the Rock-Paper-Scissors game led to selection for an intermediate level of mutation [198]. This is because cyclic dynamics in this game allow for a sub-population with multiple traits to remain resilient as the composition of the population changes. Since cyclic dynamics cannot be observed in the smaller class of games, it is likely that larger exploration rates cannot be selected for with these types of fitness function. Then with fitness function a single peak that oscillated according to some frequency, we found both attracting and repelling equilibria depending on the period. For fast changing environments, arbitrarily small exploration rates are optimal, though a sufficiently large initial exploration rate will evolve upwards. In contract, slow changing environments have intermediate optimal exploration rates, and evolution proceed towards this. As the rate of oscillation decreases further, this optimal value approaches zero [258].

Previous work on exploration/exploitation has applied a wide range of techniques to determine the optimal exploration rate, but not how it evolves. Our research answers this question in a variety of environments, finding lower rates are usually selected for, but time dependent fitness can make nonzero or even arbitrarily large exploration evolve. We expand on previous research by using a local model of exploration, continuous approximation of traits, and the approach of adaptive dynamics. The results we find complement experimental data from biology, providing additional theoretical evidence for the drift-barrier hypothesis, as seen in our consideration of frequency dependent fitness. These findings could also help explain how behaviors like foraging evolve, or the benefits and harms of different levels of conformity in social groups [75] and the dynamics that select for or against exploration. Our results

suggest that genetic algorithms [53] might be improved by generalizing the mutation to include the exploration rate itself, rather than tuning this value manually. Lastly, the findings of this study could imply that the optimal tradeoff between exploration and exploitation found in companies can be reached through market forces [132], as these are analogous to natural selection.

Section 3.5

Conclusions

Exploration and exploitation are two competing effects that combine to determine an optimal strategy. Since one can never know if their current approach is best, exploration is necessary to some degree to find better strategies. However, this often produces worse results, especially when the current method performs well, so it is also important to balance this exploration of new ideas with exploitation of the best known ones. This general dilemma is found in many areas of research, including biology, ecology, computer science, and economics.

The goal of our research was to see what resolution between these two factors is reached through natural selection, by studying the evolution of an exploration rate parameter. To answer this, we applied adaptive dynamics to the replicator-mutator equations, which gave the invasion fitness of exploration rates in various contexts. We consider fitness that depends on the relative frequencies of other traits, and fitness that varies explicitly in time. In the first case we found exploration rates often evolved towards zero, which was then an ESS, though trajectories could vary based on game parameters or initial conditions. In the second case, we observed a discontinuous transition in optimal exploration rate, from zero to an intermediate

value which decreased back to zero as environmental change slowed.

The generality of this framework makes it ideal to study several related question about exploration versus exploitation. Future study could consider other trait topologies, by adapting the matrix $Q(u)$. For example, [165] explores how mutation rates can evolve upwards even in a fixed environment. This is found not just for traits in some interval, but also in a circular space or finite strings on a finite alphabet. Preliminary results showed the HD game with $c = 0.5$ led to increasingly polarized distributions are exploration rates decreased, if exploration outside of the interval was accumulated at the endpoints. In addition, making the trait space circular caused the time averaged fitness to strictly decrease with exploration rate, indicating that in the absence of asymmetry, there is no benefit to an intermediate level of exploration. Multidimensional trait spaces could also be considered, but without dependencies between the axes, this may reduce to several copies of a one dimensional trait. The fitness functions could also be changed within this framework. For example, one could consider fitness that comes from nonlinear or multiplayer games, like the Public Goods Game, or stochastic fitness functions, such as jumping to a random position at some constant frequency, or some constant positions with some random frequency. Recent studies suggest that interesting results could been found in this area [36, 157] One could even make the exploration rate itself non-constant, possibly modeling it as a decreasing function of time, such as a linear or exponential function, and study the evolution of the parameters of these functions. Lastly, one could experiment with more intelligent agents. Whereas agents in this model explored randomly, one could use a reinforcement learning framework like Q-learning to model agents who explore based on previous knowledge. Such modifications would certainly change the balance

of importance between exploration and exploitation, likely leading to different results.

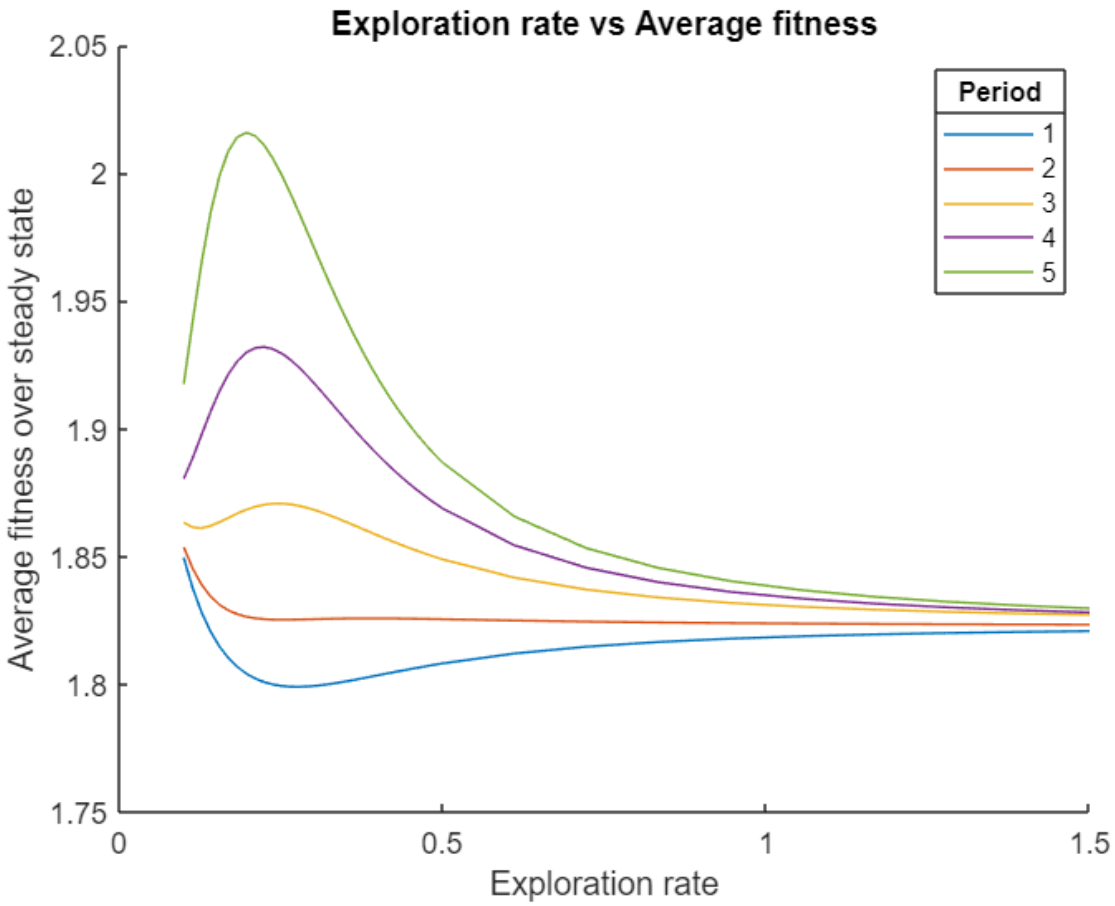


Figure 3.4: For fitness that explicitly depends on time, we see both unstable and stable equilibria in the evolution of exploration rate. When environmental change is slow, corresponding to a long period, an intermediate level is optimal. Whereas for fast environmental change arbitrarily small exploration rates are optimal, though if the initial value is large enough, they will become arbitrarily large.

Chapter 4

Evolutionary Multi-Agent Reinforcement Learning in Group Social Dilemmas

Reinforcement learning (RL) is a powerful machine learning technique that has been successfully applied to a wide variety of problems. However, it can be unpredictable and produce suboptimal results in complicated learning environments. This is especially true when multiple agents learn simultaneously, which creates a complex system that is often analytically intractable. Our work considers the fundamental framework of Q-learning in Public Goods Games, where RL individuals must work together to achieve a common goal. This setting allows us to study the tragedy of the commons and free rider effects in AI cooperation, an emerging field with potential to resolve challenging obstacles to the wider application of artificial intelligence. While this social dilemma has been mainly investigated through traditional and evolutionary game theory, our work connects these two approaches by studying agents with an

intermediate level of intelligence. We consider the influence of learning parameters on cooperation levels in simulations and a limiting system of differential equations, as well as the effect of evolutionary pressures on exploration rate in both of these models. We find selection for higher and lower levels of exploration, as well as attracting values, and a condition that separates these in a restricted class of games. Our work enhances the theoretical understanding of recent techniques that combine evolutionary algorithms with Q-learning, and extends our knowledge of the evolution of machine behavior in social dilemmas.

Section 4.1

Introduction

The world has recently seen a surge of innovations powered by advanced artificial intelligence technologies, such as Large Language Models and self-driving cars, promising to fundamentally reshape many aspects of the world. As progress in these areas continues, we will see such systems deployed more extensively throughout the world. This has already created complex systems of interacting agents with different goals and patterns of behavior. Understanding the theoretical basis of these systems will be crucial for successful implementations, and require new approaches with interdisciplinary ideas [236]. In particular, a pressing open question is how to ensure models act cooperatively while performing their given task [59]. This is related to the issue of aligning the incentives used to train AI models with those of the broader society.

Reinforcement learning (RL) is a prominent framework for AI that has been successfully applied to many challenging problems due to its exceptionally general approach [115]. While many machine learning techniques have constraints on the types

of problem they can address, this approach can be applied to a broad range of problems. The key idea behind reinforcement learning is simply that actions that have a positive payoff will be repeated more, and those with a negative payoff will be repeated less [248, 226]. This technique has its basis in models of animal psychology but has since found a series of cutting-edge applications due to its flexibility. Reinforcement learning has been applied to a wide variety of problems from solving traditional games like Go, to making stock price predictions, managing energy systems, and controlling chaotic dynamics [18, 191, 33]. But this approach can be unpredictable, especially when multiple agents learn simultaneously, as this creates a dynamic environment [138, 120, 171, 25, 28]. Such Multi-Agent Reinforcement Learning (MARL) systems have seen many applications such as navigating groups of autonomous vehicles and distributing resources through communication systems [90, 70]. This field draws attention from researchers studying complex systems, engineers seeking to improve algorithms, and policy makers trying to manage these technologies effectively.

Recently there has been growing interest in combining reinforcement learning with evolutionary algorithms, a similarly general approach with simple motivation. Evolutionary algorithms use mutation and selection to solve complex problems [162, 139, 229, 92]. Reinforcement learning has a natural connection to this method as the learning of a single agent is equivalent to evolutionary dynamics between the strategies that agent can use, under a particular form of mutation [237, 120, 15]. Thus evolution between agents themselves can be seen as a form of multi-level selection, which has long been investigated by evolutionary theorists. The majority of the work in this intersection focuses on comparing and improving different algorithms, as

opposed to understanding its theory [209, 69, 167, 145, 256, 121, 13]. Many MARL architectures have been proposed, with studies determining when each is optimal for various applications, but lacking broader conclusions [264, 144, 228, 146].

Our work seeks to enhance the theoretical understanding of these evolutionary MARL systems, eventually allowing a more principled approach to designing algorithms for MARL. We investigate the effect of selection on a parameter governing the degree to which agents explore new strategies, and how this depends on the game that governs agent interactions. Due to its relative analytic tractability, we focus on one of the foundational reinforcement learning models, Q-learning. We combine this with ideas from evolutionary theory to study dynamics within this system. In particular, we focus on the evolution of learning in a classical social dilemma, the public goods game, to investigate cooperative AI. These are a canonical example of conflicting incentives, the interests of the self and those of the collective, and are suitably broad to encompass many scenarios including the sustainability of common resources, institutional incentives such as team bonuses, and collective endeavors like combating climate change [227, 205, 101].

Most prior investigations into evolution in social dilemmas have studied simple, often static, strategies [97, 128, 164, 165]. Our work extend this to more realistic, learning agents. Likewise, evolutionary MARL generalizes techniques known as particle swarm optimization and simulated annealing to explore more efficiently [241, 215]. Previous studies into reinforcement learning in social dilemmas have mainly focused on the Prisoner's Dilemma [238, 239, 2, 94, 245, 16, 200, 111, 116, 155, 152, 243, 261, 183]. However, this game is limited to interactions between two individuals, as with most theoretical studies of MARL, limiting the complexity that can be found and

reducing the potential applicability of the findings. Some recent studies have considered MARL in the public goods game, investigating the effect of learning models and how to optimize cooperation [136, 170, 246, 244, 153]. The majority explore whether MARL can explain test subjects' behavior in experiments, [110, 107, 113, 10, 34, 55, 98]. Our study extends these preceding works by revealing a variety of evolutionary dynamics between learning agents in a group social dilemma.

Section 4.2

Model

In this work, we expand the MARL framework to include evolutionary dynamics between agents. Inspired by evolutionary game theory, we investigate population dynamics in a model where agents reproduce and die concurrently with learning. We study this through extensive agent based simulations, a stochastic model, and an evolutionary analysis of a corresponding deterministic system of ordinary differential equations.

Our model uses the foundational Q-learning algorithm to perform reinforcement learning. Each agent keeps a table of values for each possible state s of the system and action a. These are updated according to

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{4.1}$$

where s_{t+1} and a_{t+1} are the updated values given the state s_t and action a_t at time t, α is the learning rate that determines how quickly values are updated, r_t is the reward received, and γ is the discount rate governing the extent the agent cares about

future rewards. In essence, this generalizes the notion of keeping a weighted average of the rewards for a particular action, also considering the change in state an action will cause. Actions can then be selected in a variety of ways. We use the Boltzmann function to determine policies, as it is better suited to mathematical analysis. With this method, actions are selected at random proportionally to the exponential of their payoff divided by a parameter T. For a finite set of actions A, the strategy of the agent is given by the probability of choosing each action a in A given the state s, equal to

$$P(a_t = a|s_t = s) = \frac{\exp(Q(s_t, a_t)/T)}{\sum_{b \in A} \exp(Q(s_t, b)/T)} \tag{4.2}$$

This is known as the "temperature" of an agent, since it controls the degree to which the agent will randomly exploration new strategies. As T approaches zero, only the action with the highest Q-value will be selected, corresponding to purely exploiting this best-known strategies. Temperature can vary with time or payoff relative to some aspiration level, though it can also be a constant value. We focus on stateless Q-learning, as the variable group composition makes it challenging to specify meaningful states. This means there is effectively a single state, and the transition function between states is $s_{t+1} = s_t$ regardless of the choice of actions.

In this study, rewards are determined by the public goods game. This is a natural setting for arbitrary numbers of agents to interact, and has a long history of being used to understand group social dilemmas like the tragedy of the commons and free-rider effects. Each agent, of a group of N, chooses whether to pay a cost of one to contribute, or not contribute, referred to as defection. The total contribution is then scaled by a reward function $f(x)$ and distributed evenly. The payoff of individual j

is

$$\pi_j(c_1, ...c_N) = \frac{1}{N} f\left(\sum_{i=1}^{N} c_i\right) - c_i \tag{4.3}$$

where c_i is the contribution, zero or one, of individual i. Typically the reward function is linear $f(x) = kx$ with $1 < k < N$, but it can also contain more complicated nonlinear effects. For this work, we include the division by N in the definition of $f(x)$ to more easily interpret it as the reward per individual, simplifying comparison between values of N. We'll represent these functions at the possible discrete levels of contribution with vectors $[f(0), f(1), f(2), f(3), ..., f(N)]$, since the intermediate values cannot be realized so are irrelevant. This allows for more precise control of the rewards, though it has the drawback of being harder to extend to larger groups.

Our first approach is an agent based simulation of a variation on a classical stochastic model of evolution, the Moran (death-birth) process[177]. A finite number of N individuals interact to receive payoffs. At each time step, every individual has an independent probability r of dying, then being replaced by the offspring of another member of the population, where the new individual inherits all the learning parameters and Q-values of the parent. The individual who gives birth for this replacement is selected proportionally to their fitness $f_i = e^{\beta \pi_i}$, the exponential of their average payoff π_i over all rewards from previous interactions. Using the exponential of fitness is a standard technique to avoid complications from negative fitness. The parameter β gives the strength of selection, in this study we focus on $\beta = 1$ for simplicity. Concurrently with the replacement, individuals perform Q-learning under Boltzmann Selection, with individual parameters T, α, and γ, to receive rewards from the public goods game with reward function $f(x)$. This function could be linear such as $f(x) = kx$, which is most commonly considered, or contain non-linear effects such as

$f(x) = b_0x + b_1x^2$. Sample trajectories of this simulation are shown in Fig. 4.1. Death occurs uniformly at random at each iteration of the traditional Moran process, so on average individuals only learn for N iterations before being replaced. Our probabilistic modification gives agents more time to learn, an expected $1/r$ iterations, since the independent death probability means the iterations individuals learn for follows a geometric distribution with parameter r. Intuitively, the replacement rate parameter r balances between learning, when it is low, and evolution, when it is high. We can then use this to estimate the fixation probability that a mutant will replace the resident type. These determine the evolutionary trajectories under rare mutations. They can determine whether there will be positive or negative selection, or other effects like attractors and repellors of the dynamics.

Complementing this agent based simulation, our second model applies the evolutionary technique of adaptive dynamics using a limiting system of differential equations[63, 64, 43, 29]. This is described more fully in the Appendix.

Section 4.3

Results

There is a significant effect from noise in the stochastic model, due to randomness in the action choices, which agents are selected to die, and which replace them. Because of this, careful consideration must be given to the model parameters to obtain meaningful results. If the rewards are large relative to the discount rate and temperature, agents may prematurely fix their strategy and not explore sufficiently. In addition, evolutionary dynamics are inherently more volatile in smaller populations, so larger groups or more trials are necessary. We control for these and further account for the

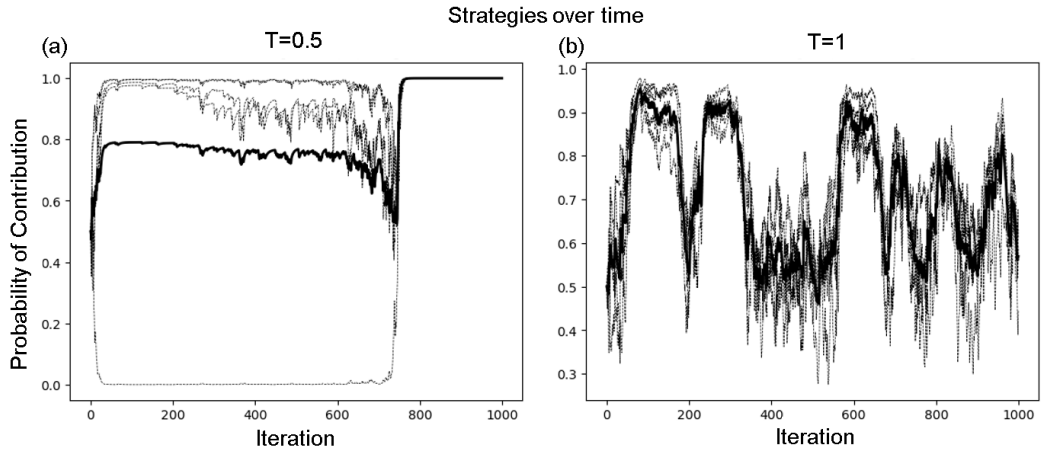


Figure 4.1: **Stochastic learning dynamics with symmetric temperature**. These plots show the trajectories of strategies in the agent-based simulation over time as dotted lines, with the average strategy in bold, where the rewards are [0, 0, 0, 2, 4, 6], $N = 5$, $\gamma = 0$, $\alpha = 0.1$, $r = 0$, and $T = 0.5$ in panel (a) and $T = 1$ in panel (b). By varying the temperature, a range of behaviors are seen. For low temperatures, relative to learning rate and rewards, agents enter a self-reinforcing cycle where they choose the most beneficial action repeatedly. For large temperatures, the strategies fail to converge. We see good alignment with the predictions of the ODE model, that strategies cluster together when the temperatures are the same. However, the averaging in the ODE model and the assumption of small replacement rate limit the direct predictions it can make of the agent-based simulation.

effect of the other parameters by holding all fixed except the parameter of interest, repeating this process for several combinations of the other parameters to ensure the results are robust. Figure 4.2 plots the average probability of contribution depending on the learning rate and discount factor, showing a large learning rate and small discount factor are ideal in these cases. In other games, it is possible large discount factors or low learning rates are better. The two games shown use linear reward functions $f(x) = kx$ with $k = 0.9$ and 1.1 so the jumps in reward for an additional contribution are slightly above and below the cost of contribution of one, so learning should always favor contributing in the former case, and not contributing in the latter. Despite this, they can result in similar levels of cooperation, as the payoff is mostly determined by the actions of the other group members given the large group size. The combined effects of learning and evolution, given by the temperature and replacement rate parameters, are presented in Figure 4.3. Depending on the temperature we can see positive or neutral effects from replacement rate in a particular game. Across the replacement rates we can see increasing temperature initially increases the contribution level, then decreases it to 0.5 as agents purely explore. This suggests intermediate temperature values perform best, which is consistent with previous findings. Even within a single game, different effects from these forces are observed.

Our deterministic model reveals a range of selection effects on temperature for different reward functions. We find examples where temperature experiences positive and negative selection, as well as attraction to an intermediate value. To understand how the game influences this, we study the direction of selection across all possible reward functions. While the invasion fitness is positive, we repeatedly take mutant values slightly above resident values. This gives a sense for the most likely evolu-

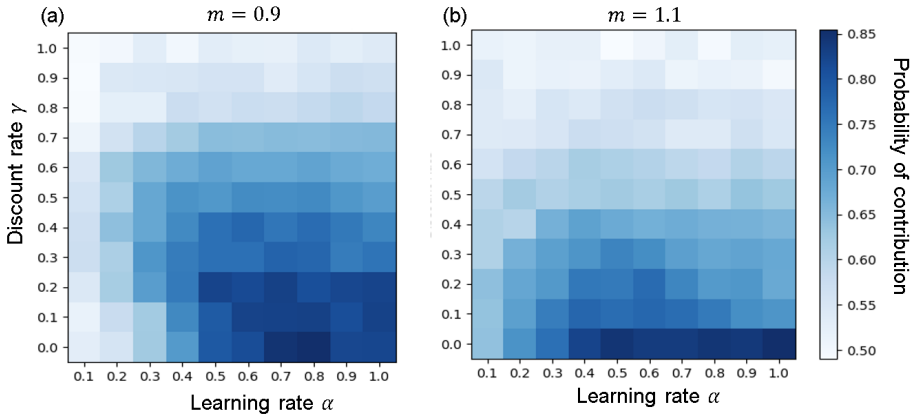


Figure 4.2: **The optimal learning parameters vary with the reward function**. This plot shows the average, over 100 runs, strategy in the group after 500 iterations where the horizontal axis is the learning rate and vertical axis is the discount factor, both between zero and one. Here $r = 0$ so there is no replacement, the temperature is $T = 0.5$, the population consists of five agents, and the reward function is linear $f(x) = kx$ with $k = 0.9$ on the left in part (a), and 1.1 on the right in part (b). In these cases the jumps are a constant of k, so in the first it is always slightly better to defect, despite this agents contribute 85% of the time with a higher learning rate and low discount rate. Similarly, on the right it is always slightly better to contribute, and a wider range of learning rates achieve a similar probability of contributing. Because of the nonzero temperature, it is impossible to achieve perfect cooperation, a strategy of one, and the achieved values are approximately the largest achievable given this temperature and the rewards for each action.

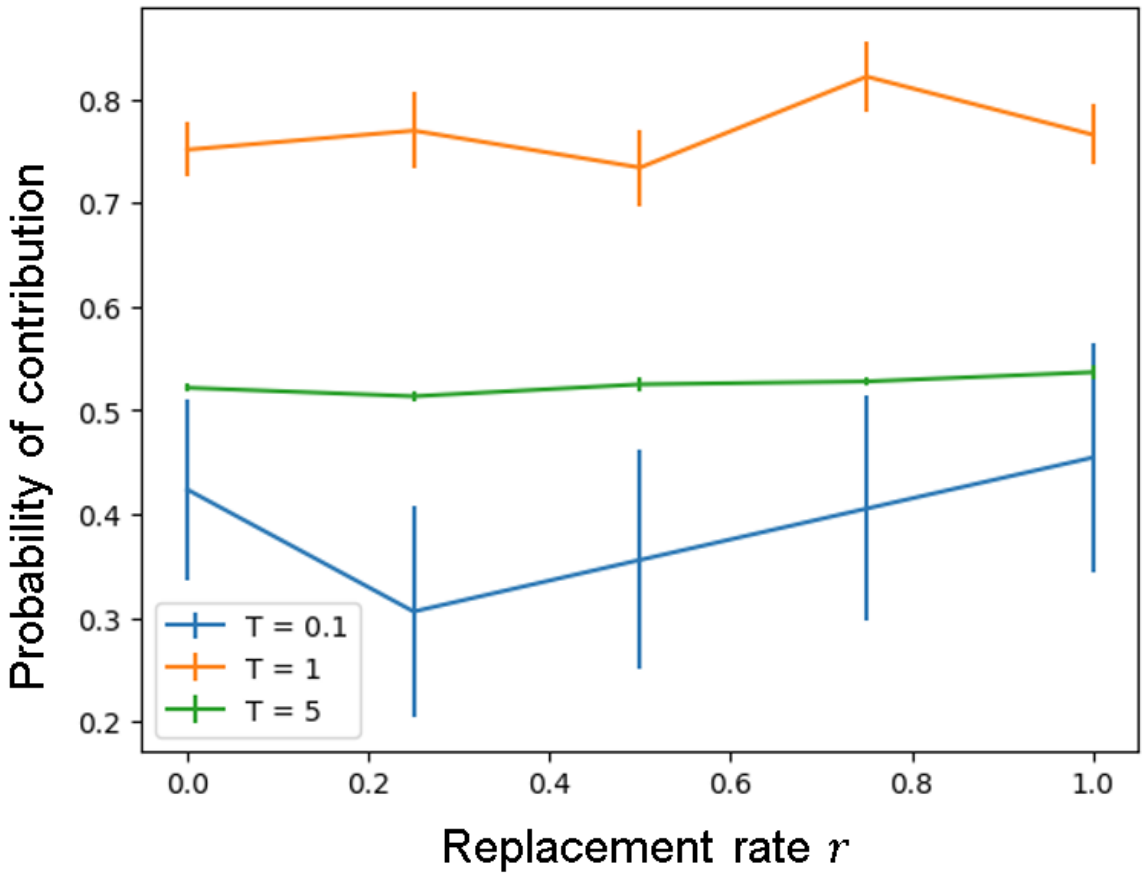


Figure 4.3: **Learning and evolution can have varying effects on contribution levels**. By varying the learning rate T and replacement probability r, one can tune the relative strength of learning and evolution. This plots the average, plus or minus one standard error, over 20 runs of group level of contribution after 1000 iterations. These simulation are initialized with a single temperature, with no mutation in temperature, so selection is only acting on the strategies.

tionary trajectory of the temperature. Since there are $n+1$ values in the reward function for a group of n players, we restrict to small cases to assist with finding patterns. In particular, we consider reward functions of the form $[0, j_0, j_0 + j_1, m]$ where $0 \leq j_0 \leq m$ and $0 \leq j_1 \leq m - j_0$. These correspond to a constraint where there is an upper limit m on the per individual reward, and that the reward function is weakly increasing. Since there are only two parameters for a fixed m, we can plot the final temperature over the j_0, j_1-plane, shown in Figure 4.4. When $m > 3$ we see there is a clear separation between regimes, selection is positive when $j_0 + j_1 = f(2)$ is above a threshold depending on m. For smaller values of m, selection is predominantly negative. The distinction is that for $m > 3$, it is possible for all jumps to be above the cost of contribution one, where contribution would always be beneficial. This corresponds to the region $j_0 > 1$ and $j_1 > 1$, which does experience consistent selection on temperature. The computation of fixation probabilities generally supports these results, though computational constraints limited further investigation into these.

Section 4.4

Discussion

Our results suggest that the relationship between learning parameters and cooperation levels is quite sophisticated, and warrants further study. The temperature of agents can evolve in different ways depending on the environment. This suggests that combination of evolutionary algorithms with reinforcement learning can be used to optimize cooperation. Our approach of computing fixations probabilities is more interpretable than traditional evolutionary algorithms, providing a broader picture of the evolutionary trajectories that might be taken. Analytic approaches that average

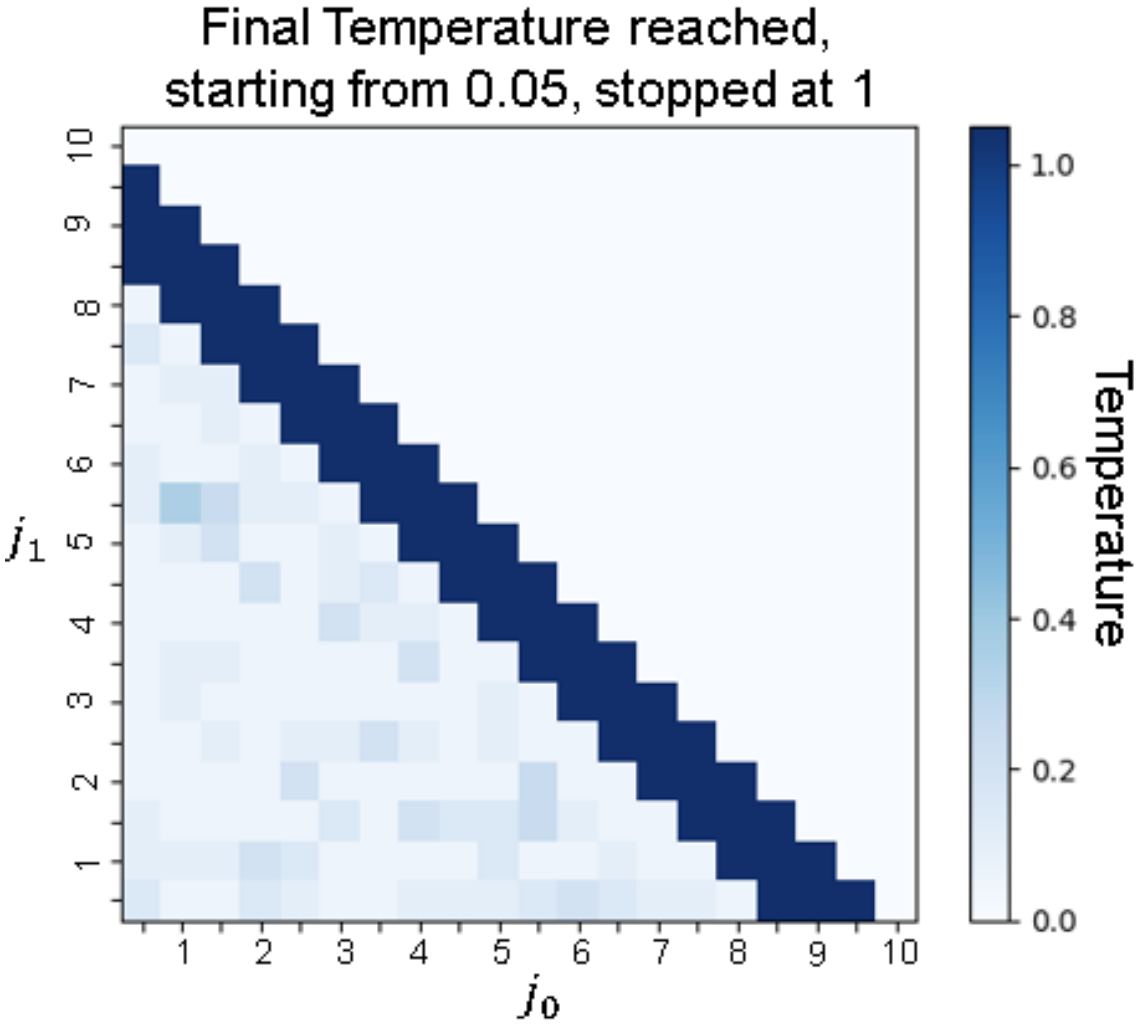


Figure 4.4: **The reward function can lead to positive or negative selection**. This plot represents the most likely outcome of the evolutionary dynamics in the temperature parameter, starting from $T = 0.05$ and up to $T = 1$, over the space of possible reward functions for the three player game, found through the adaptive dynamics approach described in the appendix. Letting m be the maximum reward for when all individuals contribute, we can specify the function as $[0, j_0, j_0 + j_1, m]$ where j_0 and j_1 are the jumps in reward when an additional individual contributes, if zero or one other had already contributed. Assuming the reward function is increasing, we have $0 \leq j_0 \leq m$ and $0 \leq j_1 \leq m - j_0$, so only the values in the lower triangle are considered. We see there is a clear transition to larger final temperatures when $j_0 + j_1$ exceeds a threshold depending on m, in this case $m = 10$.

over all possible interactions can improve the efficiency of these investigations and allow for more robust results through reducing stochastic effects.

While a multitude of reinforcement models have been proposed, we focus on the foundational Q-learning due to its theoretical guarantees and relative analytic tractability. Our analytic approach to studying these dynamics makes a few strong assumptions. We replace the reward an action receives with the average over all possible interactions, losing a large amount of data. Consequently, agents with the same temperature are expected to follow the same learning dynamics, often contrasting simulations where they can easily diverge based on initial actions. Future work could allow a more complete description of every state of the system, extending previous mean-field approaches that characterize the evolution of the probability distribution of strategies in the population [82, 259, 37]. This would undoubtedly require a far more sophisticated mathematical model, complicating the analysis. Another significant factor is the lack of states in our reinforcement learning model. The state could simply be the number who cooperated in the previous round, or an average over the last several rounds. However this is a rather inaccurate measure of the true state: the strategies of all agents.

Section 4.5

Conclusion

In this work, we extended the MARL framework to allow for reproduction, introducing evolutionary pressures between agents. While previous studies have studied evolution on individuals with static strategies, or fixed groups with variable strategies, our model combines these two dynamics to create a more realistic system than

either alone. By doing so, we are able to broaden the applications of evolutionary game theory, and bring theoretical insights into a complex learning model. We apply evolutionary techniques to study the dynamics in the temperature of agents, a key parameter governing their degree of exploration. We use agent based simulations to estimate the fixation probability of mutations in this parameters. These probabilities determine the evolutionary trajectories under the assumption of rare mutations. This assumption also allows us to apply adaptive dynamics using a system of ordinary differential equations to remove stochastic effects.

Through extensive simulations, we explored the intricate relation between learning parameters and cooperation. In particular, temperature and replacement rate could have a range of effects depending on other parameters like the reward function determining the type of public goods game being played. Depending on this, the temperature could evolve up or down, or to an intermediate levels. By studying a restricted class of these games we conjectured a condition that determines which type of selection the temperature will undergo.

There are multiple possible extensions of this work. This framework could be applied to a number of other games, such as the Iterated Prisoners Dilemma or coordination games. A round number of continuation probability could be used to determine how long the group interacts for, allow for more or less learning to occur. By restricting to simpler cases, like small population sizes, we could derive further theoretical results. For example, the Moran process we simulated in model one could be explicitly represented to analytically determine the fixation probabilities. While the analytical results are often limited to simple cases, we could extend the simulations to capture more complicated effects, for example by including mutation in

the parameter values. We could also consider a model where the initial strategy is also genetically determined, which would allow a more effective comparison between learning agents with nonzero temperature, and those who never change their strategy. We could also investigate the full replicator dynamics using the system of differential equations to determine finesses. This would provide a clearer picture of the dynamics between individuals with different temperatures. Future study in this direction has the potential to greatly enhance our understanding of the complex dynamics in Multi-agent Reinforcement Learning.

Our approach expands the theoretical understanding of combining genetic algorithms with reinforcement learning. While evolution can and has produced remarkable solutions to many challenging problems, some situations can be ill-suited to this approach. Our investigation highlights the fact that careful consideration must be given to minimize the stochastic effects that can overpower selection. Such understanding is crucial as we see a myriad of applications of reinforcement learning. Further, these results provide initial implications for the coordination of AI systems, which are trained to optimize their own performance, yet must work successfully with other individuals. As such technologies continue to develop and connect with more aspects of our world, understanding the interfaces between models of different capacities will become increasingly important.

Section 4.6

Appendix

By assuming mutations are sufficiently rare, Adaptive Dynamics simplifies evolution to competition between two types: the resident, and a mutant which either fixes to

become the new resident or becomes extinct before the next mutant emerges. Invasion fitness is the difference in payoffs $E(m, r) - E(r, r)$ between the mutant trait m and resident trait r when the mutant is initially rare, and is often used as a proxy for the fixation probability. Indeed, if this is negative, the mutant will experience negative selection and likely die out. Typically adaptive dynamics also assumes small levels of mutation to use the gradient to determine dynamics, but the lack of a closed form for our dynamics makes this infeasible, freeing up consideration to non-local mutation, as in the above model. Here we assume interaction continues for long enough for the dynamics to reach equilibrium, and remains there long enough that the equilibrium payoff approximates the average payoff accumulated throughout the whole interaction, the quantity determining fitness in the first model. Since mutation is rare, we assume the group of N consists of one individual having temperature m while the others have temperature r, and find the equilibrium numerically by solving the system over a sufficiently long time range, estimated on a case by case basis. Specifically, we use the initial condition $x(0) = 1/2$ since we assume the group forms without any prior information, so each agent initially follows a uniformly random strategy. This approach separates the timescales between learning and evolution, performing selection based on the equilibrium reached.

To derive our system of differential equations, we follow Kianercy and Galstyan[120] and consider stateless Q-learning with no discounting ($\gamma = 0$), simplifying the Q-value update equation to

$$Q_i(t + \delta t) = Q_i(t) + \alpha[r_i(t) - Q_i(t)] \tag{4.4}$$

where δt is the time step size, $r_i(t)$ is the average reward of choosing action i at time t, and α and $Q_i(t)$ are as before. Rearranging and taking the limit $\delta t \to$

0 gives $\dot{Q}_i(t) = \alpha(r_i(t) - Q_i(t))$. Since actions are chosen with the Boltzmann mechanism using temperature T, then the probability x_i of choosing action i is $x_i = \frac{\exp(Q_i(t))/T}{\sum_j \exp(Q_j(t))/T}$. Taking the time derivative we obtain $\dot{x}_i = \dot{Q}_i(t)\frac{\exp(Q_i(t))/T}{\sum_j \exp(Q_j(t))/T} - \frac{\exp(Q_i(t))/T}{(\sum_j \exp(Q_j(t))/T)^2}\sum_j \dot{Q}_j(t)\exp(Q_j(t))/T$ which simplifies to $\dot{Q}_i(t)x_i - x_i\sum_j \dot{Q}_j(t)x_j$. Equivalently, $\frac{\dot{x}_i}{x_i} = \dot{Q}_i - \sum_j \dot{Q}_j(t)x_j$. Substituting in our equation for the derivative of the Q-values we obtain $\frac{\dot{x}_i}{x_i} = \alpha(r_i(t) - Q_i(t)) - \sum_j x_j\alpha(r_j(t) - Q_j(t))$. Collecting terms and rescaling time by α for simplicity makes this $\frac{\dot{x}_i}{x_i} = \Big(r_i(t) - \sum_j x_j r_j(t)\Big) - \Big(Q_i(t) - \sum_j x_j Q_j(t)\Big)$. Since $T\ln\frac{x_i}{x_j} = T\ln\frac{\exp(Q_i(t)/T)}{\exp(Q_j(t)/T)} = Q_i(t) - Q_j(t)$, we can distribute the $Q_i(t)$ into the sum, as $\sum_j x_j = 1$, to rewrite the equation as

$$\frac{\dot{x}_i}{x_i} = \left[r_i - \sum_j x_j r_j\right] - T\sum_j x_j \ln\frac{x_i}{x_j} \tag{4.5}$$

This first term corresponds to increasing the probability of choosing an action that has an above average payoff, and the second corresponds to the energy in a statistical-mechanical system. This shows that some reinforcement learning methods can be viewed as an evolutionary process within an agent between actions. When there are just two actions, we can summarize an agent's strategy by a single number x, the probability of choosing the first action. This gives the now single equation

$$\frac{\dot{x}}{x} = [r_1 - (xr_1 + (1-x)r_2)] - T\left(x\ln\frac{x}{x} + (1-x)\ln\frac{x}{1-x}\right) \tag{4.6}$$

or equivalently

$$\dot{x} = x(1-x)\left(r_1 - r_2 - T\ln\frac{x}{1-x}\right) \tag{4.7}$$

In the game we study the actions are contribution and defection, and the difference

in expected payoffs is

$$r_1 - r_2 = \sum_{S \subseteq \{1,\ldots,N-1\}} [f(|S|+1) - 1 - f(|S|)] \prod_{i \in S} x_i \prod_{i \notin S} (1 - x_i) \tag{4.8}$$

the average difference in payoffs for contribution and defection over all subsets S of individuals contributing, weighted by the probability of that outcome. These dynamics alone can exhibit a large degree of complexity, Fig. 4.5 depicts the loci of equilibria over the space of possible reward function for just three players, and a given temperature. Finally, the full system we consider has separate equations for x_m and x_r, the strategies of those with the mutant and resident temperatures respectively, and the x_i in $r_1 - r_2$ are either x_m or x_r as appropriate.

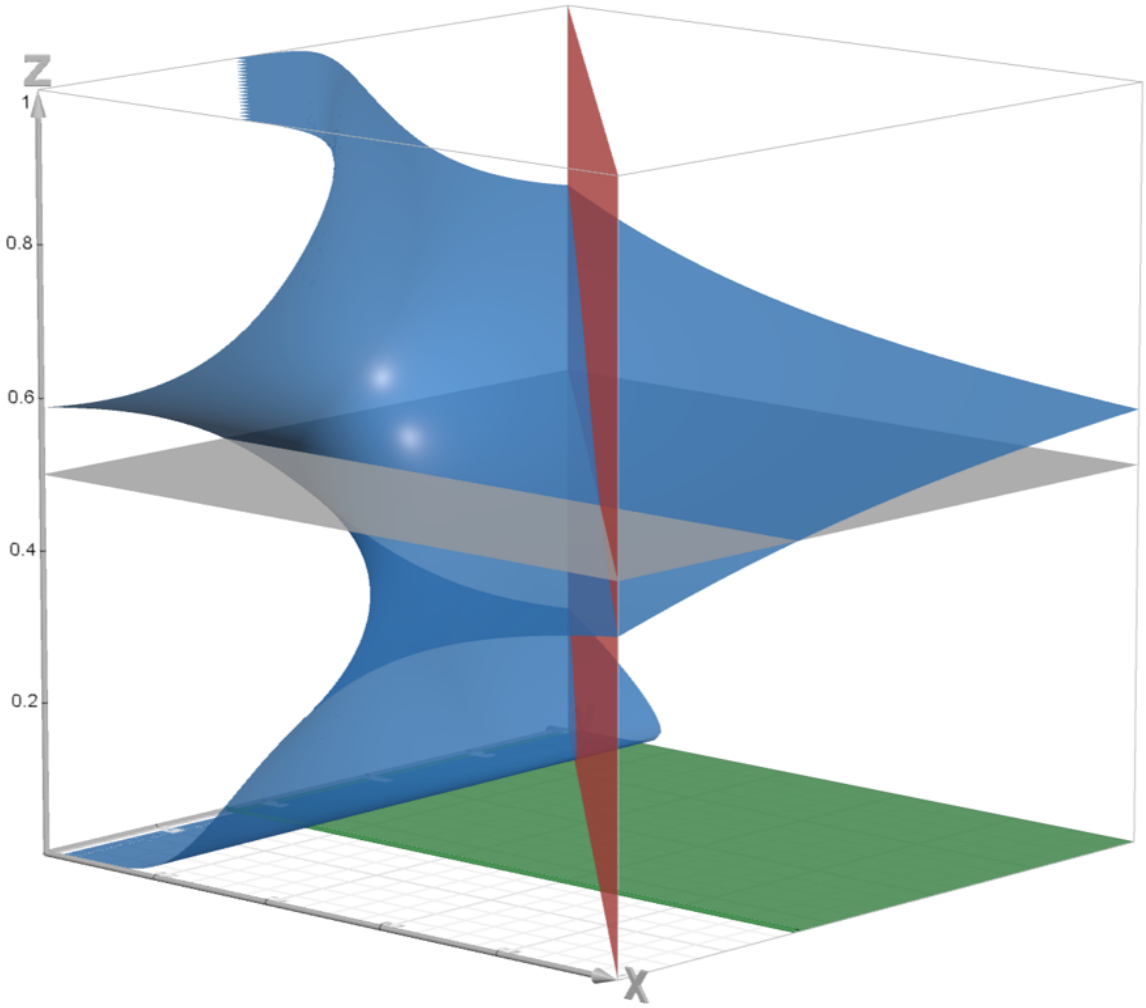


Figure 4.5: **Null-manifold of symmetric learning dynamics over the space of three player Public Goods Games**. This plots the equilibria of the learning dynamics where $j_0 = x$, $j_1 = y$, and z is the strategy, assuming $j_2 = m - j_0 - j_1$. Specifically, this plot uses $m = 3$ and $T = 0.1$. The red plane delineates the rejoin $j_0 + j_1 \leq m$, and the green region is the subset where the initial rate of change of the strategy is positive. In this case the maximum equilibrium contribution level, that is reached from an initial strategy of 0.5, occurs when y is on this boundary, and x is small. Note a large range of x, from zero to 0.1, have approximately the same level of contribution. Additionally, these values are close to having a negative initial change in the strategy, likely making them unstable for the actual dynamics, and possibly resulting in less frequent contribution.

Chapter 5

How norms shape the evolution of prosocial behavior

Compassion, Universalizability, Reciprocity, Equity: A C.U.R.E for social dilemmas

How cooperation evolves and persists widely remains an open problem for improving humanity across domains ranging from climate change to pandemic response. To shed light on how behavioral norms can resolve social dilemmas around cooperation, we present a formal mathematical model of individuals' decision making under general social norms, encompassing a variety of concerns and motivations an individual may have beyond simply maximizing their own payoff. Using the canonical Prisoner's Dilemma, we compare four norms: compassion, universalizability, reciprocity, and equity, to determine which, if any, social forces can facilitate the evolution of cooperation. We analyze our model through a variety of limiting cases, including

weak selection, low mutation, and large population sizes. This is complemented by computer simulations of population dynamics via a Fisher process, which confirm our theoretical results. We find that the first two norms lead to the emergence of cooperation in a wide range of games, but the latter two cannot. Due to our framework's generality, it can be used to investigate many other norms, and how norms themselves evolve. Our work complements recent work on fair-minded learning dynamics and provides a useful bottom-up perspective into understanding the impact of top-down social norms on collective cooperative intelligence.

Section 5.1

Introduction

Through millennia, evolution has produced incredibly sophisticated mechanisms by which organisms manage to survive and reproduce [12]. It is often difficult to explain how these arise through natural selection, as any sufficiently complex trait would likely result from a series of mutations that are likely neutral, or possibly deleterious on their own [190]. One prominent example of this is the intricate social systems seen throughout the natural world, from wolf packs to insect colonies. Vampire bats are known to cooperate in numerous ways, from sharing food to even their own blood [251]. Stickleback fish have also been observed cooperating to handle predators [163]. It remains a longstanding question how such complexity could form emerge though the simple process of mutation and selection, especially since cooperation often entails some degree of cost that may not be compensated for [178]. The canonical example of the tension between acting in a mutually beneficial way or the more compelling, selfish alternative is the Prisoner's Dilemma. Introduced in 1950 by Merrill Flood

and Melvin Dresher, but named by Albert Tucker, this game is a concrete version of this problem, assigning payoffs based on the which actions, selfish or cooperative, each individual chooses. Specifically, the below matrix gives the payoff received by a player given their action, in the row, and the other player's action, in the column. The possible actions are to cooperate or defect (C or D), yielding one of four possible payoffs $S < P < R < T$ [66]:

	C	D
C	R	S
D	T	P

(5.1)

Note these inequalities mean that regardless of the other player's choice, it is always optimal to defect, so individuals who are acting in self interest will choose to defect, leading to a worse outcome than mutual cooperation. Indeed, this is what rationality would suggest is the appropriate strategy. However, behavioral studies in humans show that people don't always follow the rational, albeit selfish strategy. Psychologists have proposed numerous explanations why people don't behavior rationally, including intelligence or personality, or that individuals adhere to some social norm [114, 27, 77, 257, 242]. In this work, we ask whether a mathematical model of four norms can promote cooperation. Specifically, we investigate compassion, universalizability, reciprocity, and equity. By prescribing a degree to which individuals follow or care about a given norm, we can model decision making that considers both individual payoff and broader social norms.

The four norms we model in this work have long histories in theories of morality. Compassion, also known as empathy, means caring about others rather than just oneself, and is one of the main factors theorized to influence cooperative behavior

[45, 19]. Universalizability is a concept introduced by the philosopher Immanuel Kant in the 18th century that claims an action's morality is determined by its effects when adopted by everyone. Immoral actions, in this framework, are those that are detrimental when they become universal [168, 195]. While this may seem like too sophisticated to apply to simpler organisms, it also makes evolutionary sense, as any successful trait should remain beneficial should it spread throughout a population. Further, it also describes the consequence of kin-interactions, as they likely share similar behavior. This norm is a bit stricter than compassion, as it additionally forbids actions like littering or free-riding, which do not harm any individual in particular, but only become truly problematic when everyone does them. Reciprocity is the exchange of beneficial actions [252], and has been found to be a key aspect in many experiments with social dilemmas, as well as an integral part of many cultures throughout the world [81, 50, 123]. Lastly, equity, or fairness, is a desire for unbiased treatment [43], and is widely valued across different communities [180]. Interestingly, even animals have been observed to care about fairness. Frans De Waals conducted a fascinating experiment which showed capuchin monkeys frustration with unequal rewards for the same task [31, 32]. This suggests that this factor may be a driving force in the evolution of cooperation, since it appears throughout the animal kingdom.

Apart from these considerations, evolutionary biologists have theorized several other mechanisms that can promote cooperation [178], including kin or group selection, direct or indirect reciprocity, and network effects among others [41, 232, 249, 102, 210]. Critically, these do not describe why pro-social behavior originates, but rather how it can spread through a population once it is present. Some studies have shown the emergence of cooperation can emerge without these factors, for example through

a combination of minimizing payoff differences between players and maximizing the sum of payoffs [158]. Other work has studied the effect of social norms, understood as a sequence of rules for updating status based on actions and the status of those interacting, finding criteria for reputation dynamics that maintain cooperation [182]. Researchers have also considered player's acting to maximize a linear combination of their payoff and that of their opponent, in a spatial game, considering a broad range of symmetric two player two action games [223, 222]. Similar studies have focused on lattice games; however all of these have the potential for spatial assortativity to select for cooperation [179, 225, 189, 204]. We add to this body of work by considering how adherence to social norms can evolve to promote cooperative behaviors in a well-mixed population, removing potential confounding effects, other factors known to allow cooperative behavior to spread through a population.

Section 5.2

Model and Methods

Individuals have a set value v that determines how important a social norm is to them, measuring either their adherence to this norm, or some notion of niceness described by the norm. Against a player with initial strategy y, an individual with value v chooses the best-response strategy x that maximizes their utility

$$x^* = \text{argmin}_x (1 - v)p(x, y) + vf(x, y) \tag{5.2}$$

where $p(x, y)$ is the payoff to an individual following strategy x with one following strategy y, and $f(x, y)$ is a function encoding the social norm, depicted in Fig. 5.1.

Each player's initial strategy x is fixed and visible to all other players, for example through publicly observing their actions over many previous interactions. However it is the realized strategy x^* that is used in the interaction to determine each players' payoff. In this way, decision making is simultaneous, and the strategy a player assumes the other player will follow is not affected by the randomness from the other player's previous interactions. We note that similar introspection dynamics has been studied in prior work [56], but our present model focuses on internal deliberations driven by norms. This framework is general enough to encompass many scenarios, including any game, encoded by $p(x, y)$. Depending on the form of $f(x, y)$, many different norms can be represented. Figure 5.1 gives a list of the four norms we consider here, though many more are possible. Our approach establishes a continuum between those who purely maximize their own individual payoff, $v = 0$, and those who solely follow the social norm, $v = 1$, allowing for a smooth transition between these extremes. In addition, the social norm $f(x, y)$ can be seen as a regularization term, which are used broadly in optimization when there is more than one goal, for example models trying to fit data with minimal complexity.

Note there are several different functions that could encode the same norm. For example, fairness could also be encoded by $f(x) = -|p(x, y)-p(y, x)|$ or the difference to an even power, or some sigmoid functions. One can also include parameters in these, for example a multiple of $x - y$ parameter in the reciprocity norm, which describes the importance of differences in strategy, the harshness of the norm, as the utility drops off faster as strategies become more distant.

The payoff we'll consider comes from the classic Prisoner's Dilemma game, which is the canonical example where cooperation is selected against. We avoid all the

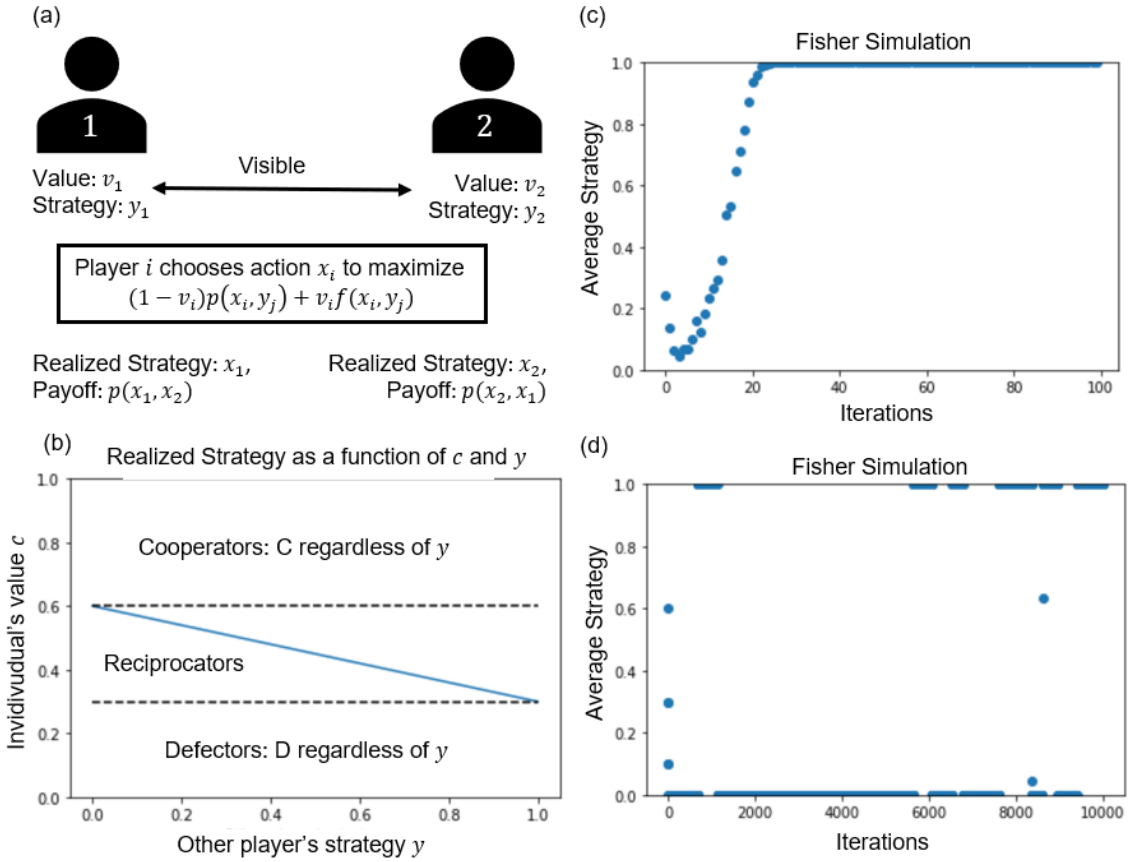


Figure 5.1: Social norms drive internal deliberation for behavioral responses. (a) we present a diagram of the model. Individuals have a fixed value v and initial strategy x. Then when interacting with an individual with initial strategy y, they choose strategy x^* to maximize their utility $(1-v)p(x^*,y)+vf(x^*,y)$, where $p(x,y)$ is the payoff from the interaction and $f(x,y)$ is some function encoding the norm. Each player follows this procedure to determine their realized strategies x^* and y^*. Specifically, we consider mixed strategies $xC+(1-x)D$ in the Prisoner's Dilemma. (b) we plot the corresponding realized strategies, the utility maximizing actions under the compassion norm as a function of y and value c, used instead of v to differentiate results for each norm. The thresholds c_0 and c_1 delineate values where players always defect or cooperate from those who match their opponent's strategy. Thus, there are four types of players, the cooperators, defectors, and reciprocators, with $y=0$ or $y=1$. In the two panels on the right, we see a Fisher simulation of this system under the compassion norm for a randomly chosen Prisoner's Dilemma, for illustration purposes. (c) shows the short-term dynamics, a population near the threshold transitioning from defection to cooperation. (d) illustrates the long-term behavior of the system under low mutation rates, showing fast transitions between cooperating and defecting states

norm	$f(x,y)$
compassion	$p(y,x)$
univeralizability	$p(x,x)$
reciprocity	$\exp(-(x-y)^2)$
equity (fairness)	$\exp(-(p(x,y)-p(y,x))^2)$

Table 5.1: Different norms can be encoded by various functions $f(x,y)$. Compassion means caring about how one's actions influence others, so individuals care somewhat about maximizing the other player's payoff $p(y,x)$. Universalizability considers how actions would perform if adopted by the whole population, where all receive payoff $p(x,x)$ since x is the common strategy. Reciprocity means acting towards others as they have acted towards you, cooperating in response to cooperation and defecting in response to defection, so your strategy should be similar to that of the other player. The given function captures this by increasing as the strategies x and y get closer. Similarly, the equity norm prefers strategies which result in a similar payoff for both players, so the same functions is used with the payoffs received instead.

features that are known to allow cooperation to emerge, to show this is new, for example spatial structure, to ensure these results are purely an effect of this model of norms. In later plots, we will normalize two parameters, $P = 0$ and $R = 1$, of the game so we can plot results in the ST-plane. Here, the strategy y is the mixed strategy $yC + (1-y)D$ which cooperates with probability y, and otherwise defects. The payoff is then the weighted sum $p(x,y) = Rxy + Sx(1-y) + T(1-x)y + P(1-x)(1-y)$ of possible outcomes, where the matrix in Eq.5.1 gives the payoff received by the player given their action, in the row, and the other player's action, in the column.

There are a few important things to note. First, individuals receive the payoff $p(x,y)$, not their utility $u(x,y)$. In this way any nonzero v will usually lead to a decrease in payoff, since only $v = 0$ guarantees the payoff maximizing strategy will be chosen. Indeed, increasing v can only decrease payoff, and likely will if $p(x,y)$ and $f(x,y)$ have different maximima. Therefore larger values should theoretically be selected against. Second, individuals cannot perceive the value v of another individual.

Consequently, this mechanism cannot be thought of as reputation or green-beard altruism, that is, individuals are somehow cooperating more with "nicer" players, those that also adhere to the norm.

One consideration in this algorithm is how strategies are updated. It may be unrealistic for strategies to change that quickly, so instead individuals could take a step of some fixed size towards the optimum, or some fixed interpolation between the two (making their strategy a geometric sum of previous optima). Another approach, that we use in some models, is to give individuals a preferred strategy that is fixed. Ultimately, strategies are visible, because in any of these cases, it could be learned by observing the individuals previous actions.

We analyze these models theoretically, and also perform a series of experiments. These are simulations of a Fisher process on n individuals. That is, each individual has fitness $e^{\beta p}$, where β is selection strength, and p is the payoff averaged over all possible interactions in the population. Then n are selected proportional to their fitness for the new population, possibly with some mutation on strategy or trait value. The code for these experiments and the corresponding figures is available at `https://github.com/bmDart/CURE`.

We begin by investigating the compassion norm, showing the continuous system has a discrete analog, which we can analyze in two simplifying limits. Then we discuss how this generalizes and the challenges in studying other norms. We will use different variables to clarify which norms each result holds for, for example c for compassion and u for universalizability.

Section 5.3

Results

5.3.1. Compassion

In the compassion norm, both $f(x, y) = p(y, x)$ and $p(x, y)$ are linear in x, so the utility function is also linear in x. Consequently the optimal actions will always be a pure strategy 0, pure defection, or 1, pure cooperation (or all actions have equal payoff, when this line has slope zero). A transition between these occurs when their payoffs are equal, when the compassion c_y satisfies $(1 - c_y)p(0, y) + c_y p(y, 0) = (1 - c_y)p(1, y) + c_y p(y, 1)$. Solving this yields

$$c_y = \frac{P - S}{T - S} + y\frac{(T + S - R - P)}{T - S} \quad \rightarrow \quad c_0 = \frac{P - S}{T - S}, \qquad c_1 = \frac{T - R}{T - S} \tag{5.3}$$

We can plot this to determine the utility maximizing responses, in figure 5.1.

We see there are three distinct regions of the compassion value, For low compassion values below c_1, no level of cooperation by the other player will get an individual to cooperate. Similarly, those with high compassion values above c_0 will cooperate regardless of the other players strategy. For intermediate levels, the utility maximizing strategy depends on the other player's probability to cooperate. However, as noted earlier, the optimal values are always either pure cooperation or pure defection. For these, all individuals with values in the range (c_1, c_0) will defect against a defector and cooperate with a cooperator. Because of this, we'll call them reciprocating strategies. One could also consider them similar to the Tit-for-Tat strategy in the iterated prisoner's dilemma which copies it's opponent's previous move [252, 207]. Thus this system with continuous values and strategies amounts to a discrete

system with four kind of players: those who always cooperate or defect, C or D, and those who reciprocate and prefer cooperation or defection, R_C or R_D. While both types cooperate with C and defect against D, the first type of reciprocator will cooperate with itself, whereas the second will not. This system is minimal enough to analyze theoretically, though we find simulation is necessary for norms exhibiting more complicated dynamics. We discuss this discrete model in the next section.

Intuitively, the reason why this model of norms may be able to promote cooperation is simple. One can imagine a population where everyone is defecting. Values below c_0 will be under neutral selection, since they will follow the same strategy, defection, as all other individuals. However, when an individual with a value above c_0 emerges, they will cooperate with everyone. Normally, this would be disadvantageous, since defectors will exploit this. However, the reciprocating players will serve as a buffer, cooperating with the cooperators to increase their payoffs and defecting against the defectors to decrease theirs. Depending on the portion of reciprocating players when a cooperator emerges, and the parameters of the game, a wide variety of dynamics can be seen. It turns out not to be sufficient for the cooperator to initially have higher average payoff than a defector, as it's possible the growing number of cooperators will make the population vulnerable to the remaining defectors. In the next two sections, we will analyze this system under two simplifying limits, low selection and large population size.

Before analyzing this further, we first note that the slope of the dividing line c_y is $\frac{(T+S-R-P)}{T-S}$, which is negative if and only if

$$T + S < R + P < 2R \tag{5.4}$$

However it's possible for this condition to fail even though the game is still a prisoner's dilemma, for example with $(S, P, R, T) = (0, 0.1, 0.2, 1)$. In such cases, individuals with intermediate values will cooperate with defectors and defect against cooperators. This counter-intuitive behavior results from the fact that c_0 and c_1 are essentially the differences $P - S$ and $T - R$, just normalized by $T - S$. These are the amounts gained by defecting against a defector or cooperator, respectively. Since compassion must outweigh the benefit of defection, these differences determine the necessary level of compassion for a player to cooperate. Therefore if it costs more to cooperate with a cooperator than a defector, players will need to be more compassionate to do so. We will not study this case, as such behavior cannot promote cooperation. Rather than acting as a buffer that promotes cooperating individuals, as in the previous case, individuals with intermediate values will promote defectors at a cost to themselves while hurting cooperators, leading to their extinction. Interestingly, this same condition delineates two types of behavior in the universalizability norm, which we discuss later.

5.3.2. Monomorphic Populations

In the limit of weak selection, Antal et al. derived a condition for traits to be favored based on the entries of a matrix of interaction payoffs under arbitrary mutation rates [9]. Specifically, they determine a necessary condition for a specific trait to be present than $1/n$, if there are n straits, in the mutation-selection equilibrium with arbitrary levels of mutation. We can think of the discrete model as one of these games with four types, with interaction payoffs given below.

	C	R_C	R_D	D
C	R	R	R	S
R_C	R	R	T	P
R_D	R	S	P	P
D	T	P	P	P

Applying their condition for low mutation rates, we see that C is favored when $S + 2R > T + 2P$. In the normalization $P = 0$, $R = 1$ this corresponds to the region $S > T - 2$, a triangle next to the origin. This makes sense, as it means the payoff to a cooperator who is defected against, S, must be large, and the payoff to a player who defects against a cooperator, T, must be small. The type D is favored in the complementary region. Next, R_C is favored when $T + 2R > S + 2P$, which is always true, and R_D is never favored. Lastly, R_C will always be more frequent than D in equilibrium. This all holds as well for the high mutation rate condition, though now the condition for C is weaker, $S+3R > T+3P$, corresponding to a larger set of R, P values. Since the conditions extend linearly to any intermediate value of mutation, this means R_C will always be favored the most, and C can be favored when D is not, depending on the payoffs and mutation level. This makes sense, as they perform well with both cooperators and defectors, and exploit R_D. This model is slightly inaccurate, as the paper assumes uniformly random mutation, which won't come from mutation on the players' values, for example D is more likely to mutate to R_D or R_C than C, as this former require less of an increase in their value.

Conversely, we can model in the limit of low mutation but with arbitrary selection [80]. In this case, the population will be mostly monomorphic, solely consisting of one of the types C, R_C, R_D, and D. When a mutant emerges, we can calculate

the fixation probability by the classic fixation formulae [235]

$$\frac{1}{1+\sum_{j=1}^{N-1}\prod_{i=1}^{j}\frac{g_i}{f_i}} \tag{5.5}$$

where f_i is the probability of transitioning from state i to state $i+1$, and g_i is the probability of transitioning from state i to $i-1$, where the state is the number of invaders. In particular, these are

$$f_i = \exp\left(\beta\frac{a(i-1)+b(N-i)}{N-1}\right)$$
$$g_i = \exp\left(\beta\frac{ci+d(N-i-1)}{N-1}\right)$$

where β is the strength of selection, and a, b are the payoffs of an invader interacting with an invader or resident, respectively, and c, d are those for a resident. Interestingly, using a linear fitness, interpolating 1 and the the term in the exponent, qualitatively similar results are obtained (the benefit of this approach is that fitness will always be positive). The cases $C \leftrightarrow D$ and $D \rightarrow C$ are classical, while $C \leftrightarrow R_C$ and $D \leftrightarrow R_D$ are neutral drift, as both types always choose the same action. We can consider $R_D \rightarrow C$ and $R_C \rightarrow D$ as neutral drift, as the invading reciprocator will switch to the resident strategy after the first interaction. The most complex is when a D invades R_C, or C invades R_D, as the other reciprocating types would be created by interactions. Nonetheless, we can approximate the fitness assuming as the expected value over interactions, for example, R_C will cooperate with themselves and defect against D, so their fitness would be a weighted average of these possibilities by the relative frequency of R_C and D. Also, R_C invading D is impossible, as emerging reciprocating players always follow the resident strategy, unless mutation occurs si-

multaneously on value and strategy. Another technical issues is that we cannot have $\beta = 1$, since the game parameters would make the expression have a division by zero term. On this timescale, the population is a Markov Chain over the monomorphic states with the transition matrix

$$\begin{bmatrix} P_{C\to C} & P_{C\to R_C} & P_{C\to R_D} & P_{C\to D} \\ P_{R_C\to C} & P_{R_C\to R_C} & P_{R_C\to R_D} & P_{R_C\to D} \\ P_{R_D\to C} & P_{R_D\to R_C} & P_{R_D\to R_D} & P_{R_D\to D} \\ P_{D\to C} & P_{D\to R_C} & P_{D\to R_D} & P_{D\to D} \end{bmatrix}$$

We can then look at it's principal eigenvector to see the proportion of time the chain spends in each state, to get a sense of which are favored. This will vary with selection strength β and the game parameters P and R, so to get a sensible plot, we'll take slices of the parameter space. This is done in figure 5.2, where first the selection strength β is varied for a fixed game, then the key game parameters S and T are varied for a fixed selection strength.

5.3.3. Large Population limit

In an infinitely large well-mixed population, the number of interacting pairs is proportional to the product of the frequencies of each individual in the pair. For example, of the C cooperators, C/N would interact with other cooperators, getting a payoff of R, R_C/N would interact with reciprocating cooperators, getting a payoff of R, and so on. This gives each type their average fitness. The new proportions will be (before

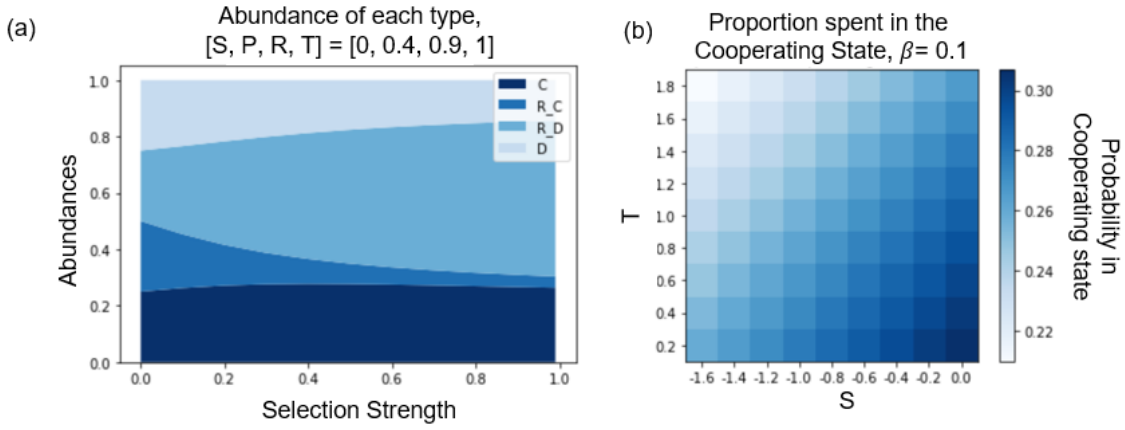


Figure 5.2: The optimal selection strength to promote unconditional cooperation. (a) plots the proportion of time the population spends in each monomorphic state for the particular game, [S,P,R,T]=[0,0.4,0.9,1], versus varying levels of selection strength. Surprisingly, there is a slightly non-monotonic relationship, at some point increasing selection causes cooperators to be favored less. Further, increasing selection seems to only result in lower quantities of cooperating players, $C + R_C$. (b) shows a heatplot of the proportion of time spent in the cooperating state with a fixed selection strength $\beta = 0.1$, over all possible games, where $P = 0$ and $R = 1$ so the space of games is two dimensional: a game is determined by the values of S and T. We see that this quantity is effectively determined by $S - T$. As expected, high values for S, the payoff of a cooperator when defected against, and low values of T, the payoff of a defector when defecting against a cooperator, yields the largest levels of cooperation.

normalization)

$$C' = RCC + RCR_C + RCR_D + SCD \tag{5.6}$$

$$R'_C = RR_CC + RR_DC + RR_CR_C + SR_DR_C \tag{5.7}$$

$$R'_D = PR_CD + PR_DD + PR_DR_D + TR_CR_D \tag{5.8}$$

$$D' = TDC + PDR_C + PDR_D + PDD \tag{5.9}$$

Each equation is a sum of terms, the coefficient of each is one of the possible payoffs $S, P, R,$ or T. Following this is a pair of variables, the first is the type of the focal player and the second is type of the other player. For example, C will get the payoff R when interacting with a C, R_C, or R_D player, giving the first three terms of the first equation, and a payoff of S when interacting with a D player, giving the last term of the first equation. The second equation indicates that new R_C players result from an R_C or R_D player interacting with a C, or two R_C's together, getting a payoff of R and becoming an R_C, while the last terms says that when an R_D player encounters a R_C player, they become an R_C and attempt to cooperate, but receive a payoff of S, as the other player saw an R_D and so defected. The next two equations can be interpreted similarly. This systems does not appear to have an analytic solution, and exhibits complicated behavior, see figure 5.3. This demonstrates that an initially higher payoff for cooperators need not guarantee they will invade successfully. Then, by comparing the final proportions of each type for various initial proportion of reciprocators, one can determine what amount of reciprocators is necessary to allow cooperators to invade a defecting population, or defectors to invade a cooperating population, in figure 5.3.

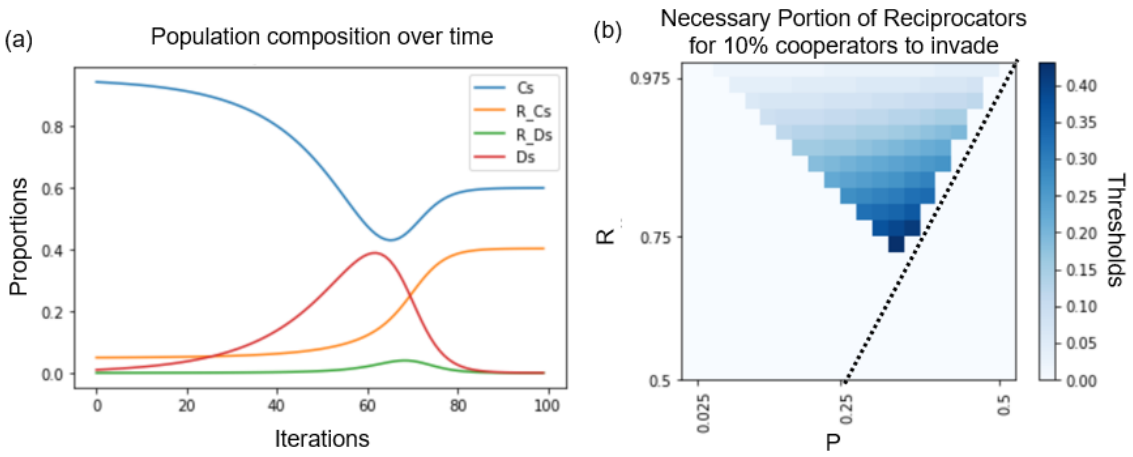


Figure 5.3: Deterministic dynamics under large population limit. (a) Here we see a model under an infinitely large population, to remove stochasticity. There are rich dynamics between the various type. In (a) we plot the proportions of each type over time for the game [S,P,R,T] = [0, 0.2, 0.9, 1], starting from a cooperating population with a small number of reciprocating players, invaded by a small proportion, five percent, of defectors. The cooperators initially have lower payoff, due to the small number of reciprocators, and decline. Defectors initially are taking over, but this creates a sufficient number of reciprocators to counteract this, around $t = 60$ iterations. The defectors then have lower payoff, allowing cooperators to recover, and the invading defectors to go extinct. Crucially, if there were too few reciprocators, the defectors would be able to fixate and replace the cooperators; this plot demonstrates the behaviour around this threshold. This threshold will vary for each game. (b) we plot this threshold for every game, normalized to have two parameters. As expected, higher R, reward for mutual cooperation, helps cooperators, so less reciprocators are necessary for them to invade a defecting population. Interestingly, there is less effect of P, the punishment payoff for mutual defection. Additionally, we see a barrier indicating some games where cooperators cannot invade, at least in this model. In addition, we see a line $R = 2P$ that separates cases where this level of cooperators can invade in this model. Thus, there are some games where no amount of reciprocating players is enough for cooperators to invade.

5.3.4. Other norms

The universalizability norm is similar. Now, the utility function is $(1-u)p(x,y) + up(x,x)$. Note that this is quadratic in x, so if it is concave up, the maximal values will be achieved at the endpoints $x = 0$ or $x = 1$. This means the same approach as in the compassion case may be used. In particular, the threshold level of universalizability u_y against a player with strategy y makes the utility equal at the endpoints, giving

$$u_y = \frac{P - S - y(R - S - T + P)}{R - S - y(R - S - T + P)}, \qquad u_0 = \frac{P - S}{R - S}, \qquad u_1 = \frac{R - T}{P - T} \tag{5.10}$$

This is part of a hyperbola, and creates the same three regions as in the compassion case. Thus, the same system of cooperators, reciprocators, and defectors may be used, the only difference being the locations of the thresholds as a function of the game parameters. Comparing these thresholds with the compassion case, we see that $u_1 < c_1$ and $u_0 > c_0$, which means that for a set game, lower values are required to cooperate with a cooperator, and higher values are needed to cooperate with a defector. Thus, individuals in this norm are quicker to help cooperators and punish defectors. Equivalently, the region of values corresponding to reciprocating players is larger, thus there is a stronger buffer effect, promoting cooperation further (at least in the region $R + P > S + T$). Indeed, simulations find that cooperation can invade a defecting population in a broader range of circumstances than in the compassion norm. This is consistent with the discussion in the introduction noting that universalizability was a stricter norm than compassion.

The concave down case is more complicated, as now intermediate strategies maximize player's utility. This occurs when the coefficient of x^2 in the utility function, $R - S - T + P$, is negative. Interestingly, this is the opposite of condition 5.4, which

delineated cases in the compassion norm. Intermediate levels of cooperation could result in a continuous transition towards cooperation. Now, the utility maximizing action as a function of u and y goes from 0 at $u = 0$ to the maximum of $p(x, x)$, which depends on the game's parameters. This is actually interesting in and of itself, as it turns out that cooperation need not be the optimal solution for a population in the prisoner's dilemma. The conclusion of the classical prisoner's Dilemma is that while defection is optimal for an individual player, it is better for the pair to mutually cooperate. It is surprising, then, when this fails to hold for the larger population. For some parameter choices, more can be gained by the exploitative defector-cooperator interactions than is lost by the mutual defection interactions (indeed, it is possible for any level of cooperation to be optimal for a population). Interestingly, the optimal value $-\frac{S+T-2P}{2(R-S-T+P)}$ of $p(x, x)$ is similar to the fixed point $\frac{P-S}{R-S-T+P}$ of the replicator dynamics of this game. Simulation shows cooperation can not emerge, even in the most favorable circumstances, but it can be maintained in some cases. While not as clearly impossible as in the compassion case, this region still suppresses the evolution of cooperation.

The reciprocity norm has utility $(1 - r)p(x, y) + re^{-(x-y)^2}$. Similar to the concave down case in universalizability, the utility maximizing strategy increases from zero at $r = 0$ to y at $u = 1$, with speeds depending on the game parameters, since $x = y$ maximizes $f(x, y)$. Thus, we study this case by simulation, finding cooperation would never invade, and indeed defection would always fixate in a cooperating population. Intuitively, high values of this norm promote taking a similar strategy as your opponent, which means defectors will be defected against and cooperated will be cooperated with. However, there is no mechanism to introduce cooperation into a

defecting population.

Lastly, the equity norm turns out to just be a sharper version of the reciprocity norm, for the given functions. Indeed, the norm can be simplified to $f(x,y) = e^{-(S-T)^2(x-y)^2}$, since $p(x,y)-p(y,x) = Sx(1-y)+T(1-x)y-(Sy(1-x)+T(1-y)x) = (S-T)[x(1-y)-y(1-x)] = (S-T)(x-y)$, canceling some terms. This is interesting in and of itself, as it shows a connection between seemingly different notions of morality. Because of this connection, equity should behave qualitatively similar to the reciprocity norm. Studying a more sensitive version of a previous norm gives initial results investigating parameterized norms and the effects of changing parameters. However, since reciprocity showed no cooperation, the same was found here.

A good way to compare these norms is by plotting the best responses in each, for a fixed game, see figure 5.4. As in 5.1, the utility maximizing action is plotted as a function of the other player's intended strategy, on the horizontal axis, and their value, on the vertical axis.

Section 5.4

Discussion and Conclusion

Much work has been done throughout philosophy, psychology, and sociology to understand how norms shape our behavior and evolve over time [81]. Our study complements this qualitative treatment with a mathematical formalism that extends previous work on the origin of cooperative behavior. We propose a model where individuals choose actions depending on the other player's strategy and a value describing their adherence to some cultural norm, to maximize a combination of their own payoff

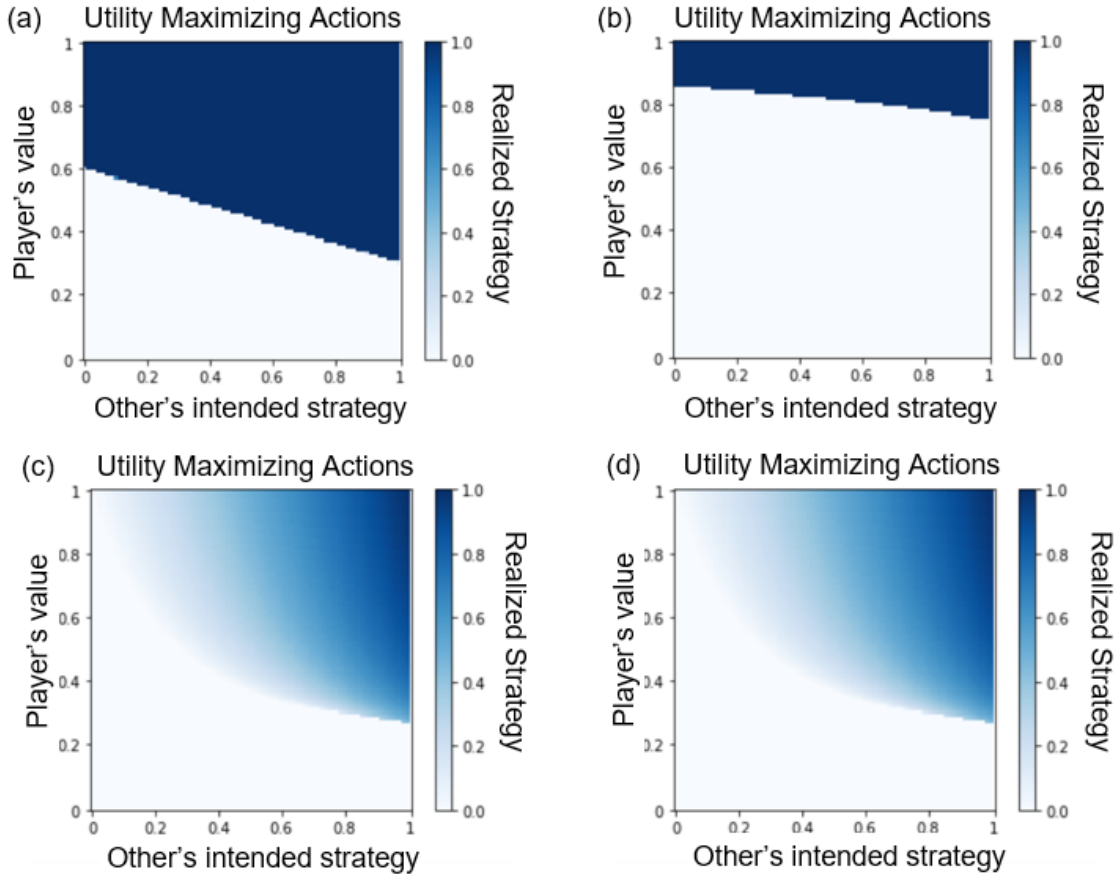


Figure 5.4: Strategy dynamics in all four norms. This plot is a comparison of the utility maximizing actions for the game [S,P,R,T] = [0, 0.6, 0.7, 1] under each norm: compassion, universalizability, reciprocity, and equity (respectively). In (a) and (b) we see the first two often have realized strategies that are pure, either zero or one, and an agreement with the earlier results. In particular, the line delineating the always cooperate, dark blue, and always defect, light blue, sections of (a) is the same line in figure 1(b). Plots (c) and (d) show that the last two norms result in mixed strategies. Here they are the same, because our choice of $f(x, y)$ makes one essentially a multiple by $S - T$ of the other, which is one for this game, so they are the same.

and a expression encoding the norm. This captures the fact that decisions are often made by considering more than just the literal payoff of one's actions. We then investigate this model from numerous angles. Using simplifying limits, we are able to obtain analytical results in a discrete description to our model in finite populations. Alternatively, we also consider infinitely large populations, where the model becomes deterministic. This is analyzed numerically to determine the necessary number proportion of our reciprocating players to allow for cooperation to spread in a population.

This work introduces a framework that encompasses decision making under a wide variety of social norms, by encoding these in a general function. We investigate whether this mechanism can promote cooperation without the presence of other factors known to do this, such as spatial structure or other forms of assortment. Further, this approach also allows us to analyze different norms from the same perspective, to compare then and even consider their evolution. As such, future work can investigate profound questions like which norms can replace others, how do norms change over time, and is there a best norm for promoting cooperation? Our framework also allows for a continuous emergence of cooperation, as compared to earlier work often considers discrete traits.

In the compassion norm, we see cooperation selected for in a large number of versions of the Prisoner's Dilemma, given by different game parameters. This shows how our proposed mechanism can resolve the dilemma in many varied cases [134, 207, 211], though does not guarantee the emergence of pro-social behavior. The essential reason we see this is that the reciprocating players serve as a buffer in the population. They simultaneously defect against defectors, and cooperate with

cooperators, suppressing the former while boosting the latter. A stronger effect is seen in the universalizability norm, due to its similarity. However, the reciprocity and equity norms are unable to promote pro-social behavior, because they provide no incentive to be more cooperative than the other player. However since the interactions are one-shot, traditional notions of reciprocity do not directly apply. Further work in this direction could address this by studying iterated games, where direct reciprocity can be explicitly modeled. Comparing this diverse set of norms, we establish the possibility, but not guarantee, for norms to allow for the emergence of pro-social behavior.

Our work presents a mathematical model of decision making under general social norms, where individuals compare the payoff they receive to how closely they achieve the orders of their norm. Quantifying this by a parameter allows for a continuum between purely selfish and purely selfless individuals, in contrast to previous models which often assume distinct behaviors between these two groups. Using this framework, we investigate whether cooperation can emerge and spread through a population depending on the social norm and game being played.

First studying the compassion norm, we note this continuous model reduces to a discrete system. This can be analyzed in the simplifying limit of weak selection, where we find a condition for the cooperators to be favored. Alternatively, under the limit of low mutation and arbitrary selection, the population is monomorphic most of the time, so we can calculate the transition probabilities between states to determine the proportion of time the population stays in each state. We find that this is essentially determined by a simple combination of the game parameters. Then we consider a large population limit, where all interactions become averaged to remove

stochastic effects. Doing so, we are able to compute the levels of reciprocating players necessary for a small number of cooperators to fixate in a population. This increases as the game becomes less favorable to cooperators, until eventually no amount of reciprocating players are sufficient. These assumptions could be relaxed in subsequent work examining these effects in finite populations with intermediate levels of selection.

We then consider how this analysis may be extended to other norms. Universalizability is a similar norm, in that a discrete systems emerges in this case as well, though the thresholds for this categorization are further apart than in the compassion norm. Consequently, the ability to promote cooperation in these cases is even greater. However, an alternative regime occurs where the optimal response varies continuously with adherence to the norm, preventing a discrete model of the system. Simulation in these cases found cooperation could not emerge. In the reciprocity norm, the actions chosen vary from defection to matching the other players strategy, as the player's value increases from zero to one. Simulation shows no cooperation is able to emerge, as there is only an incentive to cooperate as much as the other player, but no further. As a result, the players who cooperate less on average do better overall, and the level of cooperation falls over time. The same holds in the equity norm, as some straightforward algebra shows it is essentially a stronger version of the reciprocity norm, at least using the form we study. Thus, the simulation results are similar. By studying a range of different social norms grounded in the psychological tradition, we demonstrate a widely varying ability to explain the emergence of cooperation depending on the norm and game under consideration.

Future work in this framework could investigate a number of interesting questions. Thanks to our model's generality, it can be applied to practically any game, including

the Iterated Prisoner's Dilemma [47, 217] where far more complicated strategies can be used. In addition, our work could easily be extended to the ultimatum [185, 254] and public goods games [128], two other widely studied examples used to understand strategic behavior in different contexts, for example, via multilayer interactions [263, 218]. Initial simulation results suggest that the compassion norm can allow agents to find the optimal coordinate equilibrium, a result which may extend to other coordination games. Additionally, there are many functions that could encode a given social norms, and many other high-order social norms that could be considered [134, 202, 119]. The precise form of our results depend on the particulars of this function, but perhaps general insights can be gained by thinking of what forms these functions might take, such as which features are necessary to promote cooperation. Perhaps one could even evolve the norm of a function itself to see which are optimal for certain games. By putting different norms, compassion, universalizability, reciprocity, and equity, in the same framework, one can see which are better able to promote cooperation. One could even combine norms by incorporating additional values for each player governing how much they follow each norm. That way, even if one norm may be unable to promote cooperation in isolation, it may be able to in conjunction with another.

The evolution of pro-social behavior has been a longstanding question in biology with numerous explanations proposed through the years [81, 201]. By considering a well mixed population with minimal other factors, we have demonstrated the bottom-up emergence of cooperation under the influence of top-down social norms. In light of growing concerns regarding potent AI systems and their impact on humanity [59], our work paves the initial way for leveraging built-in behavioral norms to moderate

the cooperation of artificial agents [160] in hybrid AI-human systems [17, 137, 234, 42].

Chapter 6

Repulsion can create symmetry in opinion dynamics, how a single trivial issue can transform consensus into polarization and effectively prevent convergence

Opinion dynamics uses techniques from mathematics and physics to study the evolution of opinions through a population, focusing on topics like the spread of misinformation or polarization. Our work expands classical models to account for repulsion between dissimilar individuals and multiple issues of varying weights. Counterintuitively, we find special cases where adding a single issue of arbitrarily small importance

can disrupt stable opinion configurations, effectively preventing convergence, as well as strengthening consensus or mitigating polarization. To understand these effects we study a corresponding system of ordinary differential equations. The equilibria of this model exhibit unexpected symmetries that clearly explain these observations. By classifying weights intro finitely many types, we provide a complete characterization of how this effect occurs. Our work has important implications for the application of opinion dynamics models to real world cultural dynamics, and suggests promising directions for this interdisciplinary field.

This work was initiated by Daniel Simonson, and supervised by Dominik Wodarz, Feng Fu, and Natalia L. Komarova.

Section 6.1

Introduction

Opinion dynamics is a field at the intersection of mathematics, physics, and social science which seeks to understand how local social interactions create global effects on the opinions in a population. Typically, these are the degree and form of fragmentation the opinions undergo as function of the model parameters. Of particular interest are consensus states, where all individuals share a single opinion, and polarized states, where the population splits into two groups, each with a single opinion [88, 213, 95, 147]. To investigate these questions, a variety of models have been proposed. A large class of models applies ideas from statistical physics[38]. For example, voter models use discrete opinions where individuals adopt those of their neighbors at random [52, 105]. Another broad class of models, originating with the DeGroot model, assigns real numbers to describe the influence among each pair of individuals [61]. The vector

of opinions is updated through multiplication by this influence matrix, so the eigenvectors of this matrix are significant. The commonly used Hegselmann–Krause model extends this by allowing influence to vary as opinions change. Here agents give equal weight to all opinions within some distance of their own, given by some metric on the space of opinions and threshold parameter[194]. This consideration is known as bounded confidence and reflects the idea of the well known confirmation bias from psychology where individuals only consider information that already aligns well with their beliefs. One main finding in this model is that the fewer opinion clusters tend to result from larger confidence bounds [38]. Further richness has been found by examining dynamics of multiple issues simultaneously, beginning with Axelrod's model of cultural dissemination [11]. In this model, individuals exist on a spatial lattice and interact uniformly at random with their neighbors, adopting the neighbors' opinion on an issue proportional to the similarity between the opinions of the two across all issues. Through simulations, the author finds that fragmentation decreases as more issues are considered or interactions occur over a greater range, and increases with more possible opinions per issue. Several related multidimensional models have been explored since [216, 74, 20, 91, 188]. Historically models have focused on sympathetic interactions, where similar individuals update their opinions to move closer together. Recent works have also considered models that include repulsion between individuals with dissimilar opinions, investigating the phase transition between consensus and polarization[199, 54, 193, 39, 86]. One study found that these antagonistic interactions increased polarization while sympathetic interactions fostered consensus [108]. Another study observed traveling polarized states in a circular opinion space [93].

Our model investigates multidimensional opinion dynamics with attraction and

repulsion using heterogeneous issue weights, allowing us to study the often overlooked effect that issues have varying importance. Since previous studies of opinions dynamics on networks have found little effect of network topology we use a well-mixed model to make the mathematical analysis more tractable. Following the main considerations in the field, we focus on the time it takes dynamics to reach a absorbing state, polarization or consensus, and which of these states the population reaches. We find nontrivial effects in our extension, which we readily explain through symmetries in our model. One surprising finding of this work is that introducing an arbitrarily trivial issue into the opinion dynamics can force the population dynamics into consensus or out of polarization, or significantly delay the convergence time.

Section 6.2

Model

In our stochastic model, we consider a finite population of N individuals, each having J independent binary opinions, where issue i has weight w_i. At each iteration, two individuals are chosen uniformly at random to interact, discussing an issue also chosen uniformly at random. The focal individual updates their opinion on this issue based on that of the other individual, and the degree of similarity, defined as

$$\sigma(s,t) = \sum_{i=1}^{J} w_i \delta(s_i, t_i) \tag{6.1}$$

where $\delta(x,y)$ is one if $x = y$ and zero otherwise. That is, the similarity is the sum of the weights of issues where the pair agrees. If this similarity is below a lower threshold α_e, then the opinions are sufficiently different that the individuals are enemies, so

the focal individual adopts the opposite opinion of the other on the issue discussed. If instead the similarity is above an upper threshold α_f, then the individuals are similar enough to be friends, and the focal individual updates their opinion on the issue discussed to match that of the other individual. These effects can be seen as repulsion and attraction, respectively, with the parameters α_f and α_e governing the balance between these forces. The update process is repeated, until no more opinion change is possible. This occurs when all individuals have the same opinion, or the only opinions present are diametrically opposed, because the only interactions will occur between friends who have no differing opinions to change, or enemies with no opinion in common to move apart on. Note however if $\alpha_f < 0$, then even individuals with the diametrically opposite opinion will be friends, allowing for further change. Likewise if $\alpha_e > \sum_{i=1}^{J} w_i$ then even diametrically opposite opinions will be friends, allowing for further change. In either case, the condition for this model to have converged is that only a single opinion is present.

To better understand the dynamics of the stochastic model, we take a large population limit to obtain a deterministic model of a system of ordinary differential equations. Letting x_s represent the proportion of the population with opinion s and averaging over all possible interactions, the dynamics are given by

$$\frac{d}{dt}x_s = -x_s\left[\sum_t L_{s,t}x_t\right] + \sum_{i=1}^{J} x_{f_i(s)}\left[\sum_t G^i_{s,t}x_t\right] \tag{6.2}$$

$$L_{s,t} = \begin{cases} 1 - \frac{d(s,t)}{J} & \sigma(s,t) < \alpha_e \\ 0 & \alpha_e < \sigma(s,t) < \alpha_f \\ \frac{d(s,t)}{J} & \alpha_f < \sigma(s,t) \end{cases} \qquad G^i_{s,t} = \begin{cases} \frac{1}{J} & \sigma(f_i(s),t) < \alpha_e \text{ and } t_i \neq s_i \\ \frac{1}{J} & \sigma(f_i(s),t) > \alpha_f \text{ and } t_i = s_i \\ 0 & \text{otherwise} \end{cases}$$

where $f_i(s)$ is the opinion with entries equal to those of s, but has the opposite opinion on issue i, and $d(s,t)$ is the Hamming distance between two opinions, the number of issues on which they differ. This equation says that decreases in opinion s come from s discussing one of the $d(s,t)$ of J issues that they disagree on with a friend of opinion t, or one of the $J - d(s,t)$ issues they agree on with an enemy t. Increases in the proportion of type s come from one of its neighboring types $f_i(s)$ flipping it's opinion on issue i to agree with s (since only one opinion changes at each interaction). This can come from switching to disagree on issue i with an enemy t who disagrees with s on issue i, or from switching to agree on issue i with a friend t who agrees with s on issue i. We seek to understand these large systems of 2^J variables with respect to $J+2$ parameters, the J weights w_i and the similarity thresholds α_e and α_f.

While these models have continuous similarity thresholds, their behavior depends only on the relationships between opinions. There are only finitely many values for the similarity of two opinions, as these are sums of one of the 2^J subsets of the weights. Any two thresholds in an interval between adjacent subset sums will produce the same model, as every possible similarity will then either be above or below both of these thresholds and therefore those opinions will have the same relationship regardless of which threshold is used. This means there are really only discretely many cases for the threshold to consider given a list of weights.

If two weight lists have these subset sums in different orders, then they can produce different relationships between the opinions and therefore different models. Thus it suffices to determine all possible orderings of these subset sums. Despite the unaccountably many lists of J positive real numbers, there are only $2^J!$ many possible orders of the subset sums. Determining all possible orders of subset sums in a set

of J real numbers is a combinatorial problem. In some cases the subset sums might overlap, removing some possible intervals. We ignore these degenerate cases as each has a non-degenerate weight list with all the same possible opinion relationships and more, given by increasing one weight in either of the two overlapping subset sums by a number less than the minimum distance over all adjacent subset sums (so no others switch positions). This non-degeneracy means we are completing a partial order, where some elements of a set are incomparable, into a total order, where all elements can be listed in order. In general performing this enumeration is #P-complete, however in our special case of subsets of the weight list ordered by inclusion, more specific results are known, [151, 30]. We can use topological to perform this linear extension, by creating the total ordering one entry at a time, iteratively choosing a minimal element in the partial order and updating the order accordingly [221]. The conditions on the weights for each ordering are given in Fig. 6.2.

For the majority of this work, we consider $\alpha_f = \alpha_e$ for simplicity, omitting the subscripts when this is the case. We also briefly consider an extension to both of the above models where in addition to the interactions between individuals there is a small probability of the focal individual randomly changing their opinion on the selected issue, reflecting the possibility of rare innovations in opinions. This change complicates the analysis as the resulting model has no absorbing states but rather a stable distribution over all population configurations.

Section 6.3

Results

Simulations of our stochastic model show a non-uniform dependence of absorption time on the similarity threshold α, Fig. 6.1. This indicates that small changes in alpha can have large effects on the model. In contrast, the probability of consensus shows a clear relationship with α, Fig. 6.1. Larger alpha correspond to more antagonistic interactions, resulting in a polarized state. The strength of this relationship increases with population size, suggesting it will eventually reach a step function as N goes to infinity, indicating a phase transition. Intriguingly, there is a good degree of alignment between curves for different weight cases, suggesting a strong underlying pattern despite the many possible cases of heterogeneous weights. To fairly compare these cases, intervals are used instead of exact values for α. We use a relatively small population size of $N = 75$ since the absorption time is an exponential function of N, so larger populations are computationally infeasible to work with.

To understand these effect, we consider the trajectories in the deterministic system of differential equations, as shown in Fig.6.5 for various values of α_f and α_e. We see that often the proportions of various types become the same at equilibrium. That is, the equilibria exhibit various symmetries depending on the weights and similarity threshold. Figure 6.2 summarizes how the weights and similarity threshold jointly determine the symmetry in the deterministic model across the cases identified by the classification from the methods section. The symmetries are presented visually using a hyper-rectangle with solid lines connecting types with the same proportion at equilibrium, and parameterized regular expressions. For example, **zt describes

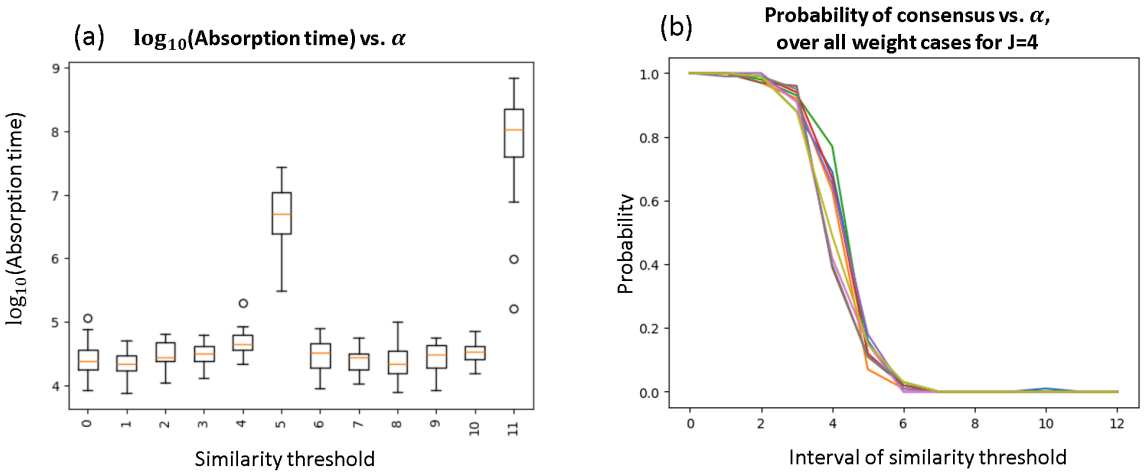
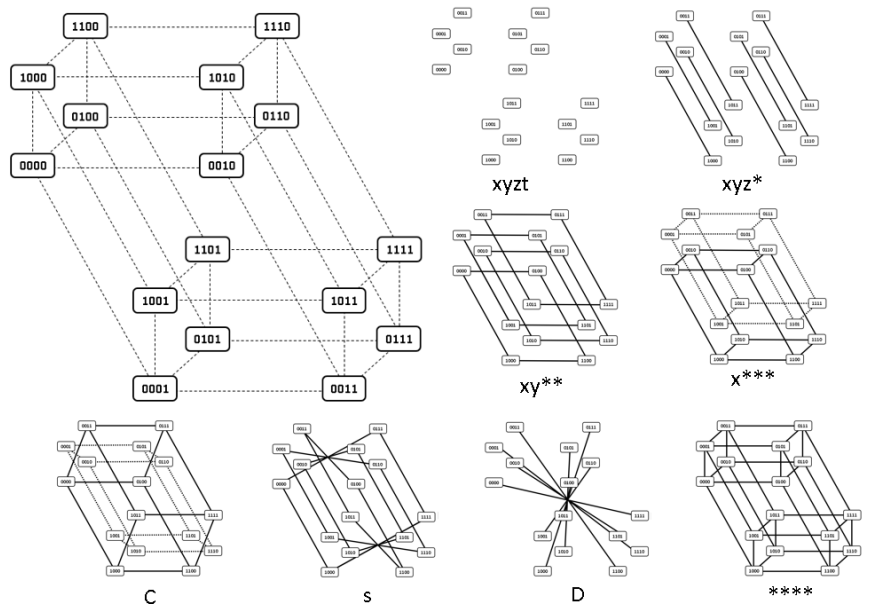


Figure 6.1: **Convergence time of the stochastic model varies significantly with α.** (a) Absorption times for 25 runs of a population with N=75 individuals starting from an initial condition selected uniformly at random, with $w = [0.18, 0.22, 0.26, 0.34]$, corresponding to the second to last row in Fig. 6.2. Note five values of the threshold are excluded, as the absorption time was significantly larger so would distort the plot. (b) The effect of alpha on the probability that the stochastic model is absorbed into a consensus state, as opposed to a polarized state, when initialized uniformly at random. Here N=50 with lines representing each case of weights.

a symmetry where proportions are the same if the opinions on the last two issues, given by z and t, are the same. That is, the opinions {0001, 0101, 1001, 1101} parameterized by $z = 0$ and $t = 1$ all have the same proportion, as do those by $z = 0$ and $t = 0$, which matches the set {0000, 0100, 1000, 1100}. We see a strong level of alignment between the symmetries present across cases, explaining the pattern seen across cases in the probability of consensus as a function of the similarity threshold. This occurs because for many similarity thresholds, related cases can have the same relationships among opinions, resulting in the same model. As an example, if α is less than the smallest weight, only those with the same opinion on all issues will be friends. Thus the order of the subsets sums, which determines the case, is irrelevant.

Using this classification, we identify an interesting case with intermediate α featuring a fully symmetric equilibrium, the uniform distribution, corresponding to $a < d+c$ for decreasing weights $[a, b, c, d]$. This allows the symmetry to change completely when issues are added, keeping α fixed, or rescaling according to the increase in the sum of the weights. This can slow convergence by several orders of magnitude, or increase the probability of consensus, Fig. 6.3. This same effect can be seen with uniform weights, for example going from $[1, 1]$ to $[1, 1, 1, 1]$ with $\alpha = 1.5$. While two issues have to be introduced with deterministic opinion updates, allowing a small probability of randomly changing an opinion on an issue creates enough noise for this effect to occur with just introducing a single opinion. Then the single additional weight can even be arbitrarily small, as the weights $[10002, 10001, 10000, 5]$ and $\alpha = 10003$ correspond to the same case. Similarly, adding the last weight to $[10008, 10004, 10000, 2]$ can increase the probability of consensus if $\alpha = 10001$, or decrease the probability of polarization if $\alpha = 10005$.



Conditions on weights $[a,b,c,d]$, $d<c<b<a$				Interval for Similarity threshold $\alpha_f = \alpha_e$														
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$d+c<b$	$d+b<a$	$a<b+c$	$a+d<b+c$	xyzt	xyzt	xyz*	xy**	xy**	xy**	xyz*	S	S	S	S	C	C	****	****	****	****
			$b+c<a+d$	xyzt	xyzt	xyz*	xy**	xy**	xy**	xyz*	S	D	S	S	C	C	****	****	****	****
		$b+c<a$	$a<d+b+c$	xyzt	xyzt	xyz*	xy**	xy**	xy**	xyz*	xyzt	D	D	S	C	C	****	****	****	****
			$d+b+c<a$	xyzt	xyzt	xyz*	xy**	xy**	xy**	xyz*	xyzt	xyzt	xyzt	S	C	C	****	****	****	****
	$a<d+b$	$a+d<b+c$		xyzt	xyzt	xyz*	xy**	xy**	xy**	C	S	S	S	C	C	C	****	****	****	****
		$b+c<a+d$		xyzt	xyzt	xyz*	xy**	xy**	xy**	C	S	S	S	C	C	C	****	****	****	****
$b<d+c$	$d+c<a$	$d+b<a$	$a<b+c$	$a+d<b+c$	xyzt	xyzt	xyz*	xy**	x***	xy**	xyz*	S	S	S	S	C	****	****	****	****	****
				$b+c<a+d$	xyzt	xyzt	xyz*	xy**	x***	xy**	xyz*	S	D	S	S	C	****	****	****	****	****
			$b+c<a$	$d<a+b+c$	xyzt	xyzt	xyz*	xy**	x***	xy**	xyz*	xyzt	D	D	S	C	****	****	****	****	****
				$a+b+c<d$	xyzt	xyzt	xyz*	xy**	x***	xy**	xy**	xyzt	xyzt	D	S	C	****	****	****	****	****
		$a<d+b$	$a+d<b+c$		xyzt	xyzt	xyz*	xy**	x***	xy**	xy**	S	S	S	C	C	****	****	****	****	****
			$b+c<a+d$		xyzt	xyzt	xyz*	xy**	x***	xy**	C	S	D	S	C	C	****	****	****	****	****
	$a<d+c$	$a+d<b+c$			xyzt	xyzt	xyz*	xy**	x***	****	C	S	S	S	C	****	****	****	****	****	****
		$b+c<a+d$			xyzt	xyzt	xyz*	xy**	x***	****	C	S	S	S	C	****	****	****	****	****	****

Figure 6.2: **Classification of symmetry type by similarity threshold interval, over all types of weight lists**. The diagram above represents the possible symmetries of the hyper-rectangle $0,1^J$, larger in the top left, seen among levels of opinions at equilibria of the deterministic model. Solid lines connect opinions with equal proportions at equilibrium, or dashed lines if there is significant overlap. Below this a table summarizing which symmetry occurs for each interval of $\alpha_e = \alpha_f$ over all of the nine cases of weights giving equivalent models for $J = 4$, indexed by conditions on the weights $w = [a, b, c, d]$ in decreasing order. The symmetries are described by parameterized regular expressions which match the subset of opinions with equal potions at equilibrium. The asterisk can be 0 or 1, and letters are parameters. The special characters D, C, and S indicate patterns not easily described by regular expressions.

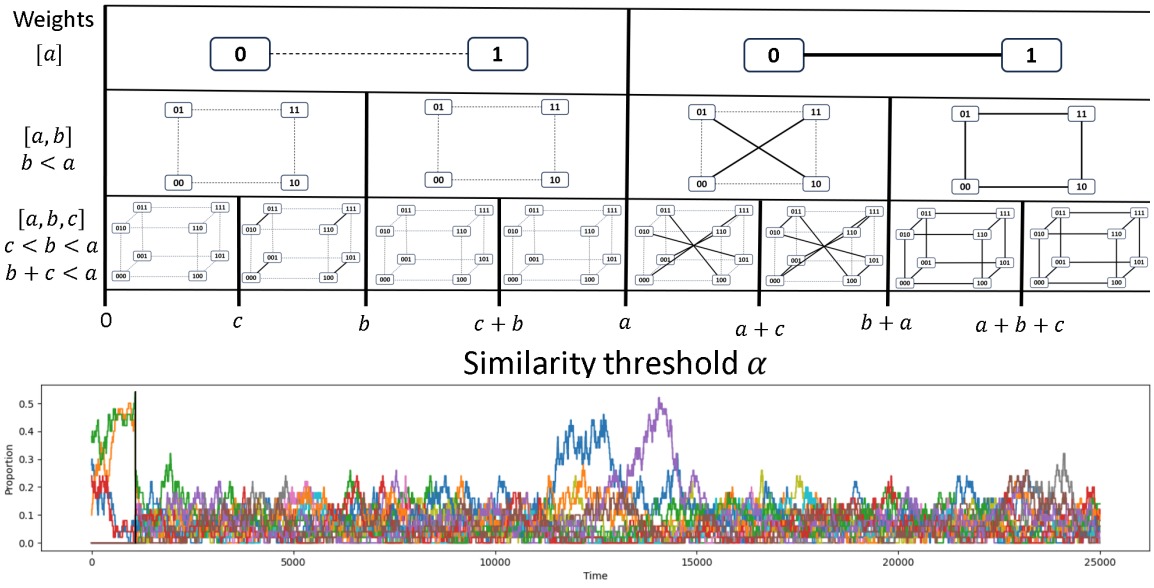


Figure 6.3: **Introducing issues of small weight can destabilize a polarized population, resulting in every opinion being held equally, and significantly delaying convergence to a new polarized, or consensus, state**. The above diagram shows how introducing weights changes the possible symmetries in the equilibria, where diagrams as in Fig. 6.2 are used for their corresponding regions of α. Using this to identify an interesting case, the below plot shows the dynamics in the stochastic model with $w = [9, 8]$, $\alpha = 9.5$, and $N = 50$. Once the population reach an absorbing state, indicated by a vertical black line around 1000 iterations, we introduce two new issues of smaller weights 6 and 4, splitting individuals with opinion s uniformly at random into the opinions $s00$, $s01$, $s10$, and $s11$. Now having the weights $[9, 8, 6, 4]$, the the system is attracted to the uniform equilibrium. The next 25,000 iterations are shown for clarity, though convergence took far longer. Repeating this 100 times, the ratio of the convergence times before and after issues are added averaged 133.36, though this varied largely, between 0.24 and 4820.

The observed symmetries explain the effects we found. When the population size is large enough, the deterministic model will be a good approximation of the stochastic model. This is evidenced in Fig. 6.3, where the equilibria satisfy $x_{00} = x_{11}$ and $x_{01} = x_{10}$ and indeed the corresponding red and blue curves are approximately equal. The agreement between the stochastic and deterministic models gets stronger as N increases, since this reduces the stochasticity in the model resulting from random interactions. The stochastic model only converges when no further opinion updates are possible, which occurs if there is only one opinion present, or two polar opposites, those which differ on each issue. This state may be reached through stochastic fluctuations while the state travels along a neutrally stable manifold of equilibria of the deterministic model, described by the symmetries. Some of these manifolds are closer to the convergence conditions, such as the D configuration where polar opposites have the same size. The relative proportions of each pair can then drift neutrally, until all but one pair becomes extinct and the population converges to a polarized state. Other symmetries, most notably the uniform distribution described by ****, is far from any absorbing state. The population must reach one of these by stochastic fluctuations to converge, despite being drawn by the dynamics back towards the uniform distribution. As such, convergence takes far longer in this case, especially for large populations. In particular, it is more likely to reach a polarized state, which requires all but two opinion types to become extinct, than the consensus state, which requires one further type to become extinct as well. This clarifies why many of the symmetry types lead to polarization instead of consensus.

The symmetries themselves arise from the relationships between opinions, which determine the model. We can summarize these by relationship networks, depicted in

Fig. 6.4. These lead to symmetries and anti-symmetries in the systems of equations. We can characterize symmetries as being fixed points of permutations. In general, when $f(\sigma(x)) = -f(x)$ for some function $f(x)$ and permutation σ, then $x = \sigma(x)$ implies

$$f(x) = f(\sigma(x)) = -f(x) \implies f(x) = 0$$

That is, all points x fixed by a permutation σ are roots of any function with σ as an anti-symmetry. Therefore if the equations for the dynamics of each opinion type follow the same anti-symmetry, then distributions that are symmetric with respect to that permutation will be equilibria. This was observed in some cases, but does not suffice to explain them all. Technically, this also only shows that such points are equilibria, not that they are the only equilibria. It is also sometimes the case that permuting the proportions x_s results in the equation for x_t. That is, $\sigma(F(x)) = F(\sigma(x))$ where $F(x)$ is a vector giving the system of differential equations, having entry $F(x)_s = \frac{d}{dt}x_s$, so equivalently $\sigma(\frac{d}{dt}x_s) = \frac{d}{dt}x_{\sigma(s)}$. Then

$$x = \sigma(x) \implies F(x) = F(\sigma(x)) = \sigma(F(x))$$

so if x is fixed by σ, so too are the dynamics. In particular, we seem to have that $\sigma(F(x)) = F(\sigma(x))$ for any permutation that transposes s and $f_i(s)$, where s has bit i flipped. By induction, compositions of these permutations also satisfy this equation. Indeed, such permutations preserve the similarity, and therefore the relationship, between opinions. These permutations also convert neighbors appropriately, as applying f_i to the set of neighbors of s gives the neighbors of $f_i(s)$. This gives a quick way to show that the uniform distribution is always an equilibrium, since it is fixed under all

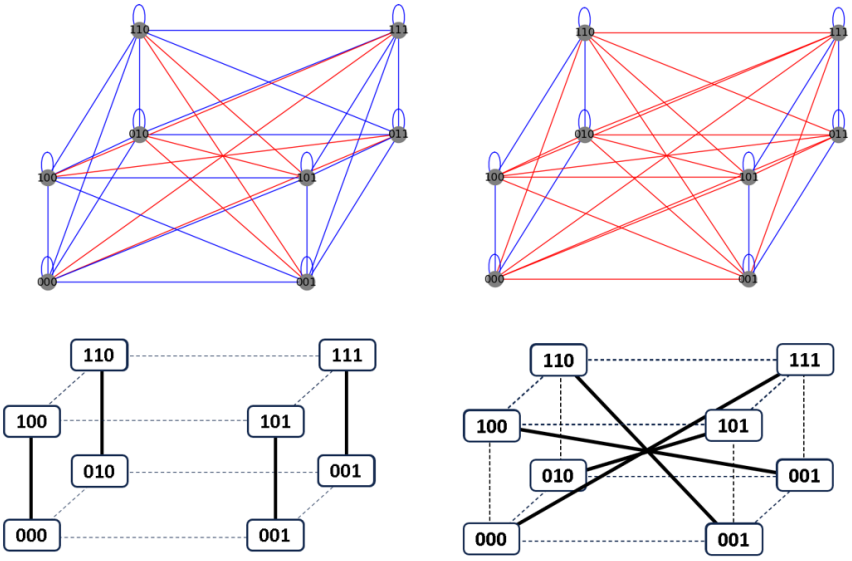


Figure 6.4: **Symmetries in the interaction network determine symmetries in the equilibria**. This shows the network of friend and enemy opinion pairs, indicated by blue and red edges respectively, using $w = [a, b, c]$ with $a + b < c$ and $a < \alpha < b$ on the left, and $c < \alpha < a + b$ on the right. Below each is the corresponding symmetry seen in the equilibria of the deterministic model in that case. For clarity only two cases with $J = 3$ are given. This explains why all the symmetric equilibria we found we are fixed by various permutations $f_i(x)$ that flip the opinion of x on issue i but leave all others the same, as well as compositions of these flips. For example, equilibria described by the symmetry *bcd are fixed by $f_1(x)$, similarly D corresponds to the product $f_1(x)f_2(x)f_2(x)$.

flip permutations, so the gradient will be too. Since all entries of the gradient are the same, each must be zero, since they sum to zero given the fixed population size. More broadly, it seems summing the equations for the rates of change over all types that are the same at equilibria of the deterministic model results in zero. Consequently, the total amount of the population contained in each equivalence class is constant. This suggests the system is somehow approaching the uniform distribution under this added constraint.

Finally, by considering the full model with $\alpha_f \neq \alpha_e$, we see that symmetries are

caused by antagonistic interactions, Fig. 6.5. Indeed, they only occur when there are enemies. While sympathetic relationships can influence the dynamics, they are not necessary for symmetries to occur. This is reflected in the fact that larger α tend to lead to a greater degree of symmetry, Fig. 6.2.

Section 6.4

Discussion

Using a multidimensional well-mixed model of opinion dynamics with heterogeneous issue weights and a combination of attraction and repulsion, we found a clear transition between consensus and polarization based on the similarity threshold delineating antagonistic and sympathetic interactions. Further, through a careful enumeration of all possible weight cases, we found instances where an intermediate value of the similarity threshold caused orders of magnitude increases in convergence times. This allowed us to find examples where introducing an issue of arbitrarily small weight would effectively prevent convergence or dramatically increase the probability of consensus, assuming a small amount of randomness is introduced. Our results show that small issues can have a disproportionate effect on these dynamics, as well as providing further evidence that sympathetic interactions lead to consensus while antagonistic interactions lead to polarization. These findings improve our understanding of opinion dynamics by highlighting the importance of antagonistic interactions and heterogeneous issue weights.

To analyze this model, we considered both a stochastic process with a finite population of agents, and a large population limit giving a system of ordinary differential equations. The stochastic model allowed us to identify meaningful phenomena, which

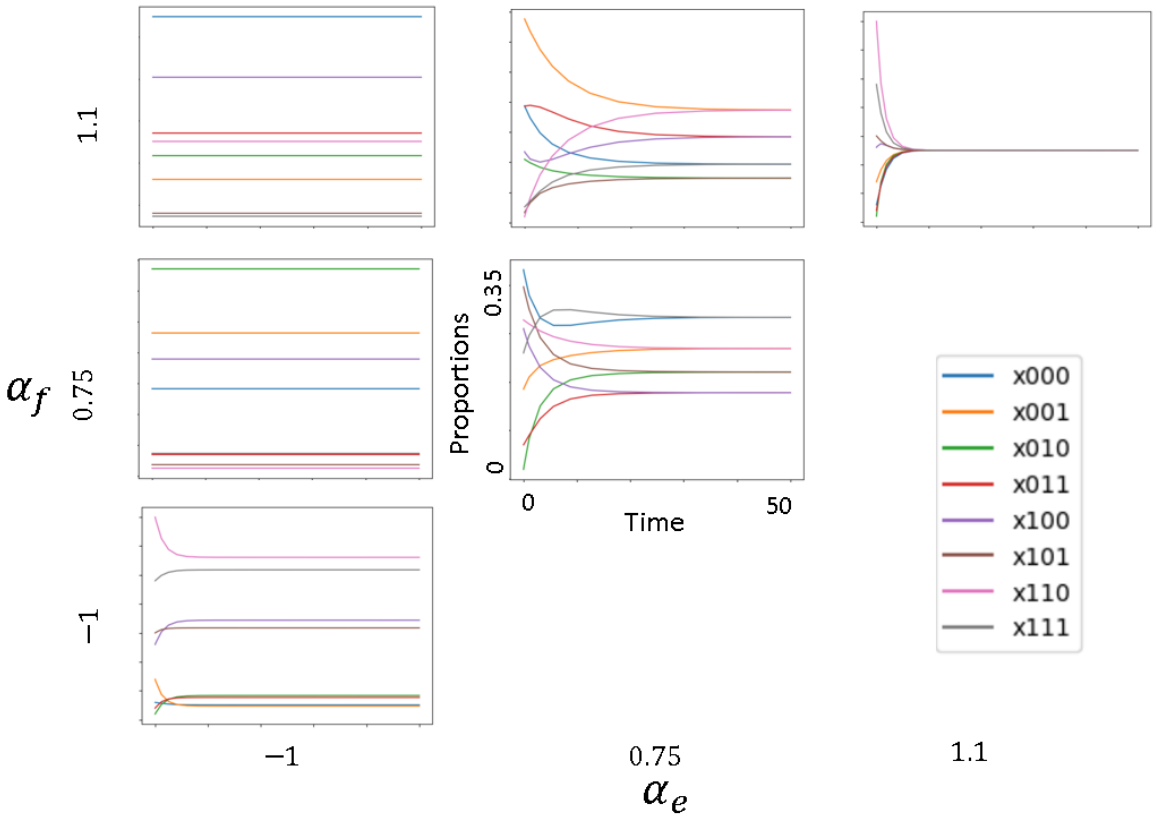


Figure 6.5: **Separating the friend and enemy similarity thresholds confirms that symmetry results from enemy interactions alone**. The preceding models focused on the case α, these parameters need not be the same in general, though $\alpha_e \leq \alpha_f$ is needed to prevent a pair from simultaneously being friends and enemies. Here we show the dynamics, of the deterministic model for clarity, using $w = [0.7, 0.2, 0.1]$ and values of α_e and α_f given along the horizontal and vertical axes respectively. The initial conditions for each are all chosen uniformly at random. Nonetheless, these are illustrative of the general case. The left column $\alpha_e < 0$ and top row $\alpha_f > \sum_{i=1}^{N} w_i$ correspond to there being no enemies or no friends in the model, respectively, since these are the lowest and highest possible similarity for this weight list. This shows that friendly interactions alone are not capable of creating symmetries in the equilibrium states, however this occurs frequently when just enemies are present.

we understood through the deterministic model. In particular, we found various kinds of symmetries that explain these effects, as they allow sufficiently large populations to neutrally drift closer to either polarization or consensus. By extending these to include rare innovations in opinions, we strengthened our results.

These findings fit into a broader body of work on opinion dynamics with multiple issues and repulsive forces. The main innovation of this work is considering heterogeneous weights, and investigating the effect of introducing new issues. By allowing a non-uniform importance of issues, we observed a range of related patterns, giving us a broader understanding of this space. By incorporating the possibility for a changing number of issues, we demonstrate that counterintuitive effects can occur where arbitrarily trivial issues can drastically change the dynamics.

Section 6.5

Conclusion

Many models consider only sympathetic interactions, and homogeneous weights. Our model relaxes these assumptions to investigate a more realistic system. We extend classical bounded confidence opinion dynamics models in a well-mixed population to include attraction and repulsion determined by the similarity of binary opinions on multiple issues with heterogeneous weights. Indeed, when antagonism and weight heterogeneity are introduced into models of opinion dynamics, surprising effects can arise in the convergence time and state. This work explicitly characterizes when, how, and why these effects occur.

We find that the absorption time of this model varies non-monotonically with the similarity threshold. By classifying all heterogeneous weight possibilities, we discover

cases where intermediate similarity thresholds lead to these orders of magnitude longer absorption times. As a result, adding issues of small combined weight, or a single issue if noise is present, can effectively prevent the convergence of a population. Likewise, depending on the parameters in our model, a single issue can substantially increase the odds of consensus, or decrease the odds of polarization. The similarity threshold has a clear relationship with whether the population reaches consensus or polarizes. More sympathetic interactions lead to consensus, while more antagonistic interactions lead to polarization. Both of these effects can be clearly explained by considering the symmetries in the equilibria of a limiting system of differential equations. These result from symmetries in the network of relationships, which cause the dynamics to exhibit symmetries and anti-symmetries. However the precise connection here is intricate and merits further study. Despite the uncountable many combinations of weights, a strong pattern emerges when they are compared. By separating the thresholds for friends and enemies, it becomes clear that the antagonistic interactions alone are driving these effects by creating symmetries. Further, the heterogeneity of the weights is critical to observing the large convergence time for an intermediate similarity threshold, a full range of the possible symmetries, and the possibility of an arbitrarily small weight completely changing the dynamics. This confirms that earlier models using only sympathetic interactions and equivalent weights miss significant effects.

This work can be easily extended in several meaningful directions. Issues could be considered proportionally to their weight, or even inversely to weight because often discussions are focused on trivial matters, and important issue could be too controversial. This would prevent the classification we used, as now the weights

would directly appear in the dynamics. As noted in the results section, further investigation can be made into precisely when and why the observed symmetries occur. Similarly, a more complete check of all the possible weight cases may be performed for larger dimensional cases. Another interesting possibility is allowing heterogeneous similarity thresholds to examine the effect of some more antagonistic individuals. This would allow investigation of whether a few who more antagonistic individuals or a few friendlier ones would have a bigger effect. Ultimately this would allow us to compare the strength of sympathy or antagonism, as well as evaluate their interaction. Such an extension would also connect well to prior work on heterogeneous agents in opinion dynamics[44, 127, 255, 108, 58]. We also only briefly explored the general model with arbitrary α_f and α_e. The same metrics of convergence time and probability of consensus could be considered, or others like the Shannon diversity index of the deterministic equilibria. We also used an asynchronous update in the stochastic model, but having all agents update their opinion simultaneously would allow the process to be put in the form $x_{t+1} = A(x_t)x_t$ where x_t is the vector of opinion frequencies and $A(x_t)$ is the matrix giving the probability of transitioning from one opinion to another. Like in the original Axelrod model, further work could consider to multiple alternatives for each issue, such as $\{-1, 0, 1\}$ for opposed, neutral, and in favor with similarity could be $\sigma(x, y) = \sum w_i x_i y_i$. Lastly small populations would have few enough states that we could explicitly find the transition matrix for our stochastic process, allowing for exact calculations of the absorption time and probability.

Beyond these, a number of more ambitious modifications could also be made, such as non-independent opinions on a more general graph of opinions[187, 175, 14, 93,

39, 187, 58, 130]. This would require a more sophisticated opinion update rule on the set of opinions V, namely a function $V \times V \to P(V)$ that returns the probability of the new opinion given the two opinions that interact. This would connect the preceding work to more general models of evolution on arbitrary mutation landscapes. An infinite dimensional model could also be considered, perhaps with weights geometrically distributed. Relating this infinite opinion string to its value in binary after the decimal point would take the discrete hyper-rectangle model to a continuous one dimensional model, the unit interval. As such, there might be a partial differential equation to describe these dynamics under a large population limit similar to mean-field models[124]. This almost recovers the classic Hegselmann-Krause model, however this mapping does not respect the distance function, for example the difference between 0.100 and 0.011 is seven eighths, despite the distance between these two numbers being an eighth. In this work we saw the significant effects caused by introducing a single issue. However it is almost universally assumed that the number of issues is a constant driving the dynamics. Thus, exploring a model where the number of issues, or their weights, can change would be significant. Like many works in opinion dynamics, another meaningful yet challenging advancement would be to connect these theoretical dynamics to observable opinion data, or learning the dynamics with a data driven approach[60, 24, 124, 86, 38]. Lastly, the assumption of a well-mixed population could be relaxed to study this model on a network to create partial observability of opinions. This network could be static[127, 187, 76, 1] or coevolving with opinions[117], creating a complex system. Dynamics could even be considered on generalized hypergraphs[108, 206] to include group interactions.

Chapter 7

Future Directions

One prominent mathematical tool that is not used in this work is partial differential equations. These can model the time evolution of a continuous distribution, for example the probability of cooperation, and as such can be seen as the limit of the preceding models when the possible traits are a continuum rather than discrete. This makes PDE's a natural extension of this work, and would broaden the potential applications. For example, the Fokker-Planck and Komolgorov equations are two PDE's that govern the evolution of probability distributions for continuous time Markov processes. They can be derived from corresponding stochastic differential equations, which combine deterministic forces with white noise. Another notable application of PDE's is mean-field games, where instead of a series of pairwise interactions, agents are assumed to interact with the population distribution as a whole, or simply its mean in some models. While this extension would offer a greater technical variety to this research, the PDE's encountered are often similarly intractable, so must be solved numerically through a variety of methods. Doing so would require a careful consideration of the convergence and error propagation properties of these methods.

The driving motivation for this work is obtaining a better understanding of the dynamics of cultural change and the origin of complex behaviors. This question is of great interest to economists, who often assume a high degree of rationality. This assumption is used to predict who individuals will collective react to various policies to determine which will be the most effective at achieving certain goals. This field is known as mechanism design, and has overlaps with algorithmic game theory, a subfield of computer science. This general goal makes it a type of engineering known as optimal control. Extensions of this thesis could consider how these models could be tuned for particular outcomes or steered under external forces from a central organization. This would connect them to a large body of literature on controlling complex systems.

While this work has explored the theoretical implications of a variety of models, there remains the question of whether these models are even accurate. As the 20th century statistician George Box said, "all models are wrong, some are useful," so it is critical to compare these to data if their predictions are to be taken seriously. This is a challenging statistical question, as the systems studied in this work are rather complex and feature a number of parameters that are not feasible to determine experimentally. In this way, it bears similarity to many ecological models, which can feature tens or hundreds of parameters, and are often used to obtain high-level insights into a system rather than concrete predictions.

Indeed, this is precisely the focus of both psychology and behavioral economics, to better inform such models. A number of biases have been empirically confirmed in human decision making, such as loss aversion, the decoy effect, and cognitive dissonance. These heuristics are patently irrational and yet have somehow were evolutionarily se-

lected for, or at least not sufficiently deleterious to be selected out of the population. Understanding the origins of these effects would provide valuable insight into when apparent irrationality is actually quite rational.

All of this work has assumed a well mixed population, making all possible interactions depend only on the frequency of each participant. This simplifying assumption allows for many formula to be derived, but it rather unrealistic. Many studies in this area instead assume a spatial or network based approach, where individuals are located at nodes of a graph and interact only with their neighbors. This approaches has tremendous flexibility, and can capture a large number of diverse effects, for example by considering real world networks like in social media, biological interactions, or research citations, or various synthetic models of networks like Erdos-Reyni random graphs or lattices. However, this introduces a substantial technical challenge to analyzing these models, as the state space is much larger. Often research in this area simulates these models to understand their dynamics, though a few approaches such as pair-approximation can be used to obtain analytic results. This technical challenge is further amplified when the network is hypergraph, including more complex group interactions than simple pairs, or co-evolves with the dynamics (indeed a static network is often itself an unrealistic assumption). Such co-evolution is exciting, as it creates a feedback loop between structure, which governing dynamics, and dynamics, which shapes structure. Simpler versions of this that are more tractable include patch models, where a small number of well mixed populations interact uniformly at random. This effectively creates multi-level selection both within and between the patches. Such models have been used in the ecology literature to study issues like dispersal and migration under various factors.

Bibliography

[1] Rediet Abebe et al. "Opinion dynamics with varying susceptibility to persuasion". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1089–1098.

[2] J Enrique Agudo and Colin Fyfe. "Reinforcement Learning for the N-Persons Iterated Prisoners' Dilemma". In: *2011 Seventh International Conference on Computational Intelligence and Security*. IEEE. 2011, pp. 472–476.

[3] Matthieu Alfaro and Rémi Carles. "Replicator-mutator equations with quadratic fitness". In: *Proceedings of the American Mathematical Society* 145 (2017), pp. 5315–5327.

[4] Matthieu Alfaro and Mario Veruete. "Evolutionary Branching via Replicator–Mutator Equations". In: *Journal of Dynamics and Differential Equations* 31 (2019), pp. 2029–2052.

[5] Matthiew Alfaro and Remi Carles. "Explicit Solutions for Replicator-Mutator Equations: Extinction Versus Acceleration". In: *SIAM Journal of Applied Mathematics* 74 (2014), pp. 1919–1934.

[6] Benjamin Allen and Daniel I. Scholes Rosenbloom. "Frequency-Dependent Selection Can Lead to Evolution of High Mutation Rates". In: *The American Naturalist* 183.5 (2014), pp. 131–153.

[7] Benjamin Allen and Daniel I. Scholes Rosenbloom. "Mutation Rate Evolution in Replicator Dynamics". In: *Bulletin of Mathematical Biolology* 74 (2012), pp. 2650–2675.

[8] Rangga Almahendra and Björn Ambos. "Exploration and exploitation: a 20-year review of evolution and reconceptualisation". In: *International Journal of Innovation Management* 19.01 (2015), p. 1550008.

[9] Tibor Antal et al. "Mutation-selection equilibrium in games with multiple strategies". In: *Journal of theoretical biology* 258.4 (2009), pp. 614–622.

[10] Jasmina Arifovic and John Ledyard. "Scaling up learning models in public good games". In: *Journal of Public Economic Theory* 6.2 (2004), pp. 203–238.

[11] Robert Axelrod. "The dissemination of culture: A model with local convergence and global polarization". In: *Journal of conflict resolution* 41.2 (1997), pp. 203–226.

[12] Robert Axelrod and William D Hamilton. "The evolution of cooperation". In: *science* 211.4489 (1981), pp. 1390–1396.

[13] Hui Bai, Ran Cheng, and Yaochu Jin. "Evolutionary reinforcement learning: A survey". In: *Intelligent Computing* 2 (2023), p. 0025.

[14] Pablo Balenzuela, Juan Pablo Pinasco, and Viktoriya Semeshenko. "The undecided have the key: Interaction-driven opinion dynamics in a three state model". In: *PloS one* 10.10 (2015), e0139572.

[15] Wolfram Barfuss. "Dynamical systems as a level of cognitive analysis of multi-agent learning: Algorithmic foundations of temporal-difference learning dynamics". In: *Neural Computing and Applications* 34.3 (2022), pp. 1653–1671.

[16] Wolfram Barfuss and Janusz M Meylahn. "Intrinsic fluctuations of reinforcement learning promote cooperation". In: *Scientific Reports* 13.1 (2023), p. 1309.

[17] Wolfram Barfuss et al. "Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse". In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12915–12922.

[18] Andrew G Barto, Philip S Thomas, and Richard S Sutton. "Some recent applications of reinforcement learning". In: *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*. 2017.

[19] C Daniel Batson and Tecia Moran. "Empathy-induced altruism in a prisoner's dilemma". In: *European Journal of Social Psychology* 29.7 (1999), pp. 909–924.

[20] Fabian Baumann et al. "Emergence of polarized ideological opinions in multidimensional topic spaces". In: *Physical Review X* 11.1 (2021), p. 011012.

[21] Richard Bellman. "On the theory of dynamic programming". In: *Proceedings of the national Academy of Sciences* 38.8 (1952), pp. 716–719.

[22] Elchanan Ben-Porath, Eddie Dekel, and Aldo Rustichini. "On the relationship between mutation rates and growth rates in a changing environment". In: *Games and Economic Behavior* 5.4 (1993), pp. 576–603.

[23] Oded Berger-Tal et al. "The exploration-exploitation dilemma: a multidisciplinary framework". In: *PloS one* 9.4 (2014), e95693.

[24] Carmela Bernardo et al. “Quantifying leadership in climate negotiations: A social power game”. In: *PNAS nexus* 2.11 (2023), pgad365.

[25] Daan Bloembergen et al. “Evolutionary dynamics of multi-agent learning: A survey”. In: *Journal of Artificial Intelligence Research* 53 (2015), pp. 659–697.

[26] Anca Bocanet and Cristina Ponsiglione. “Balancing exploration and exploitation in complex environments”. In: *Vine* (2012).

[27] Christophe Boone, Bert De Brabander, and Arjen Van Witteloostuijn. “The impact of personality on behavior in five Prisoner’s Dilemma games”. In: *Journal of Economic Psychology* 20.3 (1999), pp. 343–377.

[28] Michael Bowling and Manuela Veloso. “Multiagent learning using a variable learning rate”. In: *Artificial Intelligence* 136.2 (2002), pp. 215–250.

[29] Åke Brännström, Jacob Johansson, and Niels Von Festenberg. “The hitchhiker’s guide to adaptive dynamics”. In: *Games* 4.3 (2013), pp. 304–328.

[30] Graham Brightwell and Peter Winkler. “Counting linear extensions is# P-complete”. In: *Proceedings of the twenty-third annual ACM symposium on Theory of computing*. 1991, pp. 175–181.

[31] Sarah F Brosnan and Frans BM De Waal. “Monkeys reject unequal pay”. In: *Nature* 425.6955 (2003), pp. 297–299.

[32] Sarah F Brosnan and Frans BM de Waal. “Fairness in animals: Where to from here?” In: *Social Justice Research* 25 (2012), pp. 336–351.

[33] Michele Alessandro Bucci et al. “Control of chaotic systems by deep reinforcement learning”. In: *Proceedings of the Royal Society A* 475.2231 (2019), p. 20190351.

[34] Christoph Bühren et al. “Social preferences in the public goods game–An Agent-Based simulation with EconSim”. In: *Plos one* 18.3 (2023), e0282112.

[35] Reinhard Bürger. “Evolution of genetic variability and the advantage of sex and recombination in changing environments”. In: *Genetics* 153.2 (1999), pp. 1055–1069.

[36] Oana Carja, Uri Liberman, and Marcus W Feldman. “Evolution in changing environments: Modifiers of mutation, recombination, and migration”. In: *Proceedings of the National Academy of Sciences* 111.50 (2014), pp. 17935–17940.

[37] René Carmona, Mathieu Laurière, and Zongjun Tan. “Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning”. In: *The Annals of Applied Probability* 33.6B (2023), pp. 5334–5381.

[38] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. “Statistical physics of social dynamics”. In: *Reviews of modern physics* 81.2 (2009), pp. 591–646.

[39] Guodong Chen et al. “Deffuant model on a ring with repelling mechanism and circular opinions”. In: *Physical Review E* 95.4 (2017), p. 042118.

[40] Xiaojie Chen et al. “First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation”. In: *Journal of the royal society interface* 12.102 (2015), p. 20140935.

[41] Xiaojie Chen et al. “First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation”. In: *Journal of the royal society interface* 12.102 (2015), p. 20140935.

[42] Xingru Chen and Feng Fu. "Ensuring the greater good in hybrid AI-human systems: Comment on" Reputation and reciprocity" by Xia et al". In: *Physics of life reviews* 48 (2023), pp. 41–43.

[43] Xingru Chen and Feng Fu. "Outlearning extortioners: unbending strategies can foster reciprocal fairness and cooperation". In: *PNAS nexus* 2.6 (2023), pgad176.

[44] Jiangjiang Cheng et al. "Multidimensional opinion dynamics with heterogeneous bounded confidences and random interactions". In: *Automatica* 172 (2025), p. 112002.

[45] Gabriele Chierchia and Tania Singer. "The neuroscience of compassion and empathy and their link to prosocial motivation and behavior". In: *Decision neuroscience*. Elsevier, 2017, pp. 247–257.

[46] Manjusha Chintalapati and Priya Moorjani. "Evolution of the mutation rate across primates". In: *Current opinion in genetics & development* 62 (2020), pp. 58–64.

[47] Siang Yew Chong et al. "The iterated Prisoner's Dilemma: 20 years on". In: *Advances in Natural Competition* 4 (2007), pp. 1–22.

[48] Brian Christian and Tom Griffiths. *Algorithms to live by: The computer science of human decisions*. Macmillan, 2016.

[49] Claire de Mazancourt and Ulf Dieckmann. "Trade-Off Geometries and Frequency-Dependent Selection". In: *The American Naturalist* 164 (2004), pp. 765–778.

[50] Kenneth Clark and Martin Sefton. "The sequential prisoner's dilemma: evidence on reciprocation". In: *The economic journal* 111.468 (2001), pp. 51–68.

[51] Jens Christian Claussen and Arne Traulsen. "Cyclic Dominance and Biodiversity in Well-Mixed Populations". In: *Physical Review Letters* 100 (2008).

[52] Peter Clifford and Aidan Sudbury. "A model for spatial conflict". In: *Biometrika* 60.3 (1973), pp. 581–588.

[53] David Alexander Coley. *An introduction to genetic algorithms for scientists and engineers*. World Scientific Publishing Company, 1999.

[54] Elisabetta Cornacchia, Neta Singer, and Emmanuel Abbe. "Polarization in attraction-repulsion models". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020, pp. 2765–2770.

[55] Chenna Reddy Cotla. "Learning in Repeated Public Goods Games-A Meta Analysis". In: *Available at SSRN 3241779* (2015).

[56] Marta C Couto, Stefano Giaimo, and Christian Hilbe. "Introspection dynamics: a simple model of counterfactual learning in asymmetric games". In: *New Journal of Physics* 24.6 (2022), p. 063010.

[57] Matej Črepinšek, Shih-Hsi Liu, and Marjan Mernik. "Exploration and exploitation in evolutionary algorithms: A survey". In: *ACM computing surveys (CSUR)* 45.3 (2013), pp. 1–33.

[58] Peng-Bi Cui. "Exploring the foundation of social diversity and coherence with a novel attraction–repulsion model framework". In: *Physica A: Statistical Mechanics and its Applications* 618 (2023), p. 128714.

[59] Allan Dafoe et al. 2021.

[60] Abir De et al. "Learning and forecasting opinion dynamics in social networks". In: *Advances in neural information processing systems* 29 (2016).

[61] Morris H DeGroot. "Reaching a consensus". In: *Journal of the American Statistical association* 69.345 (1974), pp. 118–121.

[62] Ulf Dieckmann, Mikko Heino, and Kalle Parvinen. "The adaptive dynamics of function-valued traits". In: *Journal of Theoretical Biology* 241 (2006), pp. 370–389.

[63] Ulf Dieckmann et al. "Adaptive dynamics of infectious diseases". In: *Pursuit of virulence management* (2002), pp. 460–463.

[64] Odo Diekmann et al. "A beginner's guide to adaptive dynamics". In: *Banach Center Publications* 63 (2004), pp. 47–86.

[65] "Discussion of Dr Gittins' Paper". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (Dec. 2018), pp. 164–177. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1979.tb01069.x. eprint: https://academic.oup.com/jrsssb/article-pdf/41/2/164/49097406/jrsssb_41_2_164.pdf. URL: https://doi.org/10.1111/j.2517-6161.1979.tb01069.x.

[66] Michael Doebeli and Christoph Hauert. "Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game". In: *Ecology letters* 8.7 (2005), pp. 748–766.

[67] Michael Doebeli, Chirstoph Haurt, and Timothy Killingback. "The evolutionary origin of cooperators and defectors". In: *Science* 306.5697 (2004), pp. 859–862.

[68] Esteban Domingo et al. “Mutation rates, mutation frequencies, and proofreading-repair activities in RNA virus genetics”. In: *Viruses* 13.9 (2021), p. 1882.

[69] Stephen Dominic et al. “Genetic reinforcement learning for neural networks”. In: *IJCNN-91-Seattle International Joint Conference on Neural Networks*. Vol. 2. IEEE. 1991, pp. 71–76.

[70] Wei Du and Shifei Ding. “A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications”. In: *Artificial Intelligence Review* 54.5 (2021), pp. 3215–3238.

[71] Agoston E Eiben and Cornelis A Schippers. “On evolutionary exploration and exploitation”. In: *Fundamenta Informaticae* 35.1-4 (1998), pp. 35–50.

[72] Manfred Eigen and Peter Schuster. “A principle of natural self-organization”. In: *Naturwissenschaften* 64 (1977), pp. 541–565.

[73] Sigrunn Eliassen et al. “Exploration or exploitation: life expectancy changes the value of learning in foraging strategies”. In: *Oikos* 116.3 (2007), pp. 513–523.

[74] Seyed Rasoul Etesami et al. “Termination time of multidimensional Hegselmann-Krause opinion dynamics”. In: *2013 American Control Conference*. IEEE. 2013, pp. 1255–1260.

[75] Tucker Evans and Feng Fu. “Opinion formation on dynamic networks: identifying conditions for the emergence of partisan echo chambers”. In: *Royal Society open science* 5.10 (2018), p. 181122.

[76] Pengyi Fan et al. "Opinion interaction network: Opinion dynamics in social networks with heterogeneous relationships". In: *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. 2012, pp. 1–8.

[77] Ernst Fehr, Urs Fischbacher, and Simon Gächter. "Strong reciprocity, human cooperation, and the enforcement of social norms". In: *Human nature* 13 (2002), pp. 1–25.

[78] Xiaoli Feng et al. "Error thresholds for quasispecies on single peak Gaussian-distributed fitness landscapes". In: *Journal of Theoretical Biology* 246.1 (2007), pp. 28–32.

[79] Feng Fu et al. "Evolutionary dynamics on graphs: Efficient method for weak selection". In: *Physical Review E Statistical Nonlinear Soft Matter Physics* 79 (2009).

[80] Drew Fudenberg and Lorens A Imhof. "Imitation processes with small mutations". In: *Journal of Economic Theory* 131.1 (2006), pp. 251–262.

[81] Simon Gächter and Benedikt Herrmann. "Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1518 (2009), pp. 791–806.

[82] Aram Galstyan. "Continuous strategy replicator dynamics for multi-agent q-learning". In: *Autonomous agents and multi-agent systems* 26 (2013), pp. 37–53.

[83] S.A.H. Geritz et al. “Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree”. In: *Evolutionary Ecology* 12 (1998), pp. 35–37.

[84] Stefan A. H. Geritz et al. “Dynamics of Adaptation and Evolutionary Branching”. In: *Physical Review Letters* 78 (1997), pp. 2024–2027.

[85] Victor Gilsing and Bart Nooteboom. “Exploration and exploitation in innovation systems: The case of pharmaceutical biotechnology”. In: *Research policy* 35.1 (2006), pp. 1–23.

[86] Jesús Giráldez-Cru, Carmen Zarco, and Oscar Cordón. “Analyzing the extremization of opinions in a general framework of bounded confidence and repulsion”. In: *Information Sciences* 609 (2022), pp. 1256–1270.

[87] John Gittins. “A dynamic allocation index for the sequential design of experiments”. In: *Progress in statistics* (1974), pp. 241–266.

[88] Michel Grabisch and Agnieszka Rusinowska. “A survey on nonstrategic models of opinion dynamics”. In: *Games* 11.4 (2020), p. 65.

[89] Henrich R Greve. “Exploration and exploitation in product innovation”. In: *Industrial and corporate change* 16.5 (2007), pp. 945–975.

[90] Sven Gronauer and Klaus Diepold. “Multi-agent deep reinforcement learning: a survey”. In: *Artificial Intelligence Review* (2022), pp. 1–49.

[91] Dmitry A Gubanov, Ilya V Petrov, and Alexander G Chkhartishvili. “Multidimensional model of opinion dynamics in social networks: polarization indices”. In: *Automation and Remote Control* 82 (2021), pp. 1802–1811.

[92] Lingaraj Haldurai, T Madhubala, and R Rajalakshmi. "A study on genetic algorithm and its applications". In: *Int. J. Comput. Sci. Eng* 4.10 (2016), pp. 139–143.

[93] Wenchen Han et al. "Non-consensus states in circular opinion model with repulsive interaction". In: *Physica A: Statistical Mechanics and its Applications* 585 (2022), p. 126428.

[94] Marc Harper et al. "Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma". In: *PloS one* 12.12 (2017), e0188046.

[95] Hossein Hassani et al. "Classical dynamic consensus and opinion dynamics models: A survey of recent trends and methodologies". In: *Information Fusion* 88 (2022), pp. 22–40.

[96] Christoph Hauert and Michael Doebeli. "Spatial structure often inhibits the evolution of cooperation in the snowdrift game". In: *Nature* 428.6983 (2004), pp. 643–646.

[97] Christoph Hauert, Miranda Holmes, and Michael Doebeli. "Evolutionary games and population dynamics: maintenance of cooperation in public goods games". In: *Proceedings of the Royal Society B: Biological Sciences* 273.1600 (2006), pp. 2565–2571.

[98] Walid Hichri and Alan Kirman. "The emergence of coordination in public good games". In: *The European Physical Journal B* 55 (2007), pp. 149–159.

[99] Christian Hilbe, Maria Kleshnina, and Kateřina Staňková. "Evolutionary Games and Applications: Fifty Years of 'The Logic of Animal Conflict'". In: *Dynamic Games and Applications* 13.4 (2023), pp. 1035–1048.

[100] Christian Hilbe, Martin A. Nowak, and Arne Traulsen. "Adaptive Dynamics of Extortion and Compliance". In: *PLoS ONE* 8 (2013).

[101] Christian Hilbe et al. "Cooperation and control in multiplayer social dilemmas". In: *Proceedings of the National Academy of Sciences* 111.46 (2014), pp. 16425–16430.

[102] Christian Hilbe et al. "Evolution of cooperation in stochastic games". In: *Nature* 559.7713 (2018), pp. 246–249.

[103] WGS Hines. "Evolutionary stable strategies: a review of basic theory". In: *Theoretical Population Biology* 31.2 (1987), pp. 195–272.

[104] J. Hofbauer, P. Schuster, and K. Sigmund. "A note on evolutionary stable strategies and game dynamics". In: *Journal of Theoretical Biology* 81 (1979), pp. 609–612.

[105] Richard A Holley and Thomas M Liggett. "Ergodic theorems for weakly interacting infinite systems and the voter model". In: *The annals of probability* (1975), pp. 643–663.

[106] Manh Hong Duong and The Anh Han. "On Equilibrium Properties of the Replicator–Mutator Equation in Deterministic and Random Games". In: *Dynamic Games and Applications* 10 (2020), pp. 641–663.

[107] Yutaka Horita et al. "Reinforcement learning accounts for moody conditional cooperation behavior: experimental results". In: *Scientific reports* 7.1 (2017), p. 39275.

[108] Changwei Huang, Huanyu Bian, and Wenchen Han. "Breaking the symmetry neutralizes the extremization under the repulsion and higher order interactions". In: *Chaos, Solitons & Fractals* 180 (2024), p. 114544.

[109] Kazushige Ishii et al. "Evolutionarily stable mutation rate in a periodically changing environment." In: *Genetics* 121.1 (1989), pp. 163–174.

[110] Atsushi Iwasaki et al. "Does reinforcement learning simulate threshold public goods games?: a comparison with subject experiments". In: *IEICE TRANSACTIONS on Information and Systems* 86.8 (2003), pp. 1335–1343.

[111] Segismundo S Izquierdo, Luis R Izquierdo, and Nicholas M Gotts. "Reinforcement learning dynamics in social dilemmas". In: *Journal of Artificial Societies and Social Simulation* 11.2 (2008), p. 1.

[112] Segismundo S. Izquierdo and Luis R. Izquierdo. "Strictly Dominated Strategies in the Replicator-Mutator Dynamics". In: *Games* (2011), pp. 355–364.

[113] Marco A Janssen and Toh-Kyeong Ahn. "Learning, signaling, and social preferences in public-good games". In: *Ecology and society* 11.2 (2006).

[114] Garett Jones. "Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003". In: *Journal of Economic Behavior & Organization* 68.3-4 (2008), pp. 489–497.

[115] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. "Reinforcement learning: A survey". In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285.

[116] Michael Kaisers and Karl Tuyls. "FAQ-Learning in Matrix Games: Demonstrating Convergence Near Nash Equilibria, and Bifurcation of Attractors in the Battle of Sexes." In: *Interactive Decision Theory and Game Theory*. 2011.

[117] Unchitta Kan, Michelle Feng, and Mason A Porter. "An adaptive bounded-confidence model of opinion dynamics on networks". In: *Journal of Complex Networks* 11.1 (2023), pp. 415–444.

[118] Kamran Kaveh, Natalia L Komarova, and Mohammad Kohandel. "The duality of spatial death–birth and birth–death processes and limitations of the isothermal theorem". In: *Royal Society open science* 2.4 (2015), p. 140465.

[119] Taylor A Kessinger, Corina E Tarnita, and Joshua B Plotkin. "Evolution of norms for judging social behavior". In: *Proceedings of the National Academy of Sciences* 120.24 (2023), e2219480120.

[120] Ardeshir Kianercy and Aram Galstyan. "Dynamics of Boltzmann Q learning in two-player two-action games". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85.4 (2012), p. 041145.

[121] Man-Je Kim, Jun Suk Kim, and Chang Wook Ahn. "Evolving population method for real-time reinforcement learning". In: *Expert Systems with Applications* 229 (2023), p. 120493.

[122] Natalia Komarovaa. "Replicator–mutator equation, universality property and population dynamics of learning". In: *Journal of Theoretical Biology* 230.2 (2004), pp. 227–239.

[123] Samuel S Komorita, JA Hilty, and Craig D Parks. "Reciprocity and cooperation in social dilemmas". In: *Journal of Conflict Resolution* 35.3 (1991), pp. 494–518.

[124] Ivan V Kozitsin. "A general framework to link theory and empirics in opinion formation models". In: *Scientific reports* 12.1 (2022), p. 5543.

[125] Marc Krasovec, Rosalind EM Rickaby, and Dmitry A Filatov. "Evolution of mutation rate in astronomically large phytoplankton populations". In: *Genome biology and evolution* (2020).

[126] K Praveen Kumar et al. "Balancing Exploration and Exploitation in Nature Inspired Computing Algorithm". In: *Intelligent Cyber Physical Systems and Internet of Things: ICoICI 2022*. Springer, 2023, pp. 163–172.

[127] Evguenii Kurmyshev, Héctor A Juárez, and Ricardo A González-Silva. "Dynamics of bounded confidence opinion in heterogeneous social networks: Concord against partial antagonism". In: *Physica A: Statistical Mechanics and its Applications* 390.16 (2011), pp. 2945–2955.

[128] Shun Kurokawa and Yasuo Ihara. "Emergence of cooperation in public goods games". In: *Proceedings of the Royal Society B: Biological Sciences* 276.1660 (2009), pp. 1379–1384.

[129] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[130] Nicolas Lanchier and Max Mercer. "Deffuant opinion dynamics with attraction and repulsion". In: *Electronic Communications in Probability* 29 (2024), pp. 1–12.

[131] Daniella Laureiro-Martinez, Stefano Brusoni, and Maurizio Zollo. "The neuroscientific foundations of the exploration- exploitation dilemma." In: *Journal of Neuroscience, Psychology, and Economics* 3.2 (2010), p. 95.

[132] Dovev Lavie and Lori Rosenkopf. "Balancing exploration and exploitation in alliance formation". In: *Academy of management journal* 49.4 (2006), pp. 797–818.

[133] David Lazer and Allan Friedman. "The network structure of exploration and exploitation". In: *Administrative science quarterly* 52.4 (2007), pp. 667–694.

[134] Olof Leimar and Peter Hammerstein. "Evolution of cooperation through indirect reciprocity". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1468 (2001), pp. 745–753.

[135] Olof Leimar and John M McNamara. "Game theory in biology: 50 years and onwards". In: *Philosophical Transactions of the Royal Society B* 378.1876 (2023), p. 20210509.

[136] Olof Leimar and John M McNamara. "Learning leads to bounded rationality and the evolution of cognitive bias in public goods games". In: *Scientific reports* 9.1 (2019), p. 16319.

[137] Naomi Ehrich Leonard and Simon A Levin. "Collective intelligence as a public good". In: *Collective Intelligence* 1.1 (2022), p. 26339137221083293.

[138] Stefanos Leonardos and Georgios Piliouras. "Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory". In: *Artificial Intelligence* 304 (2022), p. 103653.

[139] Bin Li et al. “Machine learning-enabled globally guaranteed evolutionary computation”. In: *Nature Machine Intelligence* (2023), pp. 1–11.

[140] Ye Li and Claus O. Wilke. “Digital Evolution in Time-Dependent Fitness Landscapes”. In: *Artificial Life* 10.2 (2004).

[141] Uri Liberman, Hilla Behar, and Marcus W. Feldman. “Evolution of reduced mutation under frequency-dependent selection”. In: *Theoretical Population Biology* 112 (2016), pp. 52–59.

[142] Uri Liberman, Jeremy Van Cleve, and Marcus W Feldman. “On the evolution of mutation in changing environments: recombination and phenotypic switching”. In: *Genetics* 187.3 (2011), pp. 837–851.

[143] Erez Lieberman, Christoph Hauert, and Martin A Nowak. “Evolutionary dynamics on graphs”. In: *Nature* 433.7023 (2005), pp. 312–316.

[144] Yuanguo Lin et al. “Evolutionary Reinforcement Learning: A Systematic Review and Future Directions”. In: *arXiv preprint arXiv:2402.13296* (2024).

[145] Fei Liu and Guangzhou Zeng. “Study of genetic algorithm with reinforcement learning to solve the TSP”. In: *Expert Systems with Applications* 36.3 (2009), pp. 6995–7001.

[146] Zuxin Liu et al. “Mapper: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 11748–11754.

[147] Jan Lorenz. "Continuous opinion dynamics under bounded confidence: A survey". In: *International Journal of Modern Physics C* 18.12 (2007), pp. 1819–1838.

[148] Michael Lynch. "Evolution of the mutation rate". In: *TRENDS in Genetics* 26.8 (2010), pp. 345–352.

[149] Michael Lynch et al. "Genetic drift, selection and the evolution of the mutation rate". In: *Nature Reviews Genetics* 17.11 (2016), pp. 704–714.

[150] LK M'Gonigle, JJ Shen, and SP Otto. "Mutating away from your enemies: the evolution of mutation rate in a host–parasite system". In: *Theoretical Population Biology* 75.4 (2009), pp. 301–311.

[151] Diane Maclagan. "Boolean term orders and the root system Bn". In: *Order* 15.3 (1998), pp. 279–295.

[152] Michael W Macy and Andreas Flache. "Learning dynamics in social dilemmas". In: *Proceedings of the National Academy of Sciences* 99.suppl_3 (2002), pp. 7229–7236.

[153] U ManChon and Zhen Li. "Public goods game simulator with reinforcement learning agents". In: *2010 Ninth International Conference on Machine Learning and Applications*. IEEE. 2010, pp. 43–49.

[154] James G March. "Exploration and exploitation in organizational learning". In: *Organization science* 2.1 (1991), pp. 71–87.

[155] Naoki Masuda and Mitsuhiro Nakamura. "Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated Prisoner's dilemma". In: *Journal of theoretical biology* 278.1 (2011), pp. 55–62.

[156] Blake D Mathias, Aaron F Mckenny, and T Russell Crook. "Managing the tensions between exploration and exploitation: The role of time". In: *Strategic Entrepreneurship Journal* 12.3 (2018), pp. 316–334.

[157] Ivan Matic. "Mutation rate heterogeneity increases odds of survival in unpredictable environments". In: *Molecular cell* 75.3 (2019), pp. 421–425.

[158] Alex McAvoy et al. "Evolutionary instability of selfish learning in repeated games". In: *PNAS nexus* 1.4 (2022), pgac141.

[159] Brian J. McGill and Joel S. Brown. "Evolutionary Game Theory and Adaptive Dynamics of Continuous Traits". In: *Annual Review of Ecology, Evolution, and Systematics* 38 (2007), pp. 403–435.

[160] Luke McNally, Sam P Brown, and Andrew L Jackson. "Cooperation and the evolution of intelligence". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1740 (2012), pp. 3027–3034.

[161] G. Meszéna et al. "Evolutionary Optimisation Models and Matrix Games in the Unified Perspective of Adaptive Dynamics". In: *Selection* 2 (2005), pp. 193–220.

[162] Risto Miikkulainen and Stephanie Forrest. "A biological perspective on evolutionary computation". In: *Nature Machine Intelligence* 3.1 (2021), pp. 9–15.

[163] Manfred Milinski. "Tit for tat in sticklebacks and the evolution of cooperation". In: *nature* 325.6103 (1987), pp. 433–435.

[164] Brian Mintz and Feng Fu. "Social learning and the exploration-exploitation tradeoff". In: *Computation* 11.5 (2023), p. 101.

[165] Brian Mintz and Feng Fu. "The Point of No Return: Evolution of Excess Mutation Rate Is Possible Even for Simple Mutation Models". In: *Mathematics* 10.24 (2022), p. 4818.

[166] Christopher T Monk et al. "How ecology shapes exploitation: a framework to predict the behavioural response of human and animal foragers along exploration–exploitation trade-offs". In: *Ecology letters* 21.6 (2018), pp. 779–793.

[167] David E Moriarty, Alan C Schultz, and John J Grefenstette. "Evolutionary algorithms for reinforcement learning". In: *Journal of Artificial Intelligence Research* 11 (1999), pp. 241–276.

[168] Jan Narveson. "The how and why of universalizability". In: *Morality and universality: Essays on ethical universalizability*. Springer, 1985, pp. 3–44.

[169] Daniel J Navarro, Ben R Newell, and Christin Schulze. "Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments". In: *Cognitive psychology* 85 (2016), pp. 43–77.

[170] Heinrich H Nax and Matjaž Perc. "Directional learning and the provisioning of public goods". In: *Scientific reports* 5.1 (2015), pp. 1–6.

[171] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications". In: *IEEE transactions on cybernetics* 50.9 (2020), pp. 3826–3839.

[172] Martin Nilsson and Nigel Snoad. "Error Thresholds for Quasispecies on Dynamic Fitness Landscapes". In: *Physical Review Letters* 84.1 (2000), pp. 191–194.

[173] Martin Nilsson and Nigel Snoad. "Optimal mutation rates in dynamic environments". In: *Bulletin of Mathematical Biology* 64 (2002), pp. 1033–1043.

[174] Martin Nilsson and Nigel Snoad. "Quasispecies evolution on a fitness landscape with a fluctuating peak". In: *Physical Review E* 65 (2002).

[175] Sirikwan Noipitak and Michael A Allen. "Dynamics of interdependent multidimensional opinions". In: *Journal of Physics: Conference Series*. Vol. 1719. 1. IOP Publishing. 2021, p. 012107.

[176] Martin Nowak. *Evolutionary Dynamics; Exploring the equations of life*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 2006.

[177] Martin A Nowak. *Evolutionary dynamics: exploring the equations of life*. Harvard university press, 2006.

[178] Martin A Nowak. "Five rules for the evolution of cooperation". In: *science* 314.5805 (2006), pp. 1560–1563.

[179] Martin A Nowak and Robert M May. "Evolutionary games and spatial chaos". In: *nature* 359.6398 (1992), pp. 826–829.

[180] Martin A Nowak, Karen M Page, and Karl Sigmund. "Fairness versus reason in the ultimatum game". In: *Science* 289.5485 (2000), pp. 1773–1775.

[181] Martin A Nowak et al. "Emergence of cooperation and evolutionary stability in finite populations". In: *Nature* 428.6983 (2004), pp. 646–650.

[182] Hisashi Ohtsuki and Yoh Iwasa. "The leading eight: social norms that can maintain cooperation by indirect reciprocity". In: *Journal of theoretical biology* 239.4 (2006), pp. 435–444.

[183] Afshin Oroojlooy and Davood Hajinezhad. "A review of cooperative multi-agent deep reinforcement learning". In: *Applied Intelligence* 53.11 (2023), pp. 13677–13722.

[184] Sarah P Otto, Maria R Servedio, and Scott L Nuismer. "Frequency-Dependent Selection and the Evolution of Assortative Mating". In: *Genetics* 179 (2008), pp. 2091–2112.

[185] Karen M Page, Martin A Nowak, and Karl Sigmund. "The spatial ultimatum game". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1458 (2000), pp. 2177–2182.

[186] Karen M. Page and Martin A. Nowak. "Unifying Evolutionary Dynamics". In: *Journal of Theoretical Biology* 219 (2002), pp. 93–98.

[187] Sergey E Parsegov et al. "Novel multidimensional models of opinion dynamics in social networks". In: *IEEE Transactions on Automatic Control* 62.5 (2016), pp. 2270–2285.

[188] Lucía Pedraza et al. "An analytical formulation for multidimensional continuous opinion models". In: *Chaos, Solitons & Fractals* 152 (2021), p. 111368.

[189] Matjaž Perc and Attila Szolnoki. "Social diversity and promotion of cooperation in the spatial prisoner's dilemma game". In: *Physical Review E* 77.1 (2008), p. 011904.

[190] Matjaž Perc et al. "Statistical physics of human cooperation". In: *Physics Reports* 687 (2017), pp. 1–51.

[191] ATD Perera and Parameswaran Kamalaruban. "Applications of reinforcement learning in energy systems". In: *Renewable and Sustainable Energy Reviews* 137 (2021), p. 110618.

[192] Hart E Posen and Daniel A Levinthal. "Chasing a moving target: Exploitation and exploration in dynamic environments". In: *Management science* 58.3 (2012), pp. 587–601.

[193] Alejandro Radillo-Díaz, Luis A Pérez, and Marcelo del Castillo-Mussot. "Axelrod models of social influence with cultural repulsion". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 80.6 (2009), p. 066107.

[194] Hegselmann Rainer and Ulrich Krause. "Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation". In: *Journal of Artificial Societies and Social Simulation* 5.3 (2002).

[195] Jouni Reinikainen. "The golden rule and the requirement of universalizability". In: *J. Value Inquiry* 39 (2005), p. 155.

[196] Xiang-Yi Li Richter and Jussi Lehtonen. *Half a century of evolutionary games: a synthesis of theory, application and future directions*. 2023.

[197] Daniel Romero-Mujalli, Florian Jeltsch, and Ralph Tiedemann. "Elevated mutation rates are unlikely to evolve in sexual species, not even under rapid environmental change". In: *BMC Evolutionary Biology* 19 (2019), pp. 1–9.

[198] Daniel IS Rosenbloom and Benjamin Allen. "Frequency-dependent selection can lead to evolution of high mutation rates". In: *The American Naturalist* 183.5 (2014), E131–E153.

[199] David Sabin-Miller and Daniel M Abrams. “When pull turns to shove: A continuous-time model for opinion dynamics”. In: *Physical Review Research* 2.4 (2020), p. 043001.

[200] Tuomas W Sandholm and Robert H Crites. “Multiagent reinforcement learning in the iterated prisoner’s dilemma”. In: *Biosystems* 37.1-2 (1996), pp. 147–166.

[201] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. “The complexity of human cooperation under indirect reciprocity”. In: *Philosophical Transactions of the Royal Society B* 376.1838 (2021), p. 20200291.

[202] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. “Social norm complexity and past reputations in the evolution of cooperation”. In: *Nature* 555.7695 (2018), pp. 242–245.

[203] Francisco C Santos and Jorge M Pacheco. “Scale-free networks provide a unifying framework for the emergence of cooperation”. In: *Physical review letters* 95.9 (2005), p. 098104.

[204] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. “Evolutionary dynamics of social dilemmas in structured heterogeneous populations”. In: *Proceedings of the National Academy of Sciences* 103.9 (2006), pp. 3490–3494.

[205] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. “Social diversity promotes the emergence of cooperation in public goods games”. In: *Nature* 454.7201 (2008), pp. 213–216.

[206] Hendrik Schawe and Laura Hernández. “Higher order interactions destroy phase transitions in deffuant opinion dynamics model”. In: *Communications Physics* 5.1 (2022), p. 32.

[207] Laura Schmid et al. “A unified framework of direct and indirect reciprocity”. In: *Nature Human Behaviour* 5.10 (2021), pp. 1292–1302.

[208] Peter Schuster and Jorg Swetina. “Stationary Mutant Distributions and Evolutionary Optimization”. In: *Bulletin of Mathematical Biology* 50 (1988), pp. 635–660.

[209] Adarsh Sehgal et al. “Deep reinforcement learning using genetic algorithm for parameter optimization”. In: *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE. 2019, pp. 596–601.

[210] Yanxuan Shao, Xin Wang, and Feng Fu. “Evolutionary dynamics of group cooperation with asymmetrical environmental feedback”. In: *Europhysics Letters* 126.4 (2019), p. 40005.

[211] Longmei Shu and Feng Fu. “Determinants of successful mitigation in coupled social-climate dynamics”. In: *Proceedings of the Royal Society A* 479.2280 (2023), p. 20230679.

[212] Longmei Shu and Feng Fu. “Eco-evolutionary dynamics of bimatrix games”. In: *Proceedings of the Royal Society A* 478.2267 (2022), p. 20220567.

[213] Alina Sırbu et al. “Opinion dynamics: models, extensions and external effects”. In: *Participatory sensing, opinions and collective awareness* (2017), pp. 363–401.

[214] J Maynard Smith and George R Price. “The logic of animal conflict”. In: *Nature* 246.5427 (1973), pp. 15–18.

[215] Mei-Ping Song and Guo-Chang Gu. "Research on particle swarm optimization: a review". In: *Proceedings of 2004 international conference on machine learning and cybernetics (IEEE Cat. No. 04EX826)*. Vol. 4. IEEE. 2004, pp. 2236–2241.

[216] Serap Tay Stamoulas and Muruhan Rathinam. "Convergence, stability, and robustness of multidimensional opinion dynamics in continuous time". In: *SIAM Journal on Control and Optimization* 56.3 (2018), pp. 1938–1967.

[217] Alexander J Stewart and Joshua B Plotkin. "From extortion to generosity, evolution in the iterated prisoner's dilemma". In: *Proceedings of the National Academy of Sciences* 110.38 (2013), pp. 15348–15353.

[218] Qi Su, Alex McAvoy, and Joshua B Plotkin. "Evolution of cooperation with contextualized behavior". In: *Science advances* 8.6 (2022), eabm6066.

[219] Weiwei Sun et al. "Combination of institutional incentives for cooperative governance of risky commons". In: *Iscience* 24.8 (2021), p. 102844.

[220] Jorg Swetina and Peter Schuster. "Self replication with errors, A model for polynucleotide replication". In: *Biophysical Chemistry* 16.4 (1982), pp. 329–345.

[221] Jeremy Sylvestre. "Elementary Foundations: An Introduction to Topics in Discrete Mathematics". In: University of Alberta, 2018. Chap. 19, pp. 207–225.

[222] Gyoergy Szabo and Attila Szolnoki. "Selfishness, fraternity, and other-regarding preference in spatial evolutionary games". In: *Journal of theoretical biology* 299 (2012), pp. 81–87.

[223] Gyoergy Szabo, Attila Szolnoki, and Lilla Czako. "Coexistence of fraternity and egoism for spatial social dilemmas". In: *Journal of theoretical biology* 317 (2013), pp. 126–132.

[224] György Szabó and Gabor Fath. "Evolutionary games on graphs". In: *Physics reports* 446.4-6 (2007), pp. 97–216.

[225] György Szabó and Csaba Tőke. "Evolutionary prisoner's dilemma game on a square lattice". In: *Physical Review E* 58.1 (1998), p. 69.

[226] Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

[227] Attila Szolnoki and Matjaž Perc. "Conditional strategies and the evolution of cooperation in spatial public goods games". In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 85.2 (2012), p. 026104.

[228] Zhiping Tan and Kangshun Li. "Differential evolution with mixed mutation strategy based on deep reinforcement learning". In: *Applied Soft Computing* 111 (2021), p. 107678.

[229] Kit-Sang Tang et al. "Genetic algorithms and their applications". In: *IEEE signal processing magazine* 13.6 (1996), pp. 22–37.

[230] Peter D Taylor and Leo B Jonker. "Evolutionary stable strategies and game dynamics". In: *Mathematical biosciences* 40.1-2 (1978), pp. 145–156.

[231] Peter D. Taylor and Leo B. Jonker. "Evolutionary stable strategies and game dynamics". In: *Mathematical Biosciences* 40 (1978), pp. 145–156.

[232] Andrew R Tilman, Joshua B Plotkin, and Erol Akçay. "Evolutionary games with environmental feedbacks". In: *Nature communications* 11.1 (2020), p. 915.

[233] Ian PM Tomlinson, MR Novelli, and WF Bodmer. "The mutation rate and cancer". In: *Proceedings of the National Academy of Sciences* 93.25 (1996), pp. 14800–14803.

[234] Arne Traulsen and Nikoleta E Glynatsi. "The future of theoretical evolutionary game theory". In: *Philosophical Transactions of the Royal Society B* 378.1876 (2023), p. 20210508.

[235] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. "Stochastic dynamics of invasion and fixation". In: *Physical Review E* 74.1 (2006), p. 011909.

[236] Milena Tsvetkova et al. "A new sociology of humans and machines". In: *Nature Human Behaviour* 8.10 (2024), pp. 1864–1876.

[237] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. "A selection-mutation model for q-learning in multi-agent systems". In: *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 2003, pp. 693–700.

[238] Masahiko Ueda. "Memory-two strategies forming symmetric mutual reinforcement learning equilibrium in repeated prisoners' dilemma game". In: *Applied Mathematics and Computation* 444 (2023), p. 127819.

[239] Yuki Usui and Masahiko Ueda. "Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma". In: *Applied Mathematics and Computation* 409 (2021), p. 126370.

[240] Joe Yuichiro Wakano and Laurent Lehmann. "Evolutionary and convergence stability for continuous phenotypes in finite populations derived from two-allele models". In: *Journal of Theoretical Biology* 310 (2012), pp. 206–215.

[241] Dongshu Wang, Dapei Tan, and Lei Liu. "Particle swarm optimization algorithm: an overview". In: *Soft computing* 22.2 (2018), pp. 387–408.

[242] Juan Wang and Chengyi Xia. "Reputation evaluation and its impact on the human cooperation—A recent survey". In: *Europhysics Letters* 141.2 (2023), p. 21001.

[243] Lu Wang et al. "Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning". In: *Nonlinear Dynamics* 108.2 (2022), pp. 1837–1845.

[244] Lu Wang et al. "Synergistic effects of adaptive reward and reinforcement learning rules on cooperation". In: *New Journal of Physics* 25.7 (2023), p. 073008.

[245] Qiang Wang et al. "Evolutionary Dynamics of Population Games With an Aspiration-Based Learning Rule". In: *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[246] Xianjia Wang et al. "Enhancing cooperative evolution in spatial public goods game by particle swarm optimization based on exploration and q-learning". In: *Applied Mathematics and Computation* 469 (2024), p. 128534.

[247] Xin Wang and Feng Fu. "Eco-evolutionary dynamics with environmental feedback: Cooperation in a changing world". In: *Europhysics Letters* 132.1 (2020), p. 10001.

[248] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine learning* 8 (1992), pp. 279–292.

[249] Joshua S Weitz et al. "An oscillating tragedy of the commons in replicator dynamics with game-environment feedback". In: *Proceedings of the National Academy of Sciences* 113.47 (2016), E7518–E7525.

[250] Claus O. Wilke, Christopher Ronnewinkel, and Thomas Martinetz. "Molecular Evolution in Time-Dependent Environments". In: *Advances in Artificial Life*. Ed. by Dario Floreano, Jean-Daniel Nicoud, and Francesco Mondada. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 417–421. ISBN: 978-3-540-48304-5.

[251] Gerald S Wilkinson et al. "Non-kin cooperation in bats". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1687 (2016), p. 20150095.

[252] Max Wolf, G Sander Van Doorn, and Franz J Weissing. "On the coevolution of social responsiveness and behavioural consistency". In: *Proceedings of the Royal Society B: Biological Sciences* 278.1704 (2011), pp. 440–448.

[253] Te Wu, Feng Fu, and Long Wang. "Moving away from nasty encounters enhances cooperation in ecological prisoner's dilemma game". In: *PLoS One* 6.11 (2011), e27669.

[254] Te Wu et al. "Adaptive role switching promotes fairness in networked ultimatum game". In: *Scientific reports* 3.1 (2013), p. 1550.

[255] Tong Wu. "Heterogeneous opinion dynamics considering consensus evolution in social network group decision-making". In: *Group Decision and Negotiation* 33.1 (2024), pp. 159–194.

[256] Xiaoqiang Wu et al. "Evolutionary Reinforcement Learning with Action Sequence Search for Imperfect Information Games". In: *Information Sciences* (2024), p. 120804.

[257] Chengyi Xia et al. "Reputation and reciprocity". In: *Physics of Life Reviews* (2023).

[258] Vicky Chuqiao Yang et al. "Dynamical system model predicts when social learners impair collective performance". In: *Proceedings of the National Academy of Sciences* 118.35 (2021), e2106292118.

[259] Yaodong Yang et al. "Mean field multi-agent reinforcement learning". In: *International conference on machine learning*. PMLR. 2018, pp. 5571–5580.

[260] Gary Yen, Fengming Yang, and Travis Hickey. "Coordination of exploration and exploitation in a dynamic environment". In: *International Journal of Smart Engineering System Design* 4.3 (2002), pp. 177–182.

[261] Chao Yu et al. "Emotional multiagent reinforcement learning in spatial social dilemmas". In: *IEEE transactions on neural networks and learning systems* 26.12 (2015), pp. 3083–3096.

[262] E. C. Zeeman. "Population dynamics from game theory". In: *Lecture Notes in Mathematics* 819 (1980).

[263] Yanling Zhang et al. "Cooperation in group-structured populations with two layers of interactions". In: *Scientific Reports* 5.1 (2015), p. 17446.

[264] Qingling Zhu et al. "A survey on evolutionary reinforcement learning algorithms". In: *Neurocomputing* 556 (2023), p. 126628.

[265] Joshua Zukewich et al. “Consolidating birth-death and death-birth processes in structured populations”. In: *PLoS One* 8.1 (2013), e54639.