European Journal for Philosophy of Science (2023) 13:49
https://doi.org/10.1007/s13194-023-00553-6

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Efficiency and fairness trade-offs in two player bargaining games

David Freeborn[1]

Received: 8 February 2023 / Accepted: 28 September 2023 / Published online: 24 October 2023

Abstract
Recent work on the evolution of social contracts and conventions has often used models of bargaining games, with reinforcement learning. A recent innovation is the requirement that every strategy must be invented either through through learning or reinforcement. However, agents frequently get stuck in highly-reinforced "traps" that prevent them from arriving at outcomes that are efficient or fair to the both players. Agents face a trade-off between exploration and exploitation, i.e. between continuing to invent new strategies and reinforcing strategies that have already become highly reinforced by yielding high rewards. In this paper I systematically study the relationship between rates of invention and the efficiency and fairness of outcomes in two-player, repeated bargaining games. I use a basic reinforcement learning model with invention, and five variations of this model, designed introduce various forms of forgetting, to prioritize more recent reinforcement, or to maintain a higher rate of invention. I use computer simulations to investigate the outcomes of each model. Each models shows qualitative similarities in the relationship between the efficiency and fairness of outcomes, and the relative amount of exploration or exploitation that takes place. Surprisingly, there are often trade-offs between the efficiency and the fairness of the outcomes.

Keywords Game theory · Evolutionary game theory · Reinforcement learning · Social contract theory · Evolution of social conventions · Machine learning · Philosophy of the social sciences · Exploration-exploitation tradeoffs · Social fairness · Pareto efficiency

1 Introduction

Traditionally, evolutionary game theoretic models have assumed that agents select strategies from a pre-defined menu of options. However, this assumption is often left

✉ David Freeborn
david.freeborn@nulondon.ac.uk

[1] Northeastern University, Devon House, 58 St Katharine's Way, London E1W 1LP, UK

unjustified. In a naturalistic model, we should not always assume that agents begin with all of the possible strategies as options. In nature, any strategy must be invented, for example through learning or evolution. To this end, recent game theoretic research has relaxed this assumption, instead allowing agents to "invent" new strategies in signalling games (Alexander et al., 2012), Hawk-Dove games (Herrmann & Skyrms, 2021) and bargaining games (Freeborn, 2022).

In a bargaining game, two players compete over a divisible resource, with each player seeking to maximize their reward. Bargaining games have been widely used to model the evolution of social contracts or conventions (Binmore, 2005, 2014; Skyrms, 2014; O'Connor, 2019; Alexander & Skyrms, 1999; Axtell et al., 2000). Researchers have been particularly interested in better understanding how, and to what extent, such evolved social conventions might be efficient in their allocation of resources and fair to all players. I consider an outcome more efficient if less of the resource is wasted, with a Pareto efficient outcome dividing the entire value of the resource between two players, and a minimally efficient outcome discarding the entire value of the resource. I consider an outcome more fair if it gives a more equal share of the resource to each player, regardless of how much of the resource is wasted. An outcome in which both players received half the resource's value would be maximally fair, as would an outcome in which both players receive nothing.

Recent research (Freeborn, 2022) discusses several learning models that incorporate strategy invention as well as strategy reinforcement. However, Freeborn does not systematically examine the possible trade-offs between efficiency and fairness. Nonetheless, Freeborn finds that fairer outcomes are found to be somewhat favored over over unfair outcomes, and efficient outcomes are found to be somewhat favored over inefficient outcomes; however, there is a fairly wide variation in results. On average, the outcomes usually settle some distance from the fair solution, and some of the resource remained wasted, even after the simulations run for a large number of turns. One possible factor is that the rate of mutation decreases as strategies are reinforced. Relatedly, each strategy can be reinforced without limit, making it harder for agents to find success with alternative strategies as the simulation progressed. If one player, by luck, succeeds in getting a high demand strategy highly reinforced, the other player may be stuck without being able to get high rewards from making similarly high demands. Inefficiencies result because players draw strategies at random: the two players are not able to co-ordinate perfectly. Players frequently fall into "inefficient-unfair traps", in which strategies that lead to somewhat inefficient and unfair outcomes become ever more reinforced. As the rate of mutation falls, it becomes ever harder for the players to escape[1].

These inefficient-unfair traps in some ways resemble "polymorphic pitfalls" seen in finite-population Nash demand games with a finite population, and a finite number of available strategies, under various dynamics (see Skyrms 1994 and Alexander 2008, pages 148–198). In these pitfalls, the players are stuck in a non-perfectly coordinated trap, leading to outcomes that are inefficient and unfair. Introducing small mutation rates, allowing extinct strategies to be reintroduced, can have a variable effect: under imitate-the-best dynamics, populations are almost guaranteed to settle

[1] For similar findings with other learning dynamics, see Sugden (1986) and Skyrms (2014).

on the fair division, whereas under best-response dynamics, mutation can prevent co-ordination,leading to worse outcomes.[2] Freeborn (2022) has already shown that invention with an infinite number of possible strategies can proceed somewhat differently to mutation with a finite number of strategies, and fairer outcomes are somewhat, but not completely favoured. Furthermore, forgetting strategies seems to have an variable effect on the tendency towards a fair division, sometimes either partially favoring or disfavoring it. Such results require a more systematic investigation. Here I investigate in how, and in what ways, the invention and the types and rates of forgetting can affect these outcomes.

Perhaps keeping a higher rate of invention for longer would help players to escape from these inefficient-unfair traps. If both players keep experimenting with new strategies at a sufficiently high rate, then eventually they may discover and reinforce strategies that yield higher average rewards. On the other hand, high rates of invention also carry the risk of wasting part of the resource: whilst experimenting with random strategies, players cannot consistently coordinate. Thus it would be of interest to study the role that rates of invention in bargaining games more systematically. Of particular interest would be to better understand how they can influence the efficiency and fairness of evolved social contracts and conventions.

So there may be possible benefits or detriments to keeping the rate of invention high. Loosely, we might imagine an exploration-exploitation reinforcement learning trade-off (see Sutton and Barto 2015 for an overview).[3] In a loose sense, exploration refers to an agent widely sampling the space of strategies to learn more about which strategies yield high payoffs, whereas exploitation refers to an agent pursuing the strategy that they believe to yield the highest payoffs, based on what they have learned so far. An agent that explores too little may settle in on local maxima whilst missing potentially better strategies elsewhere. An agent that exploits too little does not sufficiently take advantage of what they have learned to receive high payoffs. Intuitively, to receive the highest overall yield, there must be some trade-off between the two approaches. In the context of this paper, exploration involves either inventing new strategies, or testing strategies that have so far received little reinforcement, whereas exploitation amounts to playing the most highly reinforced strategies. The use of the terms here is intentionally quite loose; nonetheless it captures important intuitions. This type of trade-off between exploration and exploitation has not been thoroughly investigated in bargaining games. It would be of great interest to better understand the role of this trade-off in the evolution of social conventions.

Whether or not higher rates of mutation lead to more efficient or fair equilibria, understanding how different rates of mutation may affect the dynamics is relevant to many real world systems. It is important to understand both what happens when the rate of mutation does not fall, and also how changing the rate of mutation affects the

[2] Under imitate-the-best dynamics, each agent adopt the best-performing strategies of those neighbors it can observe. Under best-response dynamics, selects the strategy that would yield them the maximum payoff, given the observed strategies that their neighbors are employing.

[3] Note that the terms *exploration* and *exploitation* have been used in various senses, some stricter, others looser. The terms were originally applied to the context of multi-armed bandit problems (Burnetas & Katehakis, 1997). However, the use of the terms here in this paper is only analogous to the exploration-exploitation trade-off in multi-armed bandit problems.

dynamics.[4] In a changing environment, learning agents might have reason to prioritize more recent knowledge over less recent, for instance through a process of forgetting. Thus it is important to understand how these factors may influence the evolution of social conventions and contracts.

In this paper, I investigate the relationship between rates of invention and the outcomes of two-player bargaining games with reinforcement learning more systematically. I consider a basic model (Section 2) and five variations (Section 3), sample the key variation parameters, and look at the fairness and efficiency of the outcomes. Each variations is designed to alter the amount of forgetting, or similarly, to prioritize more recent reinforcement over less recent reinforcement, to allow the players to keep learning for longer, or indefinitely. Alternatively, we can understand these variations as altering and degree to which agents engage in exploration compared to exploitation. Despite being built upon this common intuition, each variation uses a very different methodology to achieve this. Nonetheless, each of the variations shows qualitatively similar relationships between the fairness and efficiency of the outcomes and the dependency of these outcomes on the exploration-exploitation trade-off. This suggests some universal features of the relationship, which I analyse in Section 4. However, the relationship between the relative amount of exploration, the fairness and the efficiency of outcomes is non-trivial and there are several different regimes of behavior. I discuss some general conclusions in Section 5.

2 Model details

The basic model consists of two players, repeatedly playing a divide-the-dollar game against each other, and learning through Roth-Erev type differential reinforcement (Roth & Erev, 1995, 1998), with each strategy invented.[5] In the divide-the-dollar game, we have two identical agents, labeled players 1 and 2, who each seek to maximize their own share of a resource. Without loss of generality, we can assign the total value of the resource to be unity. Each turn, each player selects a strategy, which involves demanding some fraction of the resource. If the two players' demands sum to less than the resource's total value, each receives their demands as a reward. However, if the demands exceed the value of the resource, the players receive a fixed (typically low-value) payoff, in this case set to zero, representing the failure to come to an agreement.[6]

Each player has a list of strategies, with each strategy having an associated positive real number-valued weight. Players select a strategy each turn with probability proportional to its relative weight. At the end of each turn, each player reinforces

[4] Theoretically, the rate of mutation would fall to zero under some circumstances, for instance in a stable evolutionary environment in which faithful replication is costless. However, analytical and empirical studies (Allen & Rosenbloom, 2012) have shown that positive mutation rates can evolve in novel or fluctuating environments.

[5] The basic model is the same as that used in Freeborn (2022).

[6] In other words, the players have access to agreements in the convex feasibility set, $S \subset \mathbb{R}^2$. If the players agree on a choice within the set, then they receive the corresponding payoffs. Otherwise the players receive the payoffs corresponding to a disagreement point, which is set to (0,0).

the strategy that they chose, by increasing its weight by the quantity of the reward that they received. I also include a "mutator" strategy (which in the basic model has weight 1). If the mutator is chosen, the player "invents" a new strategy, drawing from the real numbers in the interval [0, 1], with uniform probability and adds it to their set of strategies, giving it weight 1, and then plays this strategy, reinforcing that strategy as usual. The mutator itself is not reinforced in the basic model. I assume that each player starts with no strategies other than the mutator: every other strategy must be invented.

2.1 Technical details

At every turn t, each player, $p \in \{1, 2\}$, has an ordered list of strategies, $S^{1,t} = (M^1, s_1^1, \ldots s_n^1)$ and $S^{2,t} = (M^2, s_1^2, \ldots s_m^2)$, with, $W^{1,t} = (w_M^1, w_1^{1,t}, \ldots w_n^{1,t})$ and $W^{2,t} = (w_M^2, w_1^{2,t}, \ldots w_m^{2,t})$, as the corresponding weights, and M is the mutator, $s_j^p \in [0, 1]$ refers to player p's jth strategy of demanding some fraction, s_j^p, of the total resource, and $w_j^{p,t}$ is the associated weight at turn t.

Each turn, each player draws a strategy, with a probability proportional to its weight,

$$P^t(s_j^p) = \frac{w_j^{p,t}}{w_M^p + \sum_{i=1}^n w_i^{p,t}}.$$

If the sum of both players' demands is less than or equal 1, then the players reinforce the strategy they just played by the quantity they demanded. I call this a "successful reinforcement". If the sum of the demands exceeds 1, then no strategies receive any reinforcement. For instance, suppose player p plays strategy s_j^p, with weight $w_j^{p,t}$ at turn t. Then there are two possible outcome, which may result in a different weight at turn $t + 1$. If there is successful reinforcement, then the new weight at turn $t + 1$ is $w_j^{p,t+1} = w_j^{p,t} + s_j^p$. If there is no successful reinforcement, then the weight at turn $t + 1$ is unchanged, $w_j^{p,t+1} = w_j^{p,t}$.

It may help to work through an example. Suppose at turn t, player 1 selects strategy $s_a^1 = 0.2$ (*demand 0.2*), with weight $w_a^{1,t} = 1.0$, whilst player 2 selects strategy $s_a^1 = 0.4$ (*demand 0.4*) with weight $w_b^{2,t} = 2.0$. These demands sum to $0.2 + 0.4 = 0.6 < 1.0$, so the result is a successful reinforcement. The new weights at turn $t + 1$ will be $w_a^{1,t+1} = 1.2$, $w_b^{2,t+1} = 2.4$, so the players are more now likely to choose these strategies again in the future. Alternatively, suppose that at turn t, player 1 selects strategy $s_a^1 = 0.8$ (*demand 0.8*), with weight $w_a^{1,t} = 1.0$, whilst player 2 selects strategy $s_a^1 = 0.4$ (*demand 0.4*) with weight $w_b^{2,t} = 2.0$. These demands sum to $0.8 + 0.4 = 1.2 > 1.0$, so successful reinforcement does not take place. Neither strategy's weight will change: $w_c^{1,t+1} = 1.0$, $w_d^{2,t+1} = 2.0$, so the players are not more likely to select these strategies again after this outcome.

Each turn, each player may draw the mutator, with probability,

$$P^t(M^p) = \frac{w_M^p}{w_M^p + \sum_{i=1}^n w_i^{p,t}}.$$

Then the corresponding player "invents" a new strategy, by drawing from a uniform distribution over all possible demands in the interval [0, 1].[7]

The agents begin with strategies limited to only the mutator, $S^{p,0} = (M)$, $W^{p,0} = (1)$. When a new strategy is invented, the player appends it to their ordered list of strategies and immediately plays this strategy, reinforcing the weight accordingly. So, if at turn t, player p has the set of n strategies, $S_{p,t} = (M^p, s_1^p, \ldots s_n^p)$, with weights $W^{p,t} = (w_M^p, w_1^{p,t}, \ldots s_n^{p,t})$, and draws the mutator strategy, selecting strategy s_{n+1}^p, then the new set of strategies will be $S_{p,t+1} = (M^p, s_1^p, \ldots s_n^p, s_{n+1}^p,)$, with weights $W^{p,t} = (w_M^p, w_1^{p,t}, \ldots w_{n+1}^{p,t})$.

2.2 Efficiency and fairness

We are especially interested in the efficiency and fairness of the outcomes, after some number of turns, τ. I measure the efficiency at turn τ as the proportion of the resource awarded to either player, averaged over all turns so far,

$$\text{Efficiency} = \frac{\sum_{t=1}^{\tau} \text{Reward}^{t,p_1} + \text{Reward}^{t,p_2}}{\tau}. \tag{1}$$

This will be some real number in the interval [0, 1], with higher numbers corresponding to less of the resource being wasted. If the players hit the disagreement point every turn, each receiving zero reward, then the efficiency will be 0. If the players are able to coordinate their strategies perfectly every turn, so that the entire reward is divided between the two players, then the result will be Pareto-optimal, with an efficiency of 1.

I operationalize the fairness at turn τ as the the absolute difference between the two players rewards, averaged over all turns so far,

$$\text{Fairness} = \frac{\sum_{t=1}^{\tau} |\text{Reward}^{t,p_1} - \text{Reward}^{t,p_2}|}{\tau}. \tag{2}$$

This will be a real number in the interval [0, 1], with a higher number corresponding to outcomes that are more fair.

2.3 Inefficient-unfair traps

The dynamics of this basic model were already investigated in Freeborn (2022). As the mutator strategy is not reinforced, so the probability with which the mutator is selected will decrease as the total reinforcement of other strategies increases. So, in this basic model, as new strategies are invented, or existing strategies are reinforced,

[7] This is an idealization: the computer cannot really select from a continuous interval. However, when using a double-precision floating-point number, with a 53-bit significand precision, there are 2^{53} possible numbers in the given range (IEEE Standard for Floating-Point Arithmetic, 2019), vastly greater than the maximal number of strategies that could be invented in this number of turns. So, there is a small probability than an identical strategy could be invented twice. The same situation arises for the model in Freeborn (2022).

the probability that the mutator is drawn will fall. Whilst the rate of invention will begin relatively high, this rate will gradually drop off.

The dynamics somewhat favor outcomes that are more efficient: if the outcome is inefficient then one or more players could stand to gain by making a higher demand, as long as such a strategy has been invented. However, in general, the agents will not perfectly coordinate, and there will be some inefficiency. Fair outcomes are also somewhat favored: if the outcomes are unfair, then one play will receive less reinforcement. If they receive less reinforcement, then they are more likely to experiment with other strategies or to invent new strategies, which can lead to overshooting, causing both players to receive zero reward. However, eventually, even unfair strategies may receive high levels of reinforcement, and the rate of experimentation with other strategies will fall.

As the rate of mutation falls, it is not difficult for the pair of agents to get stuck in inefficient-unfair traps. In these situations, both players play highly reinforced strategies, that nonetheless lead to inefficient and unfair outcomes. In particular, it can be hard for the players to reinforce strategies that might lead to fairer outcomes. The unfairness is familiar from bargaining games without invention, but with finitely many possible strategies (for example, see O'Connor 2019). Suppose that player 1 has highly reinforced a high demand strategy, *demand* x. Then player 2 is likely to receive zero reward for any strategy *demand* y, $y > x$, as it will probably result in overshooting. And if player 2 tends to play strategies *demand* z, $z < x$, then player 1 has no incentive to demand any less than x: such a strategy would result in lower reward for player 1.

Meanwhile the inefficiency is a result of the invention process, in which strategies are drawn at random as real numbers. As such, the two players have zero probability of coordinating their strategies exactly [8].

As I discussed in Section 1, this gives one motivation for studying dynamics in which the rate of mutation remains higher for longer. Does this lead to strategies that are more fair, or more efficient?

3 Model variations and results

Let us look at efficiency and fairness outcomes for simulated runs of the basic model and five variations of this model. In each case, I study a range of parameter values. For each case, I study 10,000 simulation runs, each over 10,000 turns, and take the average values of efficiency and fairness over those 10,000 runs. In this section, I explain the models and present the results. I save analysis until Section 4.

Three of the variations were already studied for some parameter values in Freeborn (2022), forgetting A, forgetting B and Roth-Erev discounting.[9] However, here I sample a range of parameter values, and study how this effects the efficiency and

[8] As noted in footnote 7, for these simulations, the computer algorithm cannot really selected from a continuous interval, so there is a nonzero probability in the simulations. However, the number of possible strategies that the computer can choose is much greater than the maximum number of strategies that can be invented over 10,000 turns, so the probability of the two players coordinating exactly is very small.

[9] The terms forgetting A and B originate with Alexander and Skyrms (1999), whilst Roth-Erev discounting was introduced by Roth and Erev (1995).

fairness of the outcomes. The aim is to understand how the efficiency and fairness of outcomes vary as the relative amount of exploration and exploitation changes over the parameter space. These variations introduce "forgetting" into the dynamics, in which unsuccessful strategies may become rarer or go extinct. Forgetting is likely to be realistic in many evolutionary and learning contexts. It has also been shown to lead to improved learning in many contexts (Barrett & Zollman, 2009; Schreiber, 2001; Roth & Erev, 1995; Alexander et al., 2012).

In particular, Freeborn (2022) finds that forgetting may improve the efficiency or fairness under certain conditions for learning and inventing agents in bargaining games. However, the results varied. Forgetting method A was found to lead to outcomes that were fairer and more efficient than dynamics without forgetting– at least for some parameter values. The forgetting method B was found to lead to outcomes that were of similar efficiency but slightly less fair than the no forgetting case. Roth-Erev discounting led to a trade-off between fairness and efficiency, depending on a choice of parameter values. However, only a limited range of parameter values were studied.

Two of the variations have not been previous studied, constant rates of mutation and bargaining games. These have the effect of keeping a high rate of the mutation for a longer period, or indefinitely.[10]

3.1 Basic model

The results of running the basic model, without any modifications are shown in Table 1. This provides baseline efficiency and fairness values against which each of the variations will be compared.

3.2 Forgetting *A*

For the first variation of the model, I apply forgetting method A. Each turn, forgetting takes place with probability p_f for each player. Then one of that player's strategies is chosen at random, with probability proportional to its weight. The weight assigned to this strategy is reduced by a value r_f. If the strategy's weight is already less than r_f, then the strategy's weight is set to 0. So, on average, the weight of strategies is reduced in proportion to their weight.

Freeborn (2022) finds that this form of forgetting provides a more challenging evolutionary environment, especially for successful strategies. Some strategies may fall in weight or die out because they are forgotten faster than they can be reinforced. This form of forgetting is more punishing of very high and low demand strategies, and less punishing of strategies close to *demand* 0.5.

In general, any strategy to demand less than $p_f \times r_f$ or greater than $1 - p_f \times r_f$ cannot achieve fixation[11] in the long run. To see this, first note that any fixated strategy will reduce by $p_f \times r_f$ on average each turn. Now, let us suppose that strategy *demand*

[10] To a lesser extent, the other forms of forgetting will also have this effect, in particular Roth-Erev discounting, by reducing the total reinforcement for longer; however, this will be to a much lesser extent.

[11] "Fixation" refers to the process by which a particular strategy becomes the sole version present.

Table 1 Average efficiency and fairness for the basic model, over 10,000 simulation runs, each run for 10,000 turns

Efficiency	0.78
Fairness	0.81

$k, k > 1 - p_f \times r_f$ has fixated for player 2. When player 2 plays this strategy, player 1 can receive a maximum reward of $1 - k$ each turn. But any fixated strategy for player 1 will reduce by a greater amount on average each turn, $p_f \times r_f > 1 - k$, so player 1's strategy will be forgotten faster than it is reinforced on average. But if player 1 makes demands of greater than k, then the players will overshoot and neither will receive a reward. So player 1 will not receive consistent reward either.

If $p_f \times r_f \geq 0.5$, then no strategy will be successful for long. All strategies will be forgotten faster than they are reinforced, on average. Most of the time, players will have no reinforced strategies other than drawing the mutator. I will call this regime the "invention without reinforcement" regime.

The results of running the model with forgetting A, are shown in Table 2. We can think of $p_f \times r_f$ as parametrizing the "amount of forgetting" that takes place. This can be most conveniently sampled by varying p_f. I set $r_f = 1$ and vary the value of p_f, to sample values between no forgetting taking place (identical to the basic model), $p_f = 0$ and the invention without reinforcement regime, $p_f = 0.5$ [12].

For small values of the forgetting probability, below around $p_f = 0.24$, the efficiency increases with the value of p_f. Above this value, the efficiency decreases with the probability of forgetting. The fairness decreases very slightly with increasing probabilities of forgetting, until around $p_f = 0.2$. Above this, the fairness increases continually with higher probabilities of forgetting.

3.3 Forgetting *B*

Second, I apply forgetting method B. Each turn, forgetting takes place with probability p_f for each player. If forgetting takes place for a player, then one of that player's strategies is chosen at random, with equal probability for each strategy. The weight assigned to this strategy is reduced by a value r_f. If the strategy's weight is already less than r_f, then the strategy's weight is set to 0. So, on average, the weight of each strategy is reduced in proportion to its weight. So this is similar to forgetting A, but strategies are forgotten with equal probability, rather than in proportion to their weight.

Freeborn (2022) finds that the effects of forgetting B are quite different to forgetting A, because it does not selectively punish more reinforced, or fixated strategies. However, forgetting B lengthens the time that it takes for players to settle on highly reinforced clusters of strategies. As a result, the relative probability of drawing the mutator remains higher for longer.

The effects of varying the p_f parameter by some fixed amount generally has a smaller effect with forgetting B than forgetting A. However, as $p_f \times r_f \geq 0.5$ it becomes harder any strategy to get reinforced consistently faster than it is forgotten.

[12] Note that varying r_f has as a qualitatively similar effect to varying p_f here. For clarity, only results for varying p_f are shown.

Table 2 Average efficiency and fairness for the model with forgetting A, $r_f = 1$, over 10,000 simulation runs, each run for 10,000 turns

p_f	0	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24
Efficiency	0.78	0.79	0.80	0.81	0.82	0.84	0.84	0.85	0.85	0.86	0.87	0.87	0.87
Fairness	0.81	0.81	0.81	0.81	0.81	0.81	0.82	0.82	0.82	0.82	0.83	0.83	0.84
p_f	0.26	0.28	0.30	0.32	0.34	0.36	0.38	0.40	0.42	0.44	0.46	0.48	0.50
Efficiency	0.87	0.87	0.87	0.86	0.85	0.84	0.82	0.68	0.54	0.46	0.44	0.43	0.35
Fairness	0.85	0.86	0.87	0.87	0.88	0.89	0.90	0.95	0.97	0.98	0.99	0.99	1.00

The maximum that any player can receive as a reward is 1 unit per turn. Therefore, once $p_f \times r_f \geq 0.5 = 1$, we enter a "invention without reinforcement". Note that this occurs at a value of $p_f \times r_f \geq 0.5$ twice that as for forgetting A.

The results of running the model with forgetting A, are shown in Table 3. I set $r_f = 1$ and vary the value of p_f, to sample values between no forgetting taking place (identical to the basic model), $p_f = 0$ and the invention without reinforcement regime, $p_f = 1$.

As with forgetting A, increasing the probability of forgetting B initially leads to increases in the efficiencies of the outcomes, until around $p_f = 0.20$, after which the effiency rapidly falls. Small values of p_f initially decrease the fairness of the outcomes up to around $p_f = 0.14$ after which the fairness increases monotonically. However, the values are somewhat different for those of forgetting A: forgetting B is less punishing of high and low demand strategies, for each given value of p_f, the fairness is generally similar for forgetting B than for forgetting A with the same value of p_f.

3.4 Roth-Erev

The next variation includes Roth-Erev discounting (see Roth and Erev, 1995). Unlike forgetting A and B, this is not stochastic; instead we apply a discount factor that reduces the weights of every strategy, each turn. Each weight is multiplied by a factor, $(1 - d_f)$, for some $x \in (0, 1)$. As a strategy is reinforced more, it will be discounted more, in proportion to its weight.

As with forgetting A and B, the results of Roth-Erev discounting were analysed in Freeborn (2022). Like forgetting A, this disfavors more highly reinforced strategies, in proportion to their weight. In effect, Roth-Erev discounting puts an upper limit on the total weight that any given strategy can reach. To see this, suppose that some strategy, i, has weight w_i^t at turn t and earns a reward of r_{success} if played successfully without overshooting, and an expected reward of r_μ each turn. In Roth-Erev discounting, each term the strategy weights will be discounted by $w_i^t \times (1 - d_f)$ each turn. So a strategy canot be reinforced to any higher weight once it reaches weight $w_i^t \times (1 - d_f) = r_{\text{success}}$. Furthermore, a strategy will stop increasing on average after reaching weight $w_i^t \times (1 - d_f) = r_\mu$.

However, note that an agent has no limit to how many strategies they can invent, and in principle can invent strategies arbitrarily close to a previous strategy. As a result, whilst an individual strategy may reach maximal reinforcement, the agent might continue to find success by inventing nearby strategies, which may receive similar reward on average. Thus whilst the maximum reinforcement is capped for any individual strategy[13], the set of all strategies within any finite interval does not have a maximal total reward. However, to invent and reinforce nearby strategies will take some time on average.

[13] Of course, as noted in footnotes 7 and 8, in the computer simulations there is also a positive probability that a precisely identical strategy can be invented many times. Only the reinforcement for each individual instance of that strategy would be capped.

Table 3 Average efficiency and fairness for the model with forgetting B, $r_f = 1$, over 10,000 simulation runs, each run for 10,000 turns

p_f	0	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24
Efficiency	0.78	0.83	0.86	0.87	0.88	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89
Fairness	0.81	0.79	0.78	0.76	0.75	0.74	0.74	0.74	0.75	0.75	0.75	0.76	0.79
p_f	0.26	0.28	0.30	0.32	0.34	0.36	0.38	0.40	0.42	0.44	0.46	0.48	
Efficiency	0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.87	0.87	0.86	0.81	
Fairness	0.80	0.81	0.83	0.84	0.86	0.87	0.89	0.90	0.93	0.95	0.97	0.98	
p_f	0.50	0.52	0.54	0.56	0.58	0.60	0.62	0.64	0.66	0.68	0.70	0.72	0.74
Efficiency	0.67	0.50	0.45	0.42	0.41	0.40	0.39	0.39	0.38	0.38	0.38	0.37	0.37
Fairness	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
p_f	0.76	0.78	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
Efficiency	0.37	0.37	0.37	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.35
Fairness	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Hence, in practice, Roth-Erev discounting serves to keep the total weight of strategies significantly lower, although it does not cap this total weight at any particular finite value. The result will be that the rate of mutation remains higher for longer when Roth-Erev discounting takes place, although it will eventually fall, as long as $d_f < 1$. When $d_f = 1$, we reach invention without reinforcement regime, as no strategy can be successfully reinforced at all, as it will be discounted by its entire weight each turn.

The results of running the model with Roth-Erev discounting, are shown in Table 4. I vary the value of d_f, to sample values between no forgetting taking place (identical to the basic model), $d_f = 0$ and the invention without reinforcement regime, $d_f = 1$.

The general pattern of these results is especially similar to forgetting A: Roth Erev is also more punishing of strategies in proportion to their weight. The efficiency increases with the depreciation rate until around $d_f = 0.004$, after which it falls. The fairness of the outcomes decreases slightly with the depreciation rate for small values, but above around $d_f = 0.001$ the fairness increases instead.

3.5 Constant probability of mutation

The previous forms of forgetting serve to keep the rate of invention higher for longer, although eventually the total weight of strategies will increase, and the probability of drawing the mutator will decrease. Instead, we could try fixing the probability of drawing the mutator at some constant value, m_f, regardless of the weight assigned to the other strategies. The probability of drawing any other strategy is then normalized to the relative weight of the other strategies. Note that this requires that each has at least one non-mutator strategy from the beginning with nonzero weight: I start each with agent the strategy *demand 0*.

Table 4 Average efficiency and fairness for the model with Roth-Erev discounting, over 10,000 simulation runs, each run for 10,000 turns

d_f	0	0.75^{32}	0.75^{31}	0.75^{30}	0.75^{29}	0.75^{28}	0.75^{27}	0.75^{26}	0.75^{25}
Efficiency	0.78	0.79	0.80	0.80	0.80	0.81	0.82	0.84	0.85
Fairness	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
d_f	0.75^{24}	0.75^{23}	0.75^{22}	0.75^{21}	0.75^{20}	0.75^{19}	0.75^{18}	0.75^{17}	0.75^{16}
Efficiency	0.86	0.86	0.88	0.88	0.88	0.87	0.86	0.84	0.80
Fairness	0.80	0.80	0.81	0.81	0.82	0.83	0.84	0.86	0.89
d_f	0.75^{15}	0.75^{14}	0.75^{13}	0.75^{12}	0.75^{11}	0.75^{10}	0.75^{9}	0.75^{8}	0.75^{7}
Efficiency	0.76	0.69	0.48	0.43	0.40	0.39	0.37	0.36	0.35
Fairness	0.94	0.99	0.99	0.99	0.99	099	0.99	0.99	0.99
d_f	0.75^{6}	0.75^{4}	0.75^{3}	0.75^{2}	0.75	1			
Efficiency	0.35	0.35	0.34	0.34	0.34	0.33			
Fairness	0.99	0.99	0.99	0.99	1.00	1.00			

As a result, this variation is a more fundamental alteration of the basic model. In the previous variations, the agents start with no strategies other than drawing the mutator, so the probability of drawing the mutator begins at 1, which then tends to fall as other strategies are reinforced. Here, the rate at which the mutator is drawn will not fall towards zero, the initial rate of drawing the mutator will be lower, except for the case $m_f = 1$. The agents do not begin with a period of rapid experimentation and learning. Of course, the case $m_f = 1$ represents an invention without reinforcement regime. Unlike the previous variations, there is no value of m_f equivalent to the basic model.

The results of running the model with a constant probability of mutation discounting, are shown in Table 5. I vary the value of m_f, to sample values between no invention occurs, $d_f = 0$ and the invention without reinforcement regime, $m_f = 1$.

Here the general pattern still shows some similarities to the previous results. Below around $m_f = 0.005$, increasing the mutation probability leads to higher efficiency outcomes, but above this value increasing the mutation probability leads to lower efficiency outcomes. Below around $m_f = 0.003$, increasing the mutation probability leads to increasing fairness, but above this value increasing the mutation probability leads to decreasing fairness.

3.6 Memory cutoff

One way to think about the previous variations is they systematically discount past reinforcement. In the case of Roth-Erev discounting, each unit of reinforcement degrades by a constant factor. Forgetting A has a similar effect but the forgetting is stochastic. Forgetting B instead degrades the weight equally for each strategy. This has the effect of prioritizing more recent reinforcement, which is likely to be realistic in many evolutionary contexts. In the case of a constant probability of mutation, past reinforcement is effectively discounted because we successively normalize the weights of each non-mutator strategy each turn.

The final variation goes further in prioritizing recent reinforcement over past reinforcement. With a memory cutoff, we discard all reinforcement altogether that is older

Table 5 Average efficiency and fairness for the model with Roth-Erev discounting, over 10,000 simulation runs, each run for 10,000 turns

d_f	0	0.75^{25}	0.75^{24}	0.75^{23}	0.75^{22}	0.75^{21}	0.75^{20}	0.75^{19}	0.75^{18}
Efficiency	0.0	0.62	0.65	0.69	0.72	0.73	0.75	0.76	076
Fairness	1.00	0.72	0.71	0.70	0.70	0.70	0.70	0.71	0.71
d_f	0.75^{17}	0.75^{16}	0.75^{15}	0.75^{14}	0.75^{13}	0.75^{12}	0.75^{11}	0.75^{10}	0.75^{9}
Efficiency	0.76	0.75	0.74	0.70	0.66	0.61	0.56	0.51	0.46
Fairness	0.72	0.74	0.76	0,79	0.81	0.85	0.88	0.81	0.91
d_f	0.75^{8}	0.75^{7}	0.75^{6}	0.75^{5}	0.75^{4}	0.75^{3}	0.75^{2}	0.75	1
Efficiency	0.41	0.38	0.37	36	0.35	0.34	0.34	0.34	0.34
Fairness	0.94	0.97	0.98	0.99	0.99	0.99	0.99	1.00	1.00

than a certain number of turns, t_f. In effect, the agents only remember reinforcement that took place in the last t_f turns. The effects of this will only start after we reach turn t_f: before this point all the reinforcement will be kept. Let τ be the total number of turns (recall in the simulations studied here, $\tau = 10,000$). Clearly, if $t_f = \tau$, then this will be equivalent to the basic model: bothing will be forgotten. On the other hand, $t_f = 0$ represents a regime of invention without reinforcement, as all reinforcement will be immediately discarded.

The results of running the model with a memory cutoff, are shown in Table 6. I vary the value of t_f, to sample values between no cutoff occurs, $t_f = \tau = 10,000$ and the invention without reinforcement regime, $t_f = 0$.

Once again, we see some similarities in the pattern for memory cutoff with the previous results. For memory cutoffs above around $t_f = 320$, shorter memory cutoffs lead to greater efficiency. Below this, shorter memory cutoffs lead to lower efficiency. However, the fairness increases monotonically with shorter memory cutoffs.

4 Analysis: trade-offs between efficiency and fairness

Each variation is meant to capture one plausible mechanism of forgetting, or the prioritization of more recent over less recent reinforcement. It is helpful to keep in mind the informal intuition that each of the variations should generally increase the amount of exploration that takes place relative to the basic model[14], with the consequence that the the relative amount of exploitation will decrease. Therefore, we can look at the effects of each of the variations somewhat holistically. I will refer to small values of the forgetting probability, small values of the Roth-Erev depreciation rate, small fixed mutation rates and *long* memory cutoff values as "small values of the variation parameter", corresponding to the comparatively low rates of exploration (and vice versa for "high values of the variation parameter"). I will refer to small values of the forgetting probability, small values of the Roth-Erev depreciation rate, large fixed mutation rates and *long* memory cutoff values as "small values of the variation parameter", corresponding to the comparatively low rates of exploration.

There are clear similarities in the patterns of efficiency and fairness between all of the variations. I plot the fairness against the efficiency for each variation in Fig. 1.

The basic pattern is qualitatively similar between each of the variations. Moreover, the pattern for forgetting A, forgetting B, Roth-Erev discounting and memory cutoffs is especially similar with the fixed mutation rate requiring a slightly different analysis. So let us start by considering those five most similar variations, and then treat the fixed mutation rate case separately. To better understand the pattern, I divide the relationship into several different regimes, schematized in Fig. 2.

[14] A partial exception here is the fixed mutation rate: the relative increase or decrease in the amount of exploration will depend on the turn number and the chosen mutator probability parameter. So for some values and over some turns, the amount of exploration does not increase on average relative to the basic model.

Table 6 Average efficiency and fairness with the memory cutoff model, over 10,000 simulation runs, each run for 10,000 turns

t_f	0	10	20	40	80	160	320	625	1250	2500	5000	10000
Efficiency	0.33	0.57	0.72	0.78	0.82	0.82	0.82	0.81	0.80	0.79	0.77	0.77
Fairness	1.00	0.98	0.94	0.83	0.82	0.82	0.81	0.81	0.81	0.81	0.81	0.81

Basic model regime

First, let us consider regime in which the variation parameter is effecively switched off. This corresponds to the point at which the probability of forgetting is zero, the Roth Erev depreciation rate is zero, or memory length is set to the full number of turns of the simulation (10,000 turns in this case). In this case, the model is identical to the *basic model* (green point in Fig. 2).

Low exploration regime

Second, is the case in which the variation parameter is allowed to increase above its smallest possible value. This corresponds to small forgetting probabilities, small

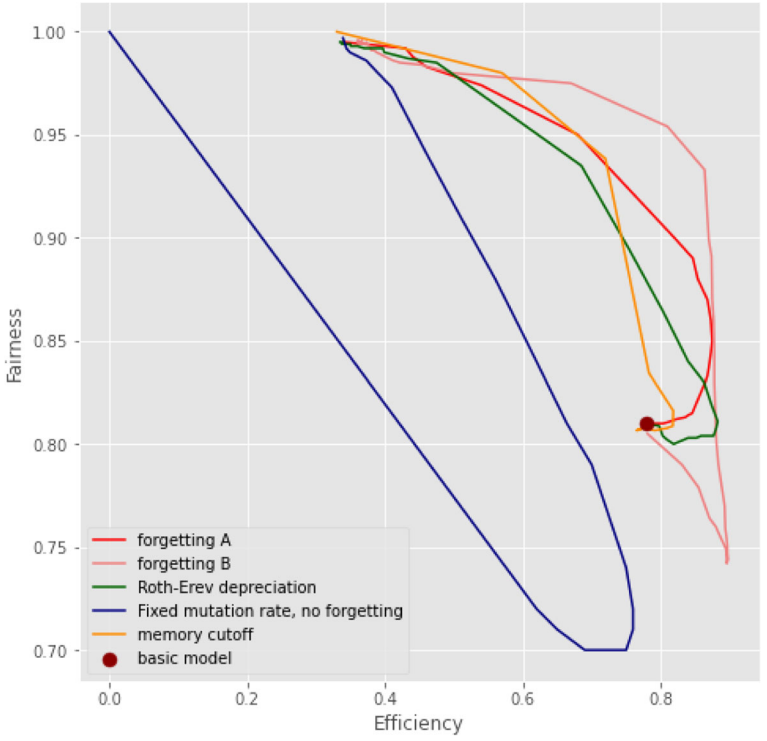

Fig. 1 Average fairness against efficiency for each of the variations described above, over 10,000 simulation runs, each for 10,000 turns

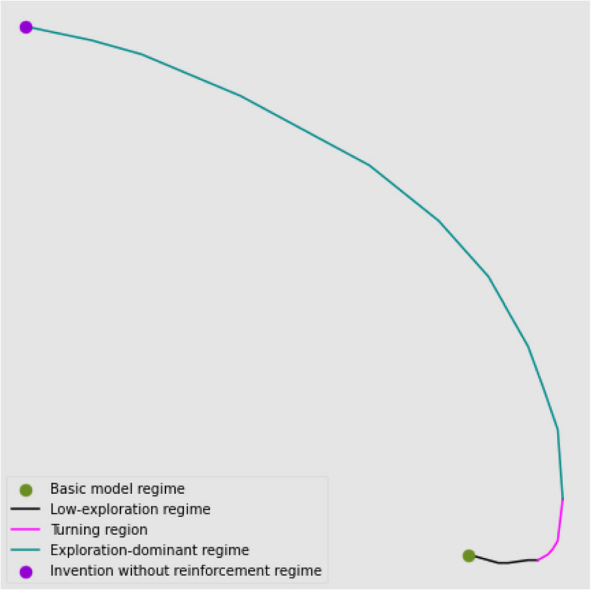

Fig. 2 Schematic illustration of the efficiency-fairness pattern for forgetting A, forgetting B, Roth-Erev discounting and memory cutoffs

Roth-Erev depreciation rates and long memory cutoffs. A relatively low amount of exploration takes place here, but it is a little more than in the basic model. In this regime, as the variation parameter grows, we see an increase in the efficiency of the outcomes, but comparatively less change to the fairness. In this regime, increasing values of the variation parameter help to knock the agents out of the inefficient-unfair traps discussed above, allowing increases in coordination, but does not lead to significant increases in fairness.

Forgetting B stands out here: the fairness falls even as the efficiency increases for values of the forgetting parameter below around $p_f = 0.20$. The reason is that forgetting method B randomly targets agents and treats all strategies equally. In the highly unequal cases, one or other agent was unlucky: Once the forgetting parameter begins to rise higher, it is less likely that bad outcomes will specifically target one agent or the other. Forgetting B caused one agent to forget high demand strategies before they were highly reinforced. As a result, the opposing agent had no high demand strategies, and the other agent received high rewards for their high demand strategies. A highly unequal outcome ensued due to the stochastic nature of forgetting B. Once the forgetting parameter begins to rise higher, it is less likely that bad outcomes will specifically target one agent or the other.

Why does fairness not clearly correlate with efficiency in this regime for the other variations? Forgetting method A is stochastic like forgetting method B, but any bad luck dealt to one agent is likely to be compensated for the general increase in fairness discussed in Section 3.2 above. The Roth-Erev depreciation affects all agents in the same way without a random component: we do not see the corresponding rise in unequal outcomes with this variation. Likewise, the memory cutoff does not lead to a

rise in unequal outcomes, as this affects both agents the same way, and furthermore, for high values will only begin to take affect later during the simulation run, once strategies are already highly reinforced.

Turning region

Next is the turning region. Surprisingly, this is the only region in which the efficiency and fairness are clearly positively correlated. This regime is best understood as an intermediate region between the low exploration regime and the exploration-dominant regime. In this regime, the fairness begins to monotonically increase as we increase the variation parameter (see below), but the efficiency continues to increase, for the same reasons as in the low exploration regime. So there is a small region of parameter space in which the efficiency and fairness increase together.

Exploration-dominant regime

Next is the regime in the variation parameter is large enough that exploration dominates over exploitation. The efficiency falls and the fairness rises as the variation parameters approach their maximum values. When agents spend more of their time exploring, the results are inherently fair: both agents are equally likely to try out new strategies, which will not benefit either agent on average. However, this fairness is not especially useful to either agent. It does not arise from agents settling on a mutually beneficial social contract, but rather from continually trying out new strategies at random. The agents spend less time exploiting successful strategies, resulting in lower efficiency.

Invention without reinforcement regime

Finally, let us consider the regime in which the variation parameter reaches its maximum value: the *invention without reinforcement* regime (purple point). In this regime, only exploration happens with no exploitation: the agents draw random strategies every turn, and no learning takes place at all. This corresponds to the regime in the forgetting probability or depreciation rate is high enough, or memory cutoff is zero, so that any reinforced strategy is instantly forgotten.

In this regime, we see an average efficiency of around $\frac{1}{3}$ and a fairness of 1. To see why, consider both players drawing a strategy at random from a uniform distribution over the interval (0, 1). They have a $\frac{1}{2}$ chance of overshooting, resulting in a reward of 0 for both players. If the players do not overshoot, then the probability density for each player's demand will be given by $f(x) = 2(1 - x)$. Hence, the expected reward is given by $2\int_0^1 x(1-x)dx = \frac{1}{3}$. So the overall expected reward for both players is $\frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$. Thus the efficiency, the total quantity of the reward earned by the two players, will be given by $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. The fairness will be 1 on average because neither player can gain an advantage over the other in the long run: both players can only draw at random from the same distribution.

Fixed memory cutoff

The fixed memory cutoff variation requires a partly separate treatment. I have divided this result into regimes in several Fig. 3. When the fixed mutation rate is set to zero, this does not correspond to the basic model, but rather to a **no learning regime**, in which both players can only play "demand zero" against each other. In this case, the agents will both receive no reward, representing a result that is wholly inefficient, but completely unfair.[15]

As we increase the fixed mutation rate to small values above zero, we arrive at the **low exploration regime**. Here, the efficiency rapidly increases, but the fairness rapidly falls. The increasing in efficiency is unsurprising: the agents now have a chance to invent better strategies and to coordinate. However, very small probabilities of mutation are likely to particularly favor one player or the other, especially if only a few strategies are invented over the course of the 10,000 turns. This accounts for the falling rate of fairness as the mutation probability increases.

Once the mutation rate is high enough, it is less likely to favor one player or the other. The pattern seen with the fixed mutation rate then looks more similar to the other variations. The remaining three regimes, the **turning region**, **exploration-dominant regime** and **invention without reinforcement regime** are completely analogous to the other model variations. As exploration increases, initially efficiency and fairness rise together. Once exploration dominates, the efficiency falls and the fairness increases, until the dynamics approach the invention without reinforcement when the probability of mutation reaches 1. However there are two things to note. First, there is no point which corresponds to the basic model case: the probability of mutation here is fixed, whereas in the basic model it varies as strategies are reinforced. Second, the curve for the fixed mutation rate is generally below and to the left of the others. In general, keeping a fixed, non-decreasing mutation rate throughout, leads to less efficient and less fair outcomes than having an initially high mutation rate that then decreases.

Thus we have seen that there are some general common trends in the relations seen between efficiency and fairness for all of the dynamics studied here. Most obviously, all dynamics show common behavior in the exploration-dominant regime. Likewise, all dynamics exhibit some turning region, where efficiency and fairness rise together as exploration increases. However, we have seen that in the low-exploration regime, the idiosyncrasies of particular models can matter more, shaping different relationships between efficiency and fairness.

5 General lessons

Of course, the most relevant regions of parameter and model space are likely to depend on the real-world evolutionary or learning systems under consideration. We cannot talk about whether efficiency and fairness are positively or negatively correlated across

[15] However, note that this result is especially sensitive to the arbitrary choice of starting strategies. For many other reasonable starting strategies, such as giving the agents a choice of *demand 1* and *demand 0* with equal weights, the outcome would look similar, but this would not be the case with all such strategies.

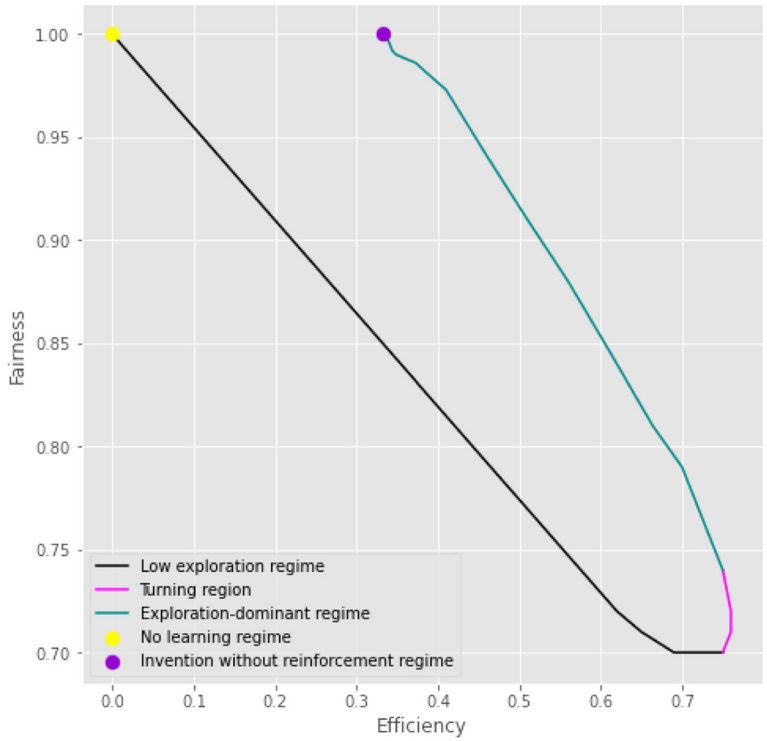

Fig. 3 Average efficiency and fairness for the model with fixed memory cutoff, with the results divided into different regimes

most of the parameter space before first putting some measure over that parameter space. Rather, these results draw attention to some general principles about the evolution of social conventions. In general terms, the relative rates of exploration and exploitation may have varying effects on the efficiency and fairness of outcomes. Furthermore, the efficiency and fairness may be either positively or negatively correlated. These results depend on both the models used and the regions of parameter space.

Nonetheless, the general qualitative pattern is mostly robust across the all of the models. When the rate of exploration is small, increasing exploration generally increases the efficiency, and the fairness may increase, decrease or remain the same. Here, exploration helps to nudge the agents out of inefficient-unfair traps, increasing co-ordination, but the effects on the fairness of the outcomes depends on the choice of model. When the rate of exploration is intermediate, increasing exploration further can increase fairness and efficiency together. However, once exploration comes to dominate, increasing exploration further will increase fairness at the expense of efficiency. The fairness that we achieve is due to continual exploration by both players, but they do not learn to exploit the strategies that they invent.

It may seem somewhat surprising that efficiency and fairness only positively correlate in the intermediate region. Intuitively, highly unfair outcomes should make co-ordination more difficult, because one player receives much lower reward. If one

of the two players learns to co-ordinate slowly, this will decrease the rate at which the other player learns as well. For example, agents will never learn to co-ordinate in a situation where one agent receives the entire reward, because the other agent will receive no reinforcement at all for a demand 0 strategy.

However, in the region where exploration dominates, the high fairness is mainly being driven by continual random invention, not by the agents learning to adopt a fair strategy. Increasing exploration will increase the fairness, because the random invention process is the same for both agents, but it will decrease the efficiency because the agents never exploit what they learn to receive higher rewards. In the low exploration region, fairness and efficiency are negatively correlated in some of the models for a different reason. Forgetting B and fixed mutation rates lead to highly stochastic exploration that relies on random probabilities. This increases the efficiency above the basic model, but when the probabilities are very low, they are likely to favor one or other agent, leading to a low fairness in the outcomes. Increasing fairness through exploration can often achieve both higher efficiencies as well as higher fairness than the lower exploration regimes.

Recall that Alexander (2008) studies finite strategy, finite population Nash demand games under various dynamics. The finite populations were placed on a lattice, in which agents could observe their neighbors' strategies. Alexander introduces small rates of mutation so that previously extinct strategies can be reintroduced. Mutation can function in a qualitatively similar way to invention and forgetting, by introducing new strategies and providing some probability that a strategy gets played away from the most dominant strategy.

It is worth noting a number of comparisons with the discussions in Alexander (2008). First, Alexander observes that mutation can sometimes drive the population either towards or away from the fair outcomes. However, these differences are driven by a qualitatively different mechanism from that observed here. For example, under imitate-the-best dynamics, small mutation rates can allow islands of fair-division play to emerge, and eventually dominate, as agents observe and adopt this successful strategy. However, under best-response dynamics, mutation can have the opposite effect, destroying islands of fair division and leading to dominance by an unfair division. The differences between imitate-the-best and best-response dynamics, are driven by neighbor-neighbor interactions, in particular along edges of regions where one strategy is dominant. There is no analogy to these edge effects here. Rather, it is only changes in the mutation rate that drive the agents towards or from fairer outcomes.

Second, in this study, we can identify a number of different regimes, including the low exploration regime, in which exploration can increase efficiency without any corresponding increase in fairness. This is driven, by exploration benefiting one agent at the expense of the other, and lacks a direct analogue in dynamics where agents copy the strategies of their neighbors. Indeed, only in the turning region, just one part of the parameter space, do fairness and efficiency both grow together, as Alexander observes in imitate-the-best dynamics.

Third, Alexander suggests that given sufficiently high rates of mutation, the "mutational noise" will prevent all agents co-ordinating. This is in fact similar to the effects observed in the exploration-dominant regime in this study, in which agents underutilize the good strategies they have already found. However, here we are able to study

the effects of high rates of mutation more systematically. It is notable that this noisy region covers a significant portion of the parameter space.

This study of the trade-offs between efficiency and exploration, and their relationship to forgetting, exploration and exploitation takes us a step further in our understanding of the evolution of social contracts and conventions. In the light of this, it is natural to ask what these results can teach us more broadly. These results corroborate the idea that fairness is one plausible outcome of (biological or cultural) evolutionary dynamics. However, they also illustrate how contingent such fair social divisions might be, favored in some, but by no means all, regimes. Indeed, whether outcomes such as fairness are achieved may be sensitive to factors such as the degree and types of experimentation or mutation that take place. We have seen that low rates of invention may lead to outcomes that favor some agents (the early adopters of successful new strategies) over others. On the other hand, high rates of invention lead to inefficiencies and lack of co-ordination. Therefore, models of this kind can only explain how intuitions of fairness specifically might have evolved if evolutionary dynamics can be well-represented by the turning regime.[16] However, models such as those studied here might offer a tentative step to better understanding the range of disparate social outcomes attitudes towards fairness seen in real world societies.

There are several directions in which this research could be naturally extended, beyond the scope of this paper. First, the results here are primarily based on simulations, rather than analytic results. Although the simulations have been studied in detail, and the full parameter space sampled, it would be interesting to see whether analytic results can be obtained here.

Second, only repeated games between two agents have been studied. However, it would be natural to extend this to bargaining games in which agents are sampled from larger populations. For example, Freeborn (2022) presents a model in which agents are randomly selected from a larger, but finite population. In such a population, we might expect outcomes to be more fair, but less efficient as agents do not always face the same competitor. It would be interesting to see the extent to which we see the same qualitative patterns can be found when the amount of exploration and exploitation is varied.

The work here explored only one dynamical reinforcement learning model. Whilst such a model provides a natural method for incorporating the invention of new strategies, it would be of interest to consider other dynamics such as fictitious play. In particular, it would be interesting to know whether the same qualitative patterns are robust in other learning dynamics. Likewise, it would be of interest to apply these learning models to other bargaining games than divide-the-dollar, with asymmetric payoff structures. Furthermore, the exploration-exploitation trade-off could also be of great interest in other game theoretic contexts where learning takes place. Signalling games provide one obvious and potentially fruitful context for further investigation.

[16] In the context of biological evolution, rates of mutation are generally low, so a turning regime could provide a plausible model. However, this will not necessarily hold in the contexts of social learning and cultural evolution, where rates of invention could be high or variable.

Declarations

Conflicts of interest The author has no conflicting interests.

References

Alexander, J. M. (2008). *The Structural Evolution of Morality*. Cambridge University Press, Cambridge (1st ed.)

Alexander, J. M., & Skyrms, B. (1999). Bargaining with neighbors: Is justice contagious? *The Journal of Philosophy, 96*(11), 588.

Alexander, J. M., Skyrms, B., & Zabell, S. (2012). Inventing new signals. *Dynamic Games and Applications, 2*(1), 129–145.

Allen, B., & Rosenbloom, D. I. S. (2012). Mutation rate evolution in replicator dynamics. *Bulletin of Mathematical Biology, 74*, 2650–2675.

Axtell, R., Epstein, J., & Young, H. (2000). The emergence of classes in a multi-agent bargaining model. *Generative Social Science: Studies in Agent-Based Computational Modeling*

Barrett, J., & Zollman, K. (2009). The role of forgetting in the evolution and learning of language. *Journal of Experimental and Theoretical Artificial Intelligence, 21*, 293–309.

Binmore, K. (2005). *Natural Justice*. Oxford University Press.

Binmore, K. (2014). Bargaining and fairness. *Proceedings of the National Academy of Sciences, 111*(Supplement 3), 10785–10788.

Burnetas, A. N., & Katehakis, M. N. (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research, 22*(1), 222–255.

Freeborn, D. (2022). The invention of new strategies in bargaining games. *Philosophy of Science*. (in press)

Herrmann, D. A. & Skyrms, B. (2021). Invention and evolution of correlated conventions. *The British Journal for the Philosophy of Science*. (in press)

IEEE Standard for Floating-Point Arithmetic (2019). Ieee std. *754-2019 (Revision of IEEE 754-2008)*, pp. 1–84

O'Connor, C. (2019). *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford University Press

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, 8*(1), 164–212.

Roth, A. E., & Erev, I. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review, 88*(4), 848–881.

Schreiber, S. (2001). Urn models, replicator processes, and random genetic drift. *SIAM Journal of Applied Mathematics, 61*, 2148–2167.

Skyrms, B. (2014). *Evolution of the Social Contract*. Cambridge University Press, (2nd ed.)

Skyrms, B. (1994). Sex and justice. *The Journal of Philosophy, 91*(6), 305–320.

Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Oxford, UK: Basil Blackwell.

Sutton, R. S. & Barto, A. G. (2015). *Reinforcement Learning: An Introduction*. A Bradford Book. The MIT Press, (2nd ed.)