Malthusian Reinforcement Learning

Joel Z. Leibo
DeepMind
London, UK
jzl@google.com

Julien Perolat
DeepMind
London, UK
perolat@google.com

Edward Hughes
DeepMind
London, UK
edwardhughes@google.com

Steven Wheelwright
DeepMind
London, UK
sjwheel@google.com

Adam H. Marblestone
DeepMind
London, UK
amarblestone@google.com

Edgar Duéñez-Guzmán
DeepMind
London, UK
duenez@google.com

Peter Sunehag
DeepMind
London, UK
sunehag@google.com

Iain Dunning
DeepMind
London, UK
idunning@google.com

Thore Graepel
DeepMind
London, UK
thore@google.com

ABSTRACT

Here we explore a new algorithmic framework for multi-agent reinforcement learning, called Malthusian reinforcement learning, which extends self-play to include fitness-linked population size dynamics that drive ongoing innovation. In Malthusian RL, increases in a subpopulation's average return drive subsequent increases in its size, just as Thomas Malthus argued in 1798 was the relationship between preindustrial income levels and population growth [24]. Malthusian reinforcement learning harnesses the competitive pressures arising from growing and shrinking population size to drive agents to explore regions of state and policy spaces that they could not otherwise reach. Furthermore, in environments where there are potential gains from specialization and division of labor, we show that Malthusian reinforcement learning is better positioned to take advantage of such synergies than algorithms based on self-play.

KEYWORDS

Intrinsic motivation; Adaptive radiation; Demography; Evolution; Artificial general intelligence

ACM Reference Format:
Joel Z. Leibo, Julien Perolat, Edward Hughes, Steven Wheelwright, Adam H. Marblestone, Edgar Duéñez-Guzmán, Peter Sunehag, Iain Dunning, and Thore Graepel. 2019. Malthusian Reinforcement Learning. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Reinforcement learning algorithms have considerable difficulty avoiding local optima and continually exploring large state and policy spaces. This is known as the problem of exploration. In single-agent reinforcement learning, the main approach is to rely on intrinsic motivations, e.g., for individual curiosity [2, 6, 25, 29, 30, 33], empowerment [22], or social influence [19].

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

However, critical adaptation events in human history are difficult to explain with intrinsic motivations. Consider the dispersal of homo sapiens out of Africa, where they first evolved between 200, 000 and 150, 000 years ago, throughout the globe, eventually occupying essentially all terrestrial climatic conditions and habitats by 14, 500 years before the present [13]. This example is relevant to AI research because intelligence is often defined as an ability to adapt to a diverse set of environments [23]. Essentially no other process on earth has led so quickly to so much adaptive diversity as did the expansion of human foragers throughout the globe[1]. Human foraging communities were capable both of discovering water-finding strategies suitable for the arid Australian desert as well as how to hunt seals hiding beneath ice sheets and keep warm in the Arctic [4]. From this perspective, the great dispersal of human foragers can be seen as some of the best evidence for an "existence proof" that intelligence, by this definition, is even possible. Yet there's no evidence that intrinsic motivations like curiosity played a role in it. Rather, considerable evidence points to a variety of extrinsic motivation mechanisms as the main drivers of human migration including climate change [11, 37] and demographic expansion [26, 31].

One perspective on the problem of exploration is that the difficulty comes from the sparseness of extrinsic rewards. If extrinsic rewards are very sparse, then it is hard to estimate state-value functions and policy gradients since returns will have very high variance. Thus intrinsic motivation methods produce frequent intermediate (dense) rewards in hopes of bridging the long gaps between extrinsic rewards [7]. Multi-agent reinforcement learning offers an alternative. Algorithms based on self-play like AlphaGo [1, 18, 34, 35, 38] are aimed at an essentially single-agent objective, e.g., defeat a specific human grandmaster. Directly training by single-agent reinforcement learning to accomplish that objective

[1]The rapid dispersal of homo sapiens across the globe and their adaptation to the full range of diverse terrestrial habitats can be regarded as a great feat of intelligence. We may even assess its universal intelligence Υ by adapting the following definition from [23], $\Upsilon(\pi) = \sum_{h \in \mathcal{H}} 2^{-K(h)} V_h^\pi$, where π is a policy, H is the set of terrestrial habitats, V_h^π is the expected value of the sum of rewards from inhabiting the habitat h, and $K(h)$ is a measure of the complexity of h. The complexity measure could be defined as any sensible ecological distance metric measuring the distance from the ancestrally adapted environment to h.

is a lost cause. Since the untrained agent would never win a game, it would never get any reward signal to learn from. Self-play, on the other hand, provides an alternative incentive for agents to explore deeply through the strategy space. In two-player zero-sum, it escapes local optima by learning to exploit them, diminishing the returns available from such strategies, and thereby extrinsically motivating new exploration. Of course, life is not a zero-sum game. Nevertheless, real-life feats of exploration like the dispersal of homo sapiens out of Africa really were motivated in part by a pressure to out-compete rivals in a struggle for scarce resources that became increasingly difficult over time due to demographic expansion. Moreover, local competition between individuals is a universal feature of natural habitats, and underlies the evolution of dispersal [20]. In this paper we investigate whether such population size dynamics can be exploited as an algorithmic mechanism in multi-agent reinforcement learning.

As an algorithmic mechanism, augmenting multi-agent reinforcement learning with population size dynamics appears to have the requisite property needed to evade local optima and traverse large state spaces. Strategies that work well at low densities do not necessarily translate well to high densities, but success at any density ensures density will increase further in the future. The rules of the game therefore naturally shift over time in a manner that depends on past outcomes. This ensures that species cannot remain too long in comfortable local optima. When resources are scarce, rising populations eventually dissipate gains from learning, forcing agents to innovate just to maintain existing reward levels [8][2].

So far we've motivated introducing population dynamics to multi-agent reinforcement learning by appealing to a competitive struggle for existence against a well-matched foe, i.e., the same argument underlying the performance of self-play in two-player zero-sum games. However, there is more to life than competition. As an algorithmic mechanism to promote learning in general-sum multi-agent environments, population dynamics may also be more suitable than self-play-based approaches. In particular, we will consider whether this approach provides greater scope for adapting to synergies between specialists, making it easier to discover joint policies involving significant division of labor.

In this work we introduce a new algorithm for multi-agent reinforcement learning based on these principles of population dynamics. It is called Malthusian reinforcement learning because improvements in returns for any subpopulation translate directly into increases in the size of that subpopulation in subsequent episodes. Thus it may be evaluated on either the individual or the group level. In this work we are interested in two specific questions:

(1) Is Malthusian reinforcement learning better at avoiding becoming stuck in bad local optima in *individual* policy space than competing algorithms based on intrinsic motivation?
(2) Is it easier to evolve joint policies to implement heterogeneous mutualism behaviors with Malthusian reinforcement learning than with alternative approaches based on self-play?

[2] [8] argues that preindustrial human populations generally oscillated around a fixed, and only very slowly increasing, carrying capacity until the industrial revolution. Similar oscillations in subpopulation sizes were recently observed in a large-scale multi-agent learning simulation by [41]. As those authors pointed out, its possible for population dynamics to endlessly oscillate rather than increasing over time. The same is true for the strategies used by learning algorithms based on self-play. One fix that was used in AlphaGo and elsewhere is to require agents to learn to defeat all previous versions of themselves, not just the most recent [34, 35]. This prevents self-play-based agents from endlessly learning and forgetting the same exploit and defense.

2 MODEL

2.1 Characteristics of the Malthusian reinforcement learning framework

Malthusian reinforcement learning differs from standard multi-agent reinforcement paradigms in a number of ways.

(1) Malthusian reinforcement learning may be seen as an algorithm for "community coevolution". It produces a set of communities, called *islands* in our terminology. Each island has a set of agents implementing policies that should, if the training was successful, function well together.
(2) Each individual is a member of a *species*. All individuals of the same species share a policy neural network.
(3) The algorithm unfolds on two timescales corresponding to (A) the population dynamics (ecological) time, and (B) policy execution (behavioral) time.
(4) The population dynamics are linked to individual reinforcement learning returns. If individuals of a given species perform well on a particular island then their population will increase there in the future.
(5) During each episode all individuals of a given species generate experience to train a common neural network via v-trace [12]. Experiences generated by individuals of a particular species are used only to update their own species neural network. After each episode the distribution over islands of each species is updated by a policy gradient-like update rule.
(6) *Conservation of compute:* Biologically realistic population dynamics all contain at least the possibility of exponential growth. In practice, they are limited by carrying capacities, i.e., by environment properties. Since Malthusian reinforcement learning is mainly intended for multi-agent machine learning applications rather than for ecological simulations we cannot rely on resource constraints in the environment to limit growth. Thus, to ensure it can be executed with bounded compute resources, the population dynamic works by maintaining probability distributions for each species over the set of all islands. The probability assigned to any given island may grow or shrink based on the individual returns achieved there, but it is always constrained to be a valid probability distribution. Compute remains bounded because a fixed number of samples are used to assign individuals to islands. The total population varies on any given island from episode to episode, but across the entire *archipelago*, the number of individuals is always constant.

2.2 Archipelagos, islands, and species

Island and Archipelago. an *island* is a multi-agent environment where a variable number of agents can interact. An *archipelago* is a set I of islands. Furthermore we will write $N_I = |I|$ the number of islands in an archipelago.

Archipelago	
I	the set of islands in the archipelago
N_I	the number of islands in an archipelago
i	indexes islands
e	indexes ecological scale time
Species	
L	the total number of species
Ψ_l	a species
π^l	the policy network of a species
θ^l	the parameters of π^l
μ^l	the distribution of species l over the archipelago
w^l	the parameters of μ^l
Δ_{N_I}	the set of distributions over the archipelago
K	total number of individuals
$M = \frac{K}{L}$	number of species l individuals across all islands
k^l	labels individual k of species l
Population	
$\Psi^l_{i,e}$	the set of individuals of species l allocated to island i at time-step e
$\phi^{l,i}_e$	the average fitness received by species l on island i at time e
$\phi_{k^l,e}$	the fitness received by individual k^l of species l at time e
η	population entropy regularization weight
α	population adaptation rate
Island	
t	indexes behavioral scale time
N	the number of agents
S	the state space
s	a state
A^i	the action space of player i
a^i	an action of player i
o^i	the observation of player i
$\psi^i(\cdot)$	the function that maps s to the observation o^i of player i
$p(s_{t+1}\|s_t, a^1, \ldots, a^N)$	is the transition kernel
$r^i(s, a^1, \ldots, a^N)$	the reward of player i

Species. A species Ψ^l is a set of individuals sharing the same policy network parameterization. There are L species indexed by l. Each species is composed of a policy network π^l with parameters θ^l which encodes the behavior of each individual of the species. The distribution of agents of a given species l over the islands is $\mu^l \in \Delta_{N_I}$ (where Δ_{N_I} is the set of distributions over islands. μ^l is defined as a softmax over weights w^l. The total number of individuals is K and the number of individuals per species is $M = \frac{K}{L}$. We will denote each individual of a species l by k^l.

The learning process unfolds over two timescales, a slow *ecological* scale which adapts the distribution of species over islands μ^l and a fast *behavioral* scale over which individuals execute their policies. Species adapt π^l to behave in the presence of others at the level of the island. The ecological scale timesteps are indexed by e, and the behavioral scale timesteps are indexed by t. The ecological scale ticks at the level of single episodes for the behavioral scale.

2.3 Population dynamics

The population dynamics govern how individuals of each species are assigned to the different islands over the ecological time scale. At a fixed ecological timestep e, individuals of each species are assigned to islands by sampling M times from the distributions μ^l. For each island, this yields an allocation $\Psi^l_{i,e}$, the set of individuals from species l playing on island i at ecological timestep e.

Each island has its own environment, in general the islands could have different environments from one another, though in this work we only consider the case where they are all the same.

Over the course of ecological time, the population evolves according to a gradient-based dynamic. At each timestep e, each individual k^l of each species l receives a fitness $\phi_{k^l,e}$, which is exactly its cumulative reward over the behavioral scale timesteps that have elapsed during one step of the ecological timescale. The per-island fitness for each species is then calculated as

$$\phi^l_{i,e} = \left(\sum_{k^l \in \Psi^l_{i,e}} \phi_{k^l,e} \right) / |\Psi^l_{i,e}| \text{ and 0 if } \Psi^l_{i,e} = \emptyset \,.$$

The distribution over islands for each species, $\mu^l(i) = e^{w^l_i} / \sum_j e^{w^l_j}$, is updated according to policy gradient with entropy regularization. Explicitly the distribution weights for species l over all islands change according to a policy-gradient update

$$w^l_{e+1} = w^l_e + \alpha \left[\sum_{i \in \{1,\ldots,N_I\}} \nabla_{w^l} \mu^l(i) (\phi^l_{i,e} - \eta \log \mu^l(i)) \right] .$$

The goal of the entropy regularization term is to enforce that some minimal population of each species remains on sub-optimal islands. Thus, the population distributions adapt over ecological time so as to minimize the following loss:

$$\left[\sum_{i \in \{1,\ldots,N_I\}} \mu^l(i) (\phi^l_{i,e} - \eta \log \mu^l(i)) \right] .$$

2.4 Multiagent Reinforcement Learning

A Partially Observable Markov Game (POMG) is sequential decision model of a multiagent environment in which N individuals interact. At each state $s \in S$ of a POMG, each agent selects an action $a^i \in A^i$ based on the observation o^i of the state of the game they have. The observation of player i is defined here as a function of the state $o^i = \psi^i(s)$. Then the state changes to $s' \sim p(.|s, a^1, \ldots, a^N)$ and the individuals receive reward $r^i(s, a^1, \ldots, a^N)$. Each species learns a policy $\pi^l(a^i|o^i)$ given the experience of each of its individuals. At each step, all individuals of a species collect trajectories of the experience gathered in the island they have been assigned to. The reinforcement learning algorithm produces gradient updates of the parameters for each individual of the species. The gradient updates are then averaged over all individuals of the species to update the parameters $\{\theta_l\}_{\{1,\ldots,L\}}$. The V-trace algorithm is used to update the parameters as described in [12] with truncation levels set to 1.

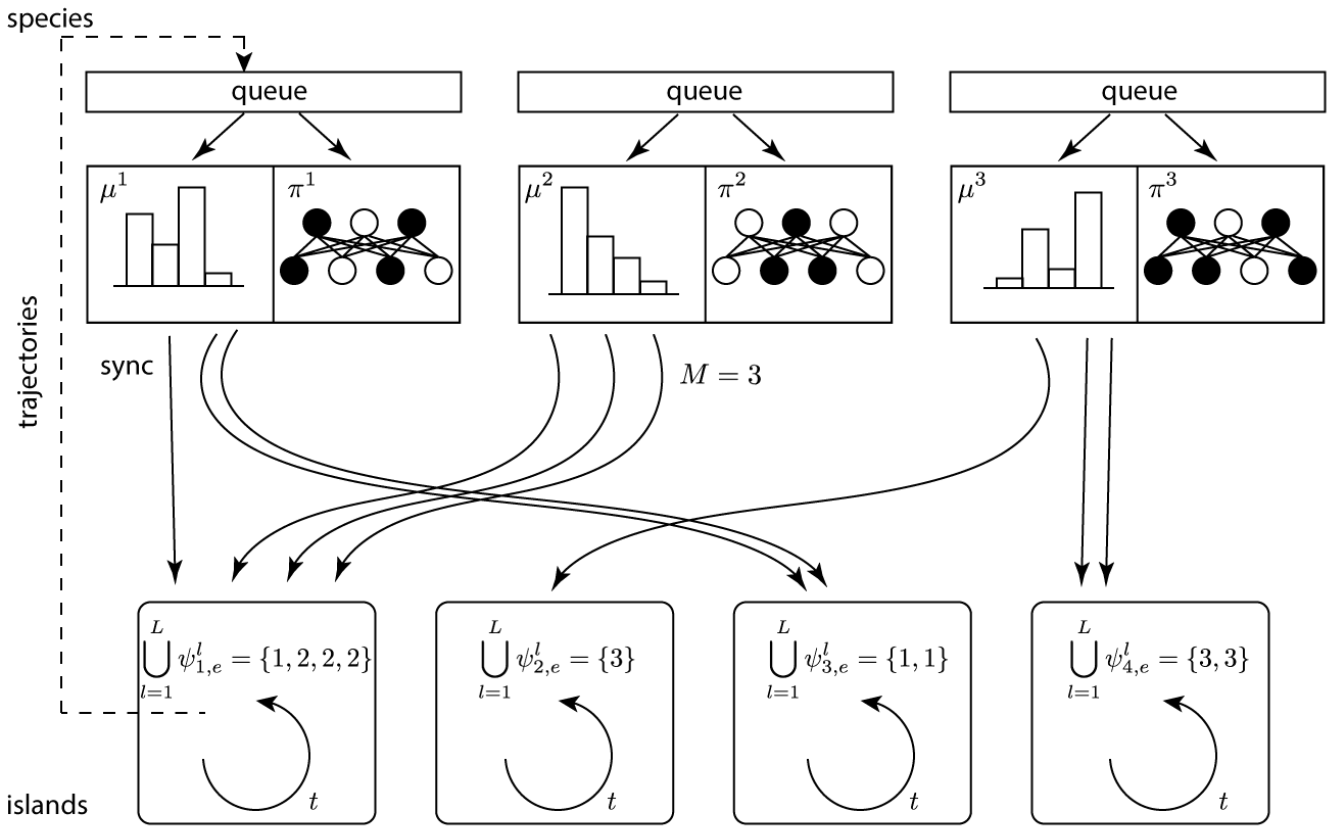

Figure 1: At each iteration on the ecological timescale e, each island samples the players to participate in its next episode according to the probability distributions μ^l over islands maintained by each species. Experienced trajectories from all conspecifics, on all islands, are used to update the same species policy network π^l. The distributions of returns to each species over islands are used to update the distributions from which to sample players for the next episode.

Note that experience (observations, actions and rewards) from all individuals of a species contribute equally, but that the individuals may be spread non-uniformly over islands. This means that the species parameter update may be disproportionately affected by the performance of the species on particular islands.

RL Agent	
LSTM Unroll Length:	20
Entropy Regularizer:	$\sim$ log-uniform$(0.00005, 0.05)$
Baseline loss scaling:	0.5
Discount:	0.99
Optimization	
RMSProp learning rate:	$\sim$ log-uniform$(0.0001, 0.005)$
RMSProp ϵ:	0.0001
RMSProp decay:	0.99
Batch size:	32

Function approximation: The neural network architecture was similar to that of [27]. It consists of a convnet with 16 channels, kernel size of 3, and stride of 1. The output of the convnet is passed to a a 1-layer MLP of size 32, followed by a recurrent module (an LSTM [17]) of size 64. The recurrent module's output is then linearly transformed into the the policy and value. All nonlinearities between layers were rectified linear units.

Distributed computing: The island simulation and the species neural network updates were implemented as separate processes, potentially running on different machines. Islands produce trajectories and send them to a circular queue on the species update process. The species update process waits until it can dequeue a complete batch of 32 trajectories, at which point it computes the v-trace update.

Environments: The games studied in this work are all partially observable in that individuals can only observe via a 15×15 RGB window, centered on their current location. The action space consists of moving left, right, up, and down, rotating left and right. Each species was assigned a unique color, shared by all conspecifics and preserved across all islands.

3 RESULTS

3.1 Exploration experiments

Given an unrefined and infrequently emitted behavior, reinforcement learning algorithms are very good at estimating its value with respect to alternatives and refining it into a well-honed strategy for achieving rewards. However, a central problem in reinforcement learning concerns the initial origin of such behaviors, especially in cases where the state space is too large for exhaustive search, and there are many local optima where the policy's reward gradient becomes zero.

This section explores how population dynamics may drive innovation in individual behavior. To study this, we introduce a new game that taxes individual exploration skills. It can be seen as a multi-agent analog of the well-known Montezuma's Revenge single player game that has often been used for studies of intrinsic motivations for single-agent exploration [2, 3, 9, 25, 29]. We hold to the game theory tradition of introducing each game with a facetious (but hopefully memorable) story, and offer the following:

In the *Clamity* game, agents begin in the trochophore stage of the bivalve mollusk lifecycle. They can freely swim around the map, a partially observed grid-world (map size = 36×60, window size = 15×15. Then whenever they are ready, they can perform the *settle* action. This action causes the agent to metamorphose into the adult clam stage of their lifecycle at their current location and removes their ability to swim. After settling, their shell grows around them, up to a maximum size. Shell growth is also restricted by the presence of adjacent shells from other clams. Each adult clam filters invisible food particles from the ocean at a rate proportional to the size of its shell, receiving reward for each food particle filtered. However, clam shells that are adjacent to the shell of another individual become unhealthy and do not filter any food. There are also nutrient patches located a considerable distance away from the starting location (more than 10 steps away, see maps in Fig. 2-A). Individuals that settle near a nutrient patch so that it is either partially or fully engulfed in their shell absorb additional nutrients from it. Episodes terminate after $T = 250$ steps. Settling immediately on the first action is a very attractive local optimum. The global optimum solution is to swim quickly out to a nutrient patch and settle there instead[3].

Single-agent reinforcement learning algorithms become stuck in the local optimum[4] and fail to ever discover the nutrient patches. To see why, consider the number of consecutive seemingly suboptimal actions that an agent would have to take in order to discover a nutrient patch. The settle action can be taken at any time, it always provides some level of rewards, and once taken, prevents movement for the rest of the episode. Thus any reasonable reinforcement learning algorithm that follows the initial gradient of its experience will reach the local optimum. If it starts out settling on step $t_s > 1$, it will receive an expected return of $T - t_s \times$ reward rate. But if it were to settle earlier instead, e.g., on step $t_s - 1$, it would receive a larger expected return. Thus there is a strong gradient from any policy initialization to the local optimum of settling on step $t_s = 1$. Furthermore. since the environment is partially observable, a single agent would need to choose to move in the same direction for several steps despite registering no change at all in its observation during that time.

On the other hand, Clamity can also be played by multiple agents simultaneously. All the trochophores begin each episode nearby one another in the center of the map. Since intersecting shells become unhealthy and provide no reward, individuals are penalized for settling too close to one another[5]. This provides a gradient that incentivizes agents to swim away from the starting location to avoid competing with one another for shell space. If the population size is large enough then this competition-motivated spreading eventually leads individuals to discover the nutrient patches[6].

3.1.1 Experimental procedure. To make like-for-like comparisons between single-agent and multi-agent training regimes, we adopt the following protocol. In parallel with the archipelago (N_I islands), we run L (the number of species) additional *solitary islands*. On the l-th solitary island, a single individual of species l plays each episode alone. All the experience generated on islands where species Ψ_l appears, even its solitary island, is used to update its policy π^l. However, the amount of each species's total experience derived from the solitary island is comparatively small since in this experiment, $M = 960$, the number of individuals of species l appearing across all islands of the archipelago. The final results are reported only from the solitary islands but reflect the policy learning accumulated in the competitive archipelago setting.

Our single agent training protocol simply sets the number of islands in the archipelago N_I to 0 and replicates each solitary island 32 times. Since there is only a single species ($L = 1$), and all solitary island replicas are the same as one another (though with different random environment seeds), the result is exactly equivalent to the A3C training regime [27].

This protocol also makes it easy to compare the proposed training regime where population sizes are dynamic and variable from episode to episode to the case of a "standard self-play" training regime, where population sizes are fixed. In this case, the archipelago contains just one island inhabited by a fixed number of individuals. As before, most of the experience is generated from the island where multiple individuals play. Results are reported only from the (single) solitary island, just as it is in the dynamic case.

3.1.2 Results. Individuals of species trained by Malthusian reinforcement learning find the globally optimum single-player solution, despite most of their experience coming from multi-player islands. Individuals trained by two baseline single-agent reinforcement learning algorithms completely fail to escape the local optimum. The first baseline we tried had all the exact same hyperparameters as in the Malthusian case, but all of its experience was in solitary islands (32 of them in parallel).[7] The second single-agent reinforcement learning algorithm baseline we tried was an implementation of the current state-of-the-art in curiosity-driven

[3] A video of the single-agent global optimum policy can be viewed here: https://youtu.be/AIT3FTC9s4s.

[4] A video of the single-agent local optimum policy can be viewed here: https://youtu.be/OHkpe9dVGyw.

[5] A video of such a multiplayer bad outcome can be viewed here: https://youtu.be/vrXOtHYMaPE.

[6] A video of a group of agents implementing a multiplayer global optimum joint policy can be viewed here: https://youtu.be/TnxMnSClBHY.

[7] These hyperparameters were not tuned for the Malthusian case—they were prespecified before the runs of both methods, and not subsequently changed.

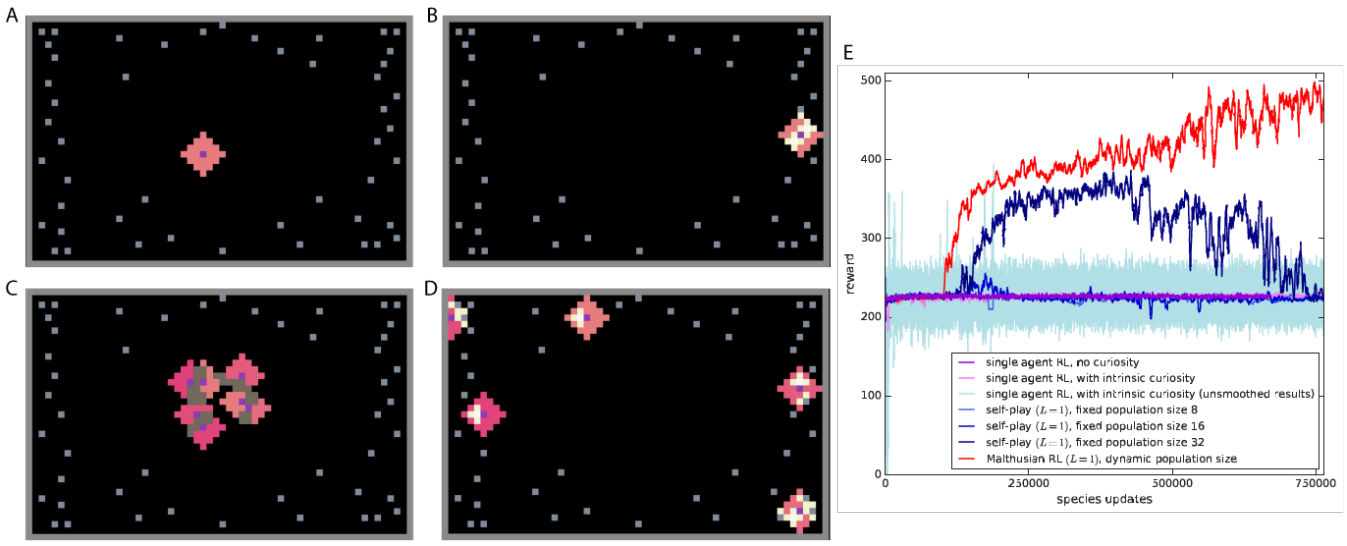

Figure 2: Experiments with extrinsically and intrinsically motivated individual exploration using the Clamity game. (A) Local optimum outcome. (B) Global optimum outcome. (C) Catastrophic multi-player outcome. (D) Multi-player global optimum. (E) Returns as a function of the number of times the species playing on the evaluated solitary island was updated. Except where indicated otherwise. reward values were smoothed over time with a window size of 100. Malthusian RL parameters were $\alpha = 0.0001$ and $\eta = 1.5$. Each episode lasted 250 behavior steps.

reinforcement learning, the intrinsic curiosity module provides a pseudo-reward to the agents based on its prediction error in predicting the next timestep in the evolution of a compressed encoding of its observations [25, 30]. In this case, augmenting the agent with the intrinsic curiosity module is still insufficient to get it it to consistently discover the nutrient patches. It does stumble upon them from time to time, especially early on in training (Fig. 2, but does not even do so consistently enough to register in a smoothed plot of rewards versus time with a 100 step smoothing window (Fig. 2). In contrast, individual members of species trained by Malthusian reinforcement learning with dynamic population sizes consistently implement globally optimal policies once they have discovered them (Fig. 2).

Next we asked whether dynamic population sizes were specifically important or whether the key was just the simultaneous training in multi-agent islands with a given, sufficiently large, population size. We noticed that most runs with dynamical population sizes converged to an island population size around 32 in the best performing islands. Thus we ran several experiments where agents trained in fixed population islands, evaluated on solitary islands as before. We found that individuals that trained in a fixed population size of 32 were able to discover the global optimum, but apparently less consistently than in the case with dynamic population size (red curve above navy curve in Fig. 2), and apparently with greater vulnerability to forgetting (the navy colored curve eventually declines back to the local optimum).

3.2 Mutualism and specialization experiments

Solution concepts for general-sum games may involve mutualistic interactions between synergistic strategies. Successful cooperative joint strategies may be either homogeneous, as in facultative mutualism, or heterogeneous. In nature, partners in mutually profitable associations are often very different from one another so that they can provide complementary capabilities to the partnership. In fact, most known mutualisms involve partners from different kingdoms, e.g., corals and their algae symbionts, vascular plants and mycorrhizal fungi, mammals and their gut bacteria, etc [5]. Moreover, division of labor and the subsequent efficiency gains from specialization are thought to be key components of complex human society [36].

However, it may be difficult to learn such mutually profitable partnerships of widely divergent strategies with self-play. All partners would need to represent all specializations, wasting valuable representation capacity. In addition, a policy learned by self-play requires a switching mechanism to break the symmetry and determine which sub-policy to emit in any given situation. For example, an agent could learn to become a blacksmith if standing on the left and a farmer if standing on the right. The complexity of the switching policy is itself related to the extent of partial observability in the environment. In some cases it may be very difficult to determine the right proportion of individuals needed to perform each part of the partnership at any given time, e.g. if the others' strategies cannot easily be observed. It would be easier to learn a heterogeneous set of policies, each one implementing only its own part of the partnership. But then, it would seem that the number of copies of each would have to be known in advance, thus adding many new difficult-to-tune hyperparameters, one for each species.

In this section we explore whether Malthusian reinforcement learning can find mutualistic partnerships more easily than other multi-agent reinforcement learning methods, especially when there

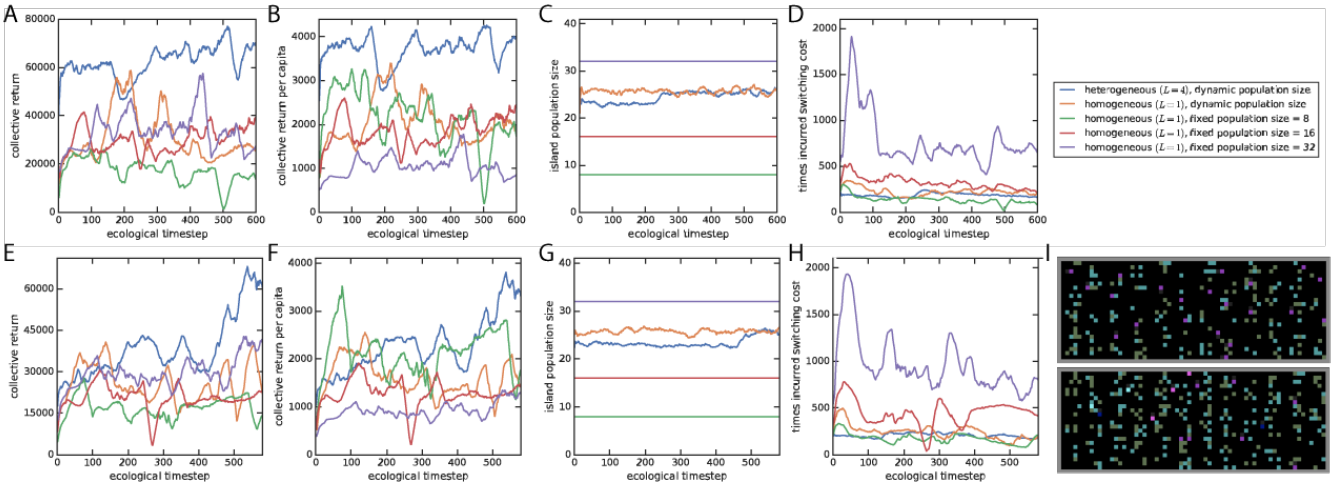

Figure 3: Experiments with the evolution of mutualism using the Allelopathy game. All results in this figure were smoothed with a window size of 25 ecological steps. (A-D) Unbiased Allelopathy game. Malthusian RL parameters were $\alpha = 1e-07$ and $\eta = 0.3$. E-H) Biased Allelopathy game. Malthusian RL parameters were $\alpha = 0.0001$ and $\eta = 0.01$. (A, E) Maximum collective return over all islands as a function of ecological time. (B, F) Maximum per capita collective return over all islands as a function of ecological time. (C, G) Maximum island population size over all islands as a function of ecological time. (D, H) Minimum number of times incurred a switching cost as a function of ecological time. ((I) Two screenshots of random procedurally generated initial map configurations. Maps were procedurally generated by randomly placing shrubs at the start of each episode. Episodes lasted 1000 behavior steps.

is a potential for gains from heterogeneous populations containing multiple specialized members. To test this, we created another partially observed Markov game environment. Again continuing the game-theoretic tradition of accompanying each game with a facetious and memorable story, we offer the following.

The Allelopathy game has two main rules. (1) shrubs grow in random positions on an open field. Shrubs allelopathically suppress one another's growth. That is, the probability that a seed of a given type grows into a shrub in any given timestep is inversely proportional to the number of nearby shrubs of other types. (2) Agents in Allelopathy are herbivorous animals that can eat many different types of shrub. However, switching frequently between digesting different shrub types imposes a metabolic cost since different enzymes must be synthesized for each. Thus, agents benefit from specialization in eating only a single type of shrub. Agents receive increasing rewards for repeatedly harvesting the same type of shrub (up to a maximum of $r = 250$). Rewards drop back down to their lowest level, ($r = 1$), when the agent harvests a different type of shrub (since that entails their switching into a different metabolic regime). Thus an agent that randomly harvests any shrub they come across is likely to receive low rewards. An agent that only harvests a particular type of shrub while ignoring others will obtain significantly greater rewards. The combined effect of these two rules is to make it so that a specialist in any one shrub type benefits from the presence of others who specialize in different shrub types since their foraging increases the growth rate of all the shrubs they do not consume.

We studied two variants of the Allelopathy game. The first variant, unbiased Allelopathy, has two shrub types A and B that appear with equal probability. In the second variant, biased Allelopathy, the two shrub types do not appear with the same frequency. Type A is significantly more common than type B. In addition, each shrub of type A consumed provides a maximum reward of 8 when at least 8 in a row are consumed. Whereas type B shrubs yield a maximum reward of 250 for any agent that manages to consume that many consecutively. Biased Allelopathy is a social dilemma since specialists in type B are clearly better off than specialists in type A, but both do better when the other is around.

3.2.1 Results. Here the critical comparison is between homogeneous ($L = 1$) and heterogeneous ($L > 1$) population dynamics. Therefore the object of study is the performance of the *islands* rather than specific individuals. The Allelopathy environment contains two niches, corresponding to specialization in consuming either shrub type A or B. In the heterogeneous case, mutualistic partnerships may develop from initial conditions where species in proximity to one another randomly fill either role. This situation features a gradient that guides each species in different directions. Whichever species begins with a propensity toward role A ends up specializing in role A. Likewise, the other species evolves to specialize in role B, to the mutual benefit of both partners. On the other hand, in the homogeneous case, it is still possible for mutualistic interactions to develop, but it is more difficulty since (1) both specialized parts of the joint policy must be represented in the same network, and (2) the policy must include a switching mechanism that breaks the symmetry, determining which sub-policy to implement in each situation. Heterogeneous species avoid the need for this symmetry breaking, and the relative proportions assigned to each role are handled naturally by the adapting relative population sizes (Fig. 4).

The total number of individuals in each experiment was $K = 960$. Thus in the homogeneous $L = 1$ case, $M = 960$, and in the heterogeneous $L = 4$ case, $M = 240$. The number of islands N_I was 60 in both dynamic population size conditions. In the fixed population size conditions the number of islands was chosen so that the total number of individual instances would still be 960, e.g., for fixed population size 32, this required $N_I = 960/32 = 30$.

Results were similar for both the biased and unbiased Allelopathy games. Heterogeneous ($L = 4$) population dynamics achieved higher returns, both per capita, and in aggregate, than the other tested methods including the homogeneous population ($L = 1$) with size dynamics (Fig. 3). Interestingly both heterogeneous and homogeneous runs converged to the same population size, but the heterogeneous case increased more slowly to that point, and did so while maintaining a higher per capita rate of return.

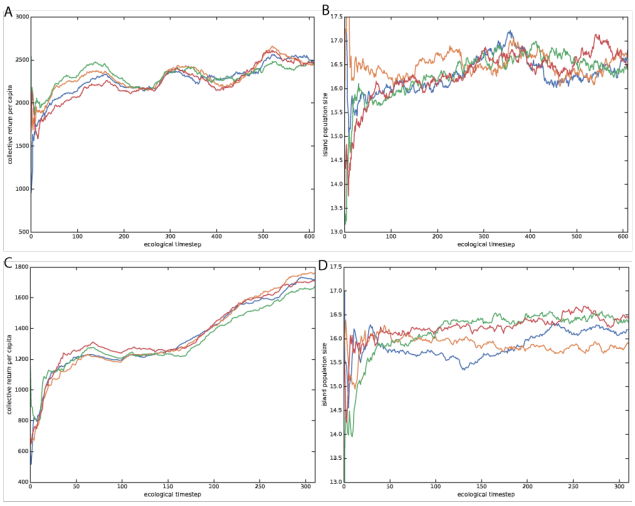

Figure 4: Representative island timecourses for the Allelopathy game. The different lines represent different islands. Notice that the results are consistent across islands. (A-B) results from the unbiased Allelopathy game. (C-D) Results from the biased Allelopathy game. (A, C) Collective return per capita as a function of ecological time for four representative islands. (B, D) Island population size as a function of ecological time for four representative islands.

4 DISCUSSION

This paper introduces Malthusian reinforcement learning, a multi-agent reinforcement learning algorithm that motivates individual exploration and takes advantage of possibilities for synergy to evolve heterogeneous mutualisms. If populations rise when returns improve then the problem itself shifts over time so no local optimum need ever be reached. This gives rise to a strategy that we may term *exploration by exploitation*. Individuals can always follow the gradient of their experience, they need never depart from their current estimate of the best policy just to explore the state space. They will naturally explore it, just by following the gradient in a changing world. Opportunities for heterogeneous mutualism may also be detected by gradient following. Initially weak specialization in one agent incentivizes its soon-to-be partner to specialize in a complementary direction, which in turn catalyzes more specialization, and so on.

How does this paper's proposed population dynamic relate to dynamics studied in evolutionary theory? Our requirement of conservation of compute, that the number of individuals of a given species on a given island may vary from episode to episode, but the total number of individuals of each species in the archipelago is always the same fixed value, implies that for populations to increase on one island they much decrease elsewhere. Thus the population dynamic introduced here may be understood as an evolutionary model of migration. Moreover, since fitnesses are computed globally, i.e. relative to the entire archipelago, it is more similar to *hard* selection models in evolutionary theory where populations are regulated globally than *soft* selection, where population regulation occurs locally within each island [14, 39, 40].

Other possible relationships between population size and innovation have appeared in the evolutionary anthropology literature. For instance, it is possible that—especially in preliterate societies—larger populations provide for more protection from forgetting of useful cultural elements since more elders, functioning as repositories of cultural knowledge, will be alive at any given time [15]. Or alternatively, larger social networks may provide more opportunities for recombination of disparate cultural elements that originated in farther and farther away contexts [4, 16, 21, 28]. As hypotheses for the origin of innovative behaviors in biology, these possibilities appear to be at odds with the mechanism implied by our algorithm, since each could explain, for instance, the same correlations between brain size and group size across the primate order [10] as well as innovation and (cultural) group size in humans [28, 32]. However, they are not mutually exclusive. In fact, all three mechanisms may even operate synergistically with one another. More research is needed in order to tease apart the precise mechanisms in biology. In computer science, we think this line of thought opens up a goldmine of new algorithmic ideas concerning the combination of population dynamics with social learning and imitation.

ACKNOWLEDGMENTS

We would like to thank Tina Zhu for coming up with the name "Clamity" and Oliver Smith for program management support.

REFERENCES

[1] Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2017. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748* (2017).

[2] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*. 1471–1479.

[3] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[4] Robert Boyd, Peter J Richerson, and Joseph Henrich. 2011. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences* 108, Supplement 2 (2011), 10918–10925.

[5] John F Bruno, John J Stachowicz, and Mark D Bertness. 2003. Inclusion of facilitation into ecological theory. *Trends in Ecology & Evolution* 18, 3 (2003), 119–125.

[6] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355* (2018).

[7] Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. 2005. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*. 1281–1288.

[8] Gregory Clark. 2008. *A farewell to alms: a brief economic history of the world*. Vol. 27. Princeton University Press.

[9] Edoardo Conti, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth O Stanley, and Jeff Clune. 2017. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. *arXiv preprint arXiv:1712.06560* (2017).

[10] RIM Dunbar and Susanne Shultz. 2017. Why are there so many explanations for primate brain evolution? *Phil. Trans. R. Soc. B* 372, 1727 (2017), 20160244.

[11] Anders Eriksson, Lia Betti, Andrew D Friend, Stephen J Lycett, Joy S Singarayer, Noreen von Cramon-Taubadel, Paul J Valdes, Francois Balloux, and Andrea Manica. 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proceedings of the National Academy of Sciences* 109, 40 (2012), 16089–16094.

[12] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, StockholmsmÃďssan, Stockholm Sweden, 1407–1416.

[13] Ted Goebel, Michael R Waters, and Dennis H O'rourke. 2008. The late Pleistocene dispersal of modern humans in the Americas. *science* 319, 5869 (2008), 1497–1502.

[14] Joseph Henrich. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. 53 (02 2004), 3–35.

[15] Joseph Henrich. 2004. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. *American Antiquity* 69, 2 (2004), 197–214.

[16] Joseph Henrich, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, et al. 2010. Markets, religion, community size, and the evolution of fairness and punishment. *science* 327, 5972 (2010), 1480–1484.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[18] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2018. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint arXiv:1807.01281* (2018).

[19] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A Ortega, DJ Strouse, Joel Z Leibo, and Nando de Freitas. 2018. Intrinsic Social Motivation via Causal Influence in Multi-Agent RL. *arXiv preprint arXiv:1810.08647* (2018).

[20] M L Johnson and M S Gaines. 1990. Evolution of Dispersal: Theoretical Models and Empirical Tests Using Birds and Mammals. *Annual Review of Ecology and Systematics* 21, 1 (1990), 449–480. https://doi.org/10.1146/annurev.es.21.110190.002313 arXiv:https://doi.org/10.1146/annurev.es.21.110190.002313

[21] Marius Kempe, Stephen J Lycett, and Alex Mesoudi. 2014. From cultural traditions to cumulative culture: parameterizing the differences between human and nonhuman culture. *Journal of theoretical biology* 359 (2014), 29–36.

[22] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. 2005. Empowerment: A universal agent-centric measure of control. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, Vol. 1. IEEE, 128–135.

[23] Shane Legg and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17, 4 (2007), 391–444.

[24] Thomas Robert Malthus. 1798. *An essay on the principle of population: or, A view of its past and present effects on human happiness*. Reeves & Turner.

[25] Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. 2017. Count-based exploration in feature space for reinforcement learning. *arXiv preprint arXiv:1706.08090* (2017).

[26] Paul Mellars. 2006. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proceedings of the National Academy of Sciences* 103, 25 (2006), 9381–9386.

[27] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[28] Michael Muthukrishna and Joseph Henrich. 2016. Innovation in the collective brain. *Phil. Trans. R. Soc. B* 371, 1690 (2016), 20150192.

[29] Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310* (2017).

[30] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, Vol. 2017.

[31] Adam Powell, Stephen Shennan, and Mark G Thomas. 2009. Late Pleistocene demography and the appearance of modern human behavior. *Science* 324, 5932 (2009), 1298–1301.

[32] Peter J Richerson, Robert Boyd, and Robert L Bettinger. 2009. Cultural innovations and demographic change. *Human biology* 81, 3 (2009), 211–235.

[33] Jürgen Schmidhuber. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2, 3 (2010), 230–247.

[34] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529 (2016), 484âĂŞ489.

[35] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

[36] Adam Smith. 1776. *An inquiry into the nature and causes of the wealth of nations: Volume One*. London: printed for W. Strahan; and T. Cadell, 1776.

[37] John R Stewart and Chris B Stringer. 2012. Human evolution out of Africa: the role of refugia and climate change. *Science* 335, 6074 (2012), 1317–1321.

[38] Gerald Tesauro. 1995. TD-Gammon, A Self-Teaching Backgammon Program, Achieves Master-Level Play. In *Applications of Neural Networks*. Springer, 267–285.

[39] Michael J. Wade. 1985. Soft Selection, Hard Selection, Kin Selection, and Group Selection. *The American Naturalist* 125, 1 (1985), 61–73. https://doi.org/10.1086/284328 arXiv:https://doi.org/10.1086/284328

[40] Stuart A. West, Ido Pen, and Ashleigh S. Griffin. 2002. Cooperation and Competition Between Relatives. *Science* 296, 5565 (2002), 72–75. https://doi.org/10.1126/science.1065507 arXiv:http://science.sciencemag.org/content/296/5565/72.full.pdf

[41] Yaodong Yang, Lantao Yu, Yiwei Bai, Ying Wen, Weinan Zhang, and Jun Wang. 2018. A Study of AI Population Dynamics with Million-agent Reinforcement Learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2133–2135.