On Reinforcement Learning, Nurturing, and the Evolution of Risk Neutrality

Kevin M. Robb
Robotics, Evolution, Adaptation, and Learning Lab
Engineering Physics Program
Gallogly College of Engineering
University of Oklahoma
kevin@js-x.com

Dean F. Hougen
Robotics, Evolution, Adaptation, and Learning Lab
School of Computer Science
Gallogly College of Engineering
University of Oklahoma
hougen@ou.edu

Abstract— Reinforcement learning depends on agents being learning individuals, and when agents rely on their instincts rather than gathering data and acting accordingly, the population tends to be less successful than a true RL population. "Riskiness" is the elementary metric for determining how willing to rely on learning an individual or a population is. With a high learning parameter, as we denote riskiness in this paper, agents find the safest option and seldom deviate from it, essentially using learning to become a non-learning individual. With a low learning rate, agents ignore recency entirely and seek out the highest reward, regardless of the risk. We attempt in this paper to evolve this "risk neutrality" in a population by adding a safe exploration nurturing period during which agents are free to explore without consequence. We discovered the environmental conditions necessary for our hypotheses to be mostly satisfied and found that nurturing enables agents to distinguish between two different risky options to evolve risk neutrality. Too long of a nurturing period causes the evolution to waver before settling on a path with essentially random results, while a short nurturing period causes a successful evolution of risk neutrality. The non-nurturing case evolves risk aversion by default as we expected from a reinforcement learning system, because agents are unable to distinguish between the good risk and bad risk, so they decide to avoid risks altogether.

I. INTRODUCTION

Reinforcement learning (RL) is the most fundamental method for evolving an intelligent AI system. RL involves each agent in a population making a choice at each trial and receiving some sort of feedback, which the system must then use to improve. The population will undergo many iterations, called "generations", throughout which the AI will ideally improve and optimize its reward function. This method of learning does not involve any specific aid or supervision and does not necessarily include nurturing in any form. Niv et al. [1] demonstrated that the RL model can be applied to animal behaviors and found a method for evolving the ideal learning rules to maximize reward in an uncertain environment. Their research set a basis for many future experiments with reinforcement learning, and their most relevant discovery to this paper is the observation that RL causes risk averse behavior to emerge as the optimal strategy in a population. It has been quantified and further demonstrated that RL without nurturing consistently causes the evolution of risk aversion, meaning the agents who choose the safe option outperform agents who choose the risky option, and as a result pass on their genes more frequently.

Risk aversion refers to the tendency of an agent to create or fall into a positive feedback loop in which the safe option is chosen the vast majority of the time, in order to keep receiving reward, even if the average reward for the risky option is higher; however, Roberts [5] demonstrated that there is a threshold for the average reward difference, above which a population which formerly developed risk aversion will no longer always divert to the safe option, and may tend towards the risky option. Risk neutrality refers to a tendency to choose based on average or expected reward, rather than the actual probability of receiving zero reward from each option. Risk averse populations place more weight on recent results, and risk neutral populations place more weight on the overall trends of their results, rather than recency.

A risk-seeking individual would choose the risky option without regard to its average or expected reward, meaning even a state in which the risky option has a lower average than the safe option, but the nonzero outcome is higher than the safe option, could cause a risk-seeking population to diverge on the risky option. This type of behavior can manage to take over a population if a few individuals get lucky with the risky option, and then multiply by more than a factor of two each generation (due to a large tournament size). We want to avoid this effect in our experiment and only promote the evolution of risk neutrality or risk aversion. We explain how we plan to avoid this in the experimental design, section III.

A relatively recent emergence in the artificial intelligence field is the introduction of nurturing, in forms such as direct supervision, education, safe exploration, etc. A virtuous cycle of nurturing and

learning has been proposed, in which the evolution of nurturing promotes the evolution of learning, which in turn promotes greater nurturing, continuing the loop [2-4]. In this paper, we are exploring the ways in which the addition of a safe exploration period may affect the evolution of a "learning parameter" which is correlated to the riskiness level of a certain agent. This learning parameter is described in more detail in the experimental design, section III.

Safe exploration is a form of nurturing in which each agent is free to explore its options for a certain number of trials near the start of its life without any punishment; this allows it to alter expected rewards and learn a strategy which is likely to lead to success once the safe exploration period has ended and fitness calculation has begun. It was shown by Hoke [4] that the addition of a safe exploration period can cause learning itself to be more likely to evolve, so we are curious what effect safe exploration will have on a population with a set learning procedure and variable riskiness. Shah's results [2] indicate that nurturing promotes learning only in certain environments where it is desirable, which suggests that the effect of nurturing on a learning population is not always predictable. We hypothesize in our study that risk aversion will be less likely to emerge as the optimal strategy in the nurturing case, since the population will have time to learn that the risky option has a higher expected reward.

Conversely, it is not necessarily evident that risk aversion must evolve in any non-nurturing population. Roberts [5] demonstrated that reinforcement learning with knowledge sharing causes risk neutrality to emerge in a population, rather than the risk aversion which prevails when each agent is on its own. This experiment suggests a basis for enhanced knowledge leading to a population drifting toward risk neutrality. Since safe exploration similarly leads to an increase in knowledge confidence, this follows the same line of reasoning as our first hypothesis.

II. HYPOTHESES

Reinforcement learning has been shown to cause a drift toward risk aversion on its own [1], due to the impact each choice has on the continued success of the individual and the population. If this impact is removed for the majority of the individual's lifetime with the addition of nurturing [4], it would follow that more risks could be taken safely, and risk aversion would not emerge until near the end of its lifetime, if at all. Consequently, a long enough safe exploration period could directly cause a population to become risk neutral. Thus, hypothesis 1 follows:

H_1: Reinforcement learning with nurturing in the form of a long safe exploration period leads to the evolution of a risk neutral learning parameter.

Accordingly, if our assumptions about the effect of nurturing are valid, the complement to hypothesis 1 should also be true, leading to hypothesis 2:

H_2: The absence of nurturing will cause the learning parameter to evolve to be risk averse.

It is not necessarily accurate to state that both of these hypotheses must be true or false together, because even if the learning parameter evolves in a direction that agrees with our first hypothesis, the addition of nurturing may have had a negligible impact on the actual evolution of the learning parameter. It should be evident in our results that the difference in nurturing between cases is the source of the trends we see in the evolution of the learning parameter, and the differences in the outcome of the learning parameter should be statistically significant.

III. EXPERIMENTAL DESIGN

At the start of each trial, every individual will be presented with a choice between three options, A, B, and C. Option A is the "safe" option, with a guaranteed turnout and a low average reward. Option B and C are the "risky" or "uncertain" options, with a 50% probability of turnout for both. B has a higher average reward than A, and C has a lower average reward than A. Each individual will evolve their learning parameter between trials using reinforcement learning, and this parameter will be used to evaluate risk and choose option A, B, or C during each trial. The learning parameter L will be in the range (0, 1). An L value of 0 describes complete risk neutrality, where risk, reward, and recency are completely ignored when making a choice, while an L value of 1 represents complete risk aversion, where only the most recent trial is considered. We expect that with an L value very close to 1, there will be no deviation once the safe option is chosen, regardless of the potential reward from the other option. At the end of each generation, a variety of data is written to a summary file, including the average, minimum, and maximum of all L values for the population, the average fitness of the population, and the ratio of choices made in favor of each option. There is also a graphing method which has various different sets of outputs that are compatible for graphing; the output set is defined by combinations of three Boolean parameters in the setup file. After writing to the file, the next generation will be formed.

The individual will have N trials in which to acquire fitness before the next generation is formed. This resource gathering period will be called its "lifetime" for simplicity, although it is implied that the individuals in the nurturing case would persist for one more "lifetime" to provide the nurturing period for the next generation. The reward gained by an individual in a specified ith trial is represented by R_i. For our initial set of data, there will be 500 trials in one lifetime. The reward for all 500 will be averaged to obtain the fitness

F for the lifetime of a non-nurturing individual, as in Eq. (1), and for a nurturing individual the total fitness *F* will be the average reward gained during the final 350 trials, ignoring the first 150, as in Eq. (2).

$$F_{non-nurturing} = \frac{1}{N}\sum_{i=0}^{N-1} R_i = \frac{1}{100}\sum_{i=0}^{499} R_i \quad (1)$$

$$F_{nurturing} = \frac{4}{N}\sum_{i=\frac{3N}{10}}^{N-1} R_i = \frac{1}{25}\sum_{i=150}^{499} R_i \quad (2)$$

The procedure for an agent making its choice between options A, B, and C at each trial uses a Boltzmann algorithm. This algorithm uses an agent's expected reward for each option to generate a probability of how likely each is to produce the greater comparative reward, and then boosts the chance of the agent choosing the option of greater expected value. For example, if the options are 80%, 10%, and 10%, the agent is realistically more certain to receive a greater reward from the 80% option, so the algorithm makes it *more likely* that the 80% option is chosen (i.e., its chance of being chosen is greater than 80%). This method leaves a small chance of the lesser option(s) being chosen to prevent a ceiling or floor effect, and to account for random chance. The Boltzmann algorithm uses a temperature variable *T* in its calculations, representing how certain the individual is in its choice. A higher value of T represents lower certainty in expected rewards, and a T value approaching 0 represents complete certainty, and would cause the individual to pick whichever option has a higher expected reward 100% of the time. It would be interesting to scale the temperature variable throughout an individual's lifetime, where it would start out high when an individual is young and doesn't have much sample data, but would decrease and cause an agent to be very confident in its data by the end of life. For this experiment, however, we will follow Roberts' procedure [5] and use a constant temperature value of $T = 20$, rather than scaling it with an agent's lifetime. This is done with the intent of avoiding the introduction of confounding variables to our study which could alter our results.

Each agent's current expected reward values are used by the Boltzmann algorithm to generate probabilities and ultimately make each choice. Every time a choice is made, the expected reward for the option that was chosen is updated. The expected reward *E* is changed following the same update rule as used by Eskridge [3], in which a simple weighted average is calculated, using the learning parameter to define the weights. The learning parameter *L* denotes the weight of the new reward *R* just obtained, versus the current expected reward E_0. This update rule is shown in Eq. (3). It is evident from this equation that an *L* value approaching 1 would place very little weight on previous data and all weight on the most recent result; conversely, an L value approaching 0 would have the opposite effect, where essentially no weight is placed on the most recent result, and previous data is carried through.

$$E_{new} = (1 - L)E_0 + LR \quad (3)$$

There will be 500 generations in total, each containing 50 individuals. Each new generation will be formed with tournament-style selection, where two individuals of the current population of fifty are selected at random, and the member of that group of two with the highest fitness is selected to be copied to the next generation, allowing the same individual to be selected more than once, even in the same tournament. This tournament size is so low because a higher tournament size (tested up to five) disproportionately favors lucky risks from individuals, and causes all cases to regress toward risk-seeking individuals, as we discussed in section I. This process will be done fifty times to form a new population of 50 individuals. This formation method translates to an individual's selection likelihood being correlated to its end-of-life fitness and allows for a small chance of low-fitness individuals to also advance, while maintaining a high chance to prune off gene lines of low fitness. This selection method also removes the necessity of calculating a cost-of-living fitness and killing off low performing individuals, as the punishment for low fitness is simply that they are less likely to reproduce and pass on their genes.

Selected individuals will be assigned a mutation value M from a normal distribution with mean 0 and standard deviation 0.05, so $M \sim N(0, 0.05^2)$. The value of *M* will be added to *L* to generate the learning parameter value for the new individual, as in Eq. (4). The value of *L* will not be altered during an agent's lifetime, and only changes due to mutations between generations, which is why we have a relatively small population size and a large number of generations. This method of evolution serves two important purposes: it simulates natural selection by eliminating individuals from the gene pool who significantly underperformed compared to the rest of the population, and it simulates the carrying capacity of the environment by maintaining the size of the population.

$$L_{new} = L_{old} + M \quad (4)$$

Instincts are represented by the initial values of weights such as the learning parameter and anticipated values of options A, B, and C, which would affect risk evaluation by skewing the expected values of each option, altering the individual's willingness to undertake risks. We anticipate that if the instincts for each individual in a population were independently random, the generational algorithm would cause the same effect as if all were initially average, and eventually weed out poorly performing genes. It

would be interesting to see how a systematic initial instinctual preference in the population would affect the evolution of the learning parameter, and how it would distort the overall success of the population through several generations; however, in this experiment we will not perform tests with instinctual differences. At the start of every generation, all expected values will be set to 100 to ensure that these values are learned every generation, and the individual's performance will be based solely on their learning parameter value, rather than on any initial preference.

At the start of the first generation, the learning parameter of each agent will be set to 0.5, starting the individuals out with no risk preference. This will allow the parameter to dip towards 0 or 1 with equal likelihood due to random chance, and evolution will proceed further from 0.5 until a balance is reached. We hope to avoid a ceiling or floor effect, where the result of both the nurturing and non-nurturing case are so close to the same boundary (0 or 1) as to be indistinguishable. All values discussed in this section are represented in Table 1 for the nurturing case.

Setup Parameter	**Value**
Initial Learning Parameter	0.5
Standard Deviation of Mutation	0.05
Tournament Size	2
Num. of Nurturing Trials	150
Num. of Total Trials per Gen.	500
Num. of Agents per Gen.	50
Number of Generations	500
Value of Choice A	100
Value of Choice B	0 or 220
Value of Choice C (symmetric)	0 or 180
Value of Choice C (asymmetric)	0 or 150

Table 1: The set of parameters which are used to initialize and run the simulation in the nurturing case. The only difference in the non-nurturing case is that the number of nurturing trials is set to 0. Notice the value for A is 100, which is higher than the average value of C (90 or 75), and lower than the average value of B (110).

IV (a). RESULTS (TWO-STATE)

Figures 1 and 2 represent data taken over much fewer generations and trials than in Table 1, and in a two-state environment containing only options A and B. Contrary to what we expected to see from our results, there seems to be no definitive set of parameters for which the addition of nurturing causes a consistent change from risk aversion to risk neutrality. Our data showed a progression of the average learning parameter value that depended entirely on the random choices and rewards in the early generations, rather than the fitness calculation differences between the nurturing and non-nurturing cases. Ten runs of our simulation showed five which approached 0 and five which approached 1. These runs are shown together in Figure 1, clearly demonstrating the random variations which in some cases cause the average learning parameter to fluctuate wildly throughout the dataset. This fluctuation and the aspect of randomness seem to be far less apparent and influential in the non-nurturing case than in the nurturing case. Thirty runs of the simulation are shown in Figure 2, and it is clear that there is almost no variation in the evolutionary process or the results. Every non-nurturing dataset resulted in an evolution of risk neutrality, even in runs which approached 1 in early generations.

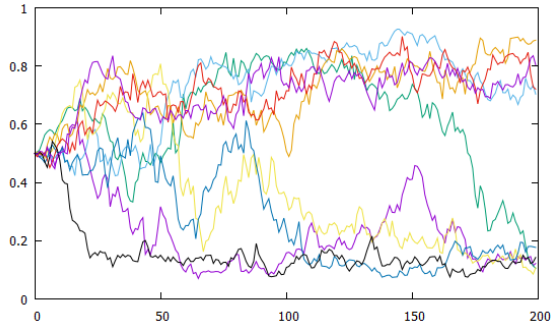

Figure 1: Our simulation, with the same parameters each time, produces results which essentially do not depend on the nurturing factor. Shown above are 9 nurturing runs, with average learning parameter plotted as a function of generation number. It is clear that these results are virtually random in their evolution.

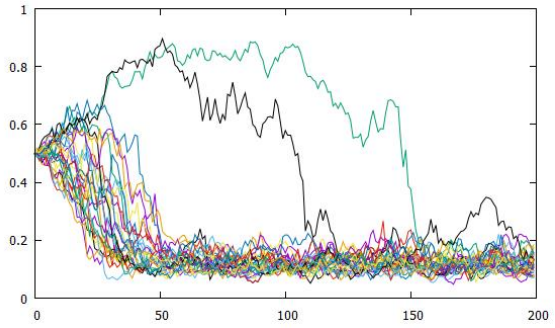

Figure 2: Thirty non-nurturing runs which for the most part show a consistent progression toward risk neutrality in every run. This is in opposition to the nurturing case, which has very indeterminate results.

When we increase the number of generations and trials (but keeping the two-state environment), we see the results shown in Figures 3 and 4, where the results are opposite our expectations.

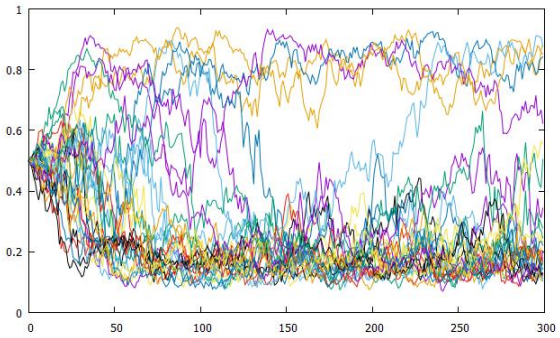

Figure 3: Average L as a function of generation number. Thirty non-nurturing runs which are very random but mostly end up with an average learning parameter value around 0.2 (very risk neutral).

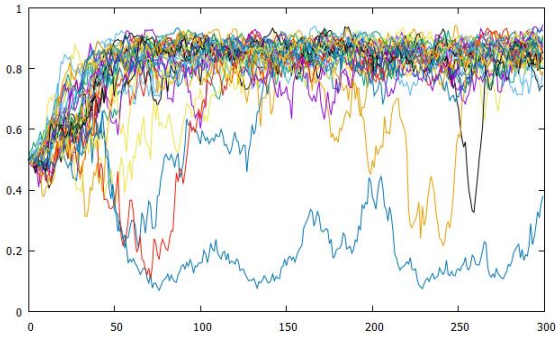

Figure 4: Average L as a function of generation number. Thirty nurturing runs which almost entirely end up with an average learning parameter value around 0.8 (very risk averse).

We anticipated that the safe exploration period would allow a cost-free method for an agent to learn that the risky option B has a higher average value and should be the obvious choice, evolving risk neutrality. We also expected in the non-nurturing case that individuals who evolved risk neutrality would've lost too much fitness in the process of discovering the higher average value and would be eliminated from the population during tournament selection. What actually seems to occur is that individuals in the non-nurturing population who happen to get lucky a few times with the risky option make it through the tournament selection and are passed on multiple times, which snowballs each generation until the entire population evolves to be risk-neutral. In the nurturing case, some agents discover that option A is more reliable and have success choosing this safe option every time, whereas other agents discover that option B has a higher expected reward and have success with choosing this risky option the majority of the time. This effect causes a wide spread of the L values during any given generation, with an average that fluctuates around the center until dipping towards 0 or 1. Figure 5 shows the progression of this spread by displaying the minimum, maximum, and average L values for each generation, as well as the L values representing the first and third quartile in a data set for a nurturing case. These data are a clear indication of the obtuse lack of precision when compared to the non-nurturing data in Figure 6.

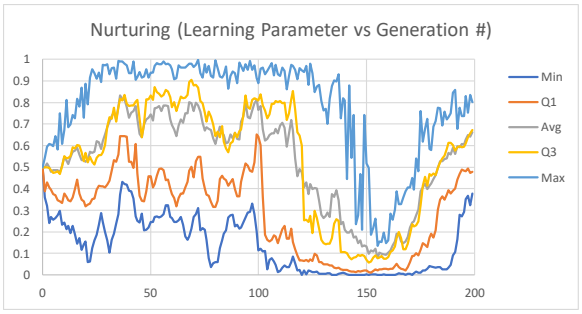

Figure 5: Progression of 5 measures for learning parameter in one nurturing case. Notice the wide disparity between the minimum and maximum values, as well as between the first and third quartile.

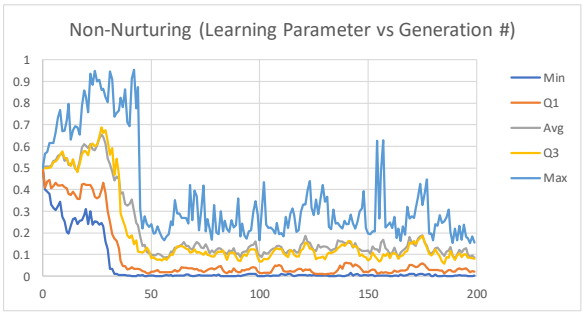

Figure 6: Progression of the same 5 measures as in Fig. 3, but for a non-nurturing case. Data spread is far more precise and finds the optimum learning parameter very early without much fluctuation.

On an individual level rather than a generational level, the data also show unexpected trends. It is evident by the plots in Figures 7 and 8 that although a decrease in average learning parameter of the population does not change the average fitness by much, it is enough to noticeably increase variability in the spread of fitness values.

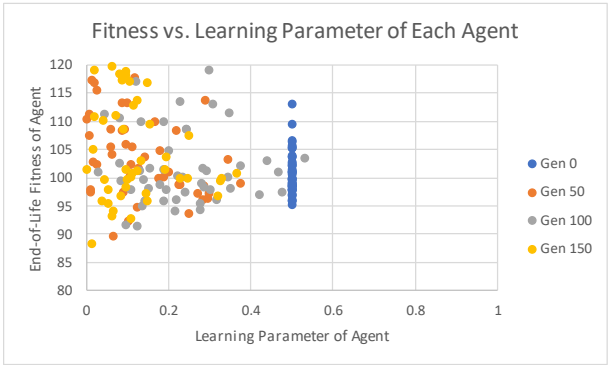

Figure 7: Non-nurturing case. Most agents quickly evolve risk neutrality (L→0) and the general spread remains the same for generations 50-200.

In terms of expected values, the agents are calculating and updating their expectations as we would expect based on their actions, as shown in Figure 7. They are also changing their behavior to match their expectations, choosing option B more often the higher its expected value is. There seems to be a linear relationship between the proportion of an agent's choices during its fitness-collection period, which is

acceptable and predictable given the Boltzmann algorithm we are using.

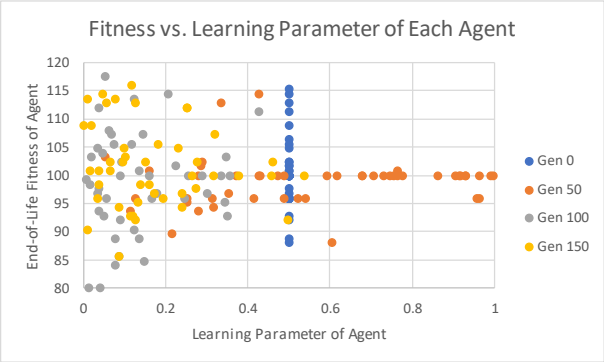

Figure 8: Nurturing case. In gen 50, most agents are consistently (mildly) successful with risk aversion (to the right on the graph) but few individuals obtain higher fitness by going left. By gen 100, these few have dominated the gene pool and nearly all agents are risk neutral, with the same spread as in Figure 5. The data in the nurturing case are sometimes mirrored from this plot, when risk aversion evolves.

We originally thought all this to mean our hypotheses were wrong, however we decided to try reducing the randomness and inconsistency before making any definitive claims about these results. This is why we introduced a third state, option C.

IV (b). RESULTS (THREE-STATE)

With all three states in play, the average between the two "uncertain" options (B: 110 and C: 90) average the same as the "safe" option (A: 100). This means that being random is *not* beneficial anymore, because being risky is on average *the same* as being risk averse, and is therefore deterred, because with no change in reward, individuals will want to decrease variability.

The graphs shown in Figures 9 and 10 each represent a single agent in a single generation. This is the first iteration of our three-state environment. Ideally the estimations will each converge to their real mean values, but this does not happen successfully yet. Significant spikes and drops mean the agent has a high L and is very risk averse, whereas if the lines experience minimal change, the agent has a low L and is very risk neutral. As these two figures show, our results still oppose our hypotheses, but more importantly, the agents still fail to distinguish between the two "uncertain" options, B and C.

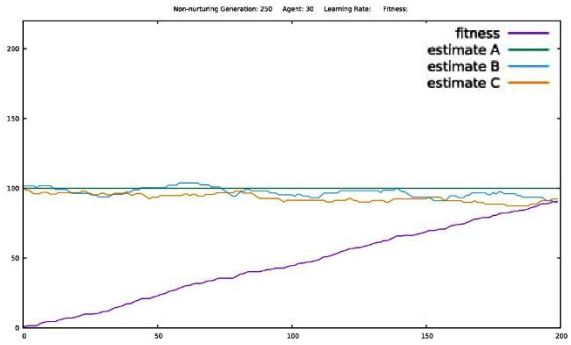

Figure 9: Non-nurturing case. Fitness and estimations as a function of trial number. 200 trials per generation.

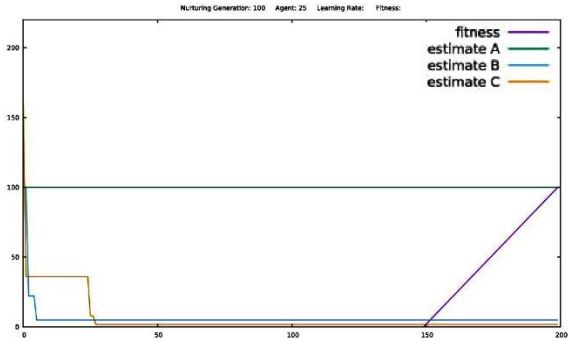

Figure 10: Nurturing case. Fitness and estimations as a function of trial number. 200 trials per generation. The flat segment of the fitness line is the nurturing period, wherein fitness is not tracked. This nurturing period lasts for the first 150 trials.

In the physical world, nurtured organisms (e.g., humans) live well beyond their nurturing period. For this reason, we increased the total number of trials to 500, leaving the nurturing period at 150 trials. This stretches the relative length of the nurturing period from 75% down to 30% of an agent's life. This change causes nurtured agents to also evolve risk neutrality, and both sets of agents are more successful, as shown in Figures 11 and 12.

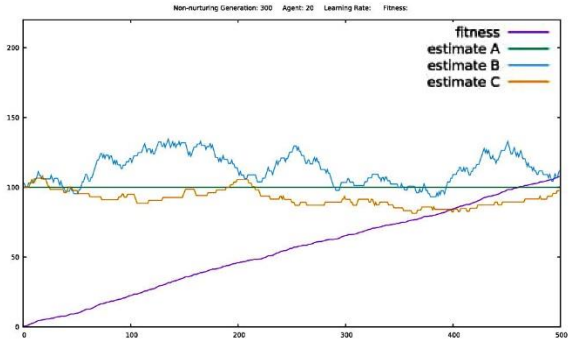

Figure 11: Non-nurturing case. Fitness and estimations as a function of trial number. 500 trials per generation.

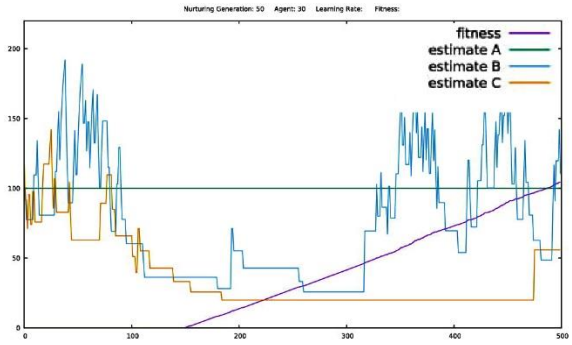

Figure 12: Nurturing case. Fitness and estimations as a function of trial number. 500 trials per generation. Nurturing period is the first 150 trials, so no fitness is tracked during this time.

The next development was the reintroduction of asymmetry to the states, but rather than in the two-state system, the safe option is favored over blanket riskiness, and even over randomness. We achieve this by lowering the average value of option C to 75, while leaving the average for B at 110. This means the average of B and C (i.e., what an agent would get if they are risk neutral but cannot distinguish between the two choices) is 92.5, i.e., lower than option A (100). Because it is lower, *being risky is bad unless an agent can tell the difference between B and C*. This is the key; option B choosers will end up with 110 fitness, which puts them at the top of the population, and therefore the B choosers will take over the gene pool over enough generations.

What we see in our final results (Figures 13 and 14) is that nurturing allows agents to learn to choose B and avoid C, evolving risk neutrality; however, without nurturing, agents avoid both B and C, evolving risk aversion and choosing A the majority of the time. This behavior satisfies both our initial hypotheses, and the environment fits our criteria, producing reliable and consistent data.

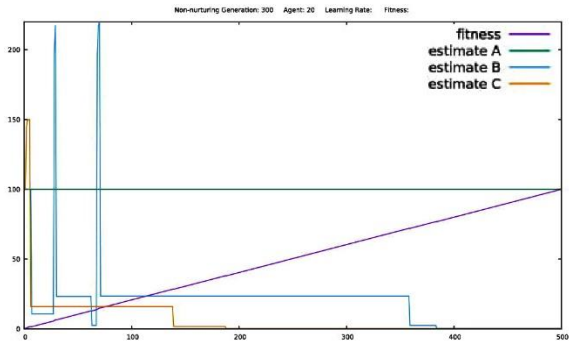

Figure 13: Non-nurturing case. Fitness and estimations as a function of trial number. 500 trials per generation. Negatively skewed asymmetry. Risk averse, as seen by the large spikes in estimations. This population ended closely gathered around a mean fitness of 100.

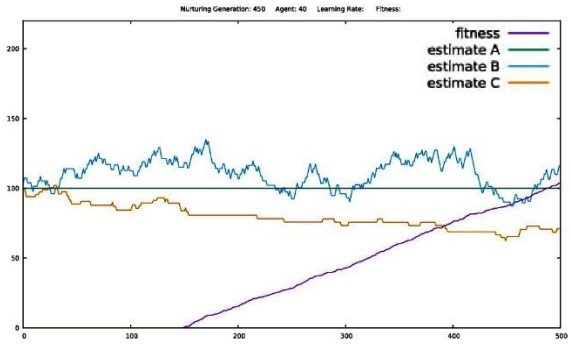

Figure 14: Nurturing case. Fitness and estimations as a function of trial number. 500 trials per generation. Nurturing period is the first 150 trials, so no fitness is tracked during this time. Negatively skewed asymmetry. Risk neutral, as seen by small changes in estimations with each trial. This population ended very spread out around a mean fitness greater than 100.

V. DISCUSSION

Our final results with the simulation demonstrate that our second hypothesis, H_2, is correct. Agents in the non-nurturing case are unable to distinguish between options B and C, so they develop risk aversion in order to skip both of these options entirely.

The story with our first hypothesis, H_1, is a bit more complicated. Agents in the nurturing case *are* able to choose B while avoiding C and are thus able to evolve risk neutrality. However, this does not happen with a "long" safe exploration period, but rather happens with a short nurturing period lasting only 30% of the agent's life. This is, as we discovered in the course of running this experiment, due to the simple statistical fact that a fitness collection period which is too short gives too much power to random chance, rather than actual decision-making, causing the large spread in the data shown in Figures 7 and 8. We initially thought H_1 to have been proven false because of its underlying predictions about reinforcement learning, but it turned out to be false because of our requirement of a long nurturing period. A modified H_1 which is true would be as follows:

H_1(modified): Reinforcement learning with nurturing in the form of a short safe exploration period leads to the evolution of a risk neutral learning parameter.

Moving on from our hypotheses, the situation our simulation models in our final setup can be described as a "fitness landscape." This fitness landscape (Figure 15) has two optima in the learning parameter (L) scale, and which one gets settled by the population depends on whether nurturing is present.

- Risk neutrality is the global optimum, and can be found when the population has a low average L. This optimum has a small basin of attraction, meaning it is hard to find.

- Risk aversion is a local optimum but is less lucrative than the global optimum. It can be found when the population has a high average L. This optimum has a large basin of attraction, meaning it is easy to find.

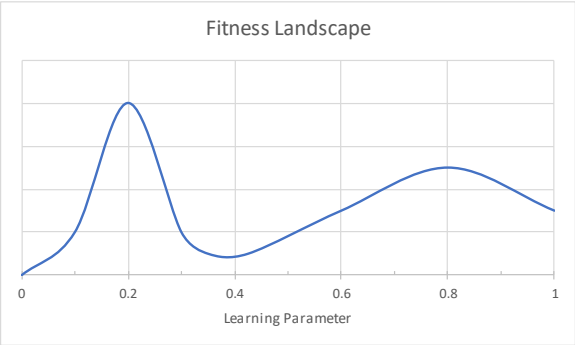

Figure 15: Fitness landscape of our simulation's environment. It is evident that the global optimum at L=0.2 is much narrower and is thus harder to find. The simulation starts at L=0.5, so agents will follow the upwards curve to the local optima at L=0.8; however, nurturing enables agents to experiment and dip into the global optimum's basin of opportunity

This experiment has served to demonstrate the type of environment (i.e., fitness landscape and fitness function) that is required in order for a population to be able to develop risk neutrality and/or risk aversion, and for the insertion of nurturing to be able to make a difference to the results of the evolution. This will enable us to have some idea of what to expect when attempting to evolve a neural network for a specific purpose, and even allow us to guide its evolution by predefining a certain fitness landscape.

VI. FUTURE WORK

The most significant work still to be done is to increase the number of choices available to each agent, as we did from two to three while refining this experiment. Being able to generalize these findings and this model to any number of choices would be outstanding, and this is what my primary goals for my own personal future work entail. It is also a subject of interest to make some choices become unavailable and reopen over the course of the agent's lifetimes or across generations, to force adaptation and see how well a nurturing versus non-nurturing population would do in surviving the changes.

It would be feasible to put nurturing and non-nurturing agents in the same simulation, such that the two populations are opposed and must find separate niches to both survive. This would turn the competition from a data comparison into a real-time test of natural selection and would really reveal the intricacies of each group's behavior.

In future experiments, it would be interesting to include additional factors for evolution such as instincts which carry over between generations, an initial instinctual preference, and varying initial values of the learning parameter within the same generation or different runs. We assumed in this experiment that variance in the first generation's initial expected values and learning parameter would lead to the same progression as a population with each individual beginning with average values, given the same evolution parameters. A possible follow-up experiment would be to determine the accuracy of this assumption, and the conditions under which it fails. We wonder if a threshold exists for the variance in the initial instincts, and if it can pass a point after which its behavior changes and is no longer predictable by our model.

We also assume that a significant difference in the initial average value of the learning parameter will affect the direction towards which it drifts through evolution (towards 1 for risk aversion or towards 0 for risk neutrality). This assumption was not relevant in this experiment, as we began each trial with the same learning parameter value which allowed a drift to occur in either direction, but this is something which could most certainly stand to be challenged or proven in future experiments.

This experiment could be repeated with an aspect of variance in some of our parameters which remained unchanged throughout the entirety of the experiment, such as the standard deviation of the mutation value, the number of agents in a population, the tournament size, and the temperature value used in the Boltzmann algorithm. Specifically, we proposed in the experimental design (Section III) that the temperature value could be decreased through the duration of every individual's lifetime, such that they become more confident in their answers as their lives go on. With this strategy, the temperature value would not undergo evolution, but rather would start at a constant value of 20 at the beginning of every generation, and would be decreased by the same amount each trial such that it will reach 0 in the final trial of every lifetime. Evolution with this sort of condition would place disproportionate weight on earlier trials compared to later, as the expected values are less able to change later on in an agent's life.

A different standard deviation of the mutation value would alter the weight of each specific generation, as the learning parameter would change more or less each generation, causing a general less or more drastic change of the average L value as generations proceed. Further experimentation could be conducted in which the standard deviation is changed between data sets, and there is a possibility of altering this value within a data set to simulate changes in the environment which either necessitate more rapid adaptation or constrict the population to less rapid adaption.

VII. REFERENCES

[1] Yael Niv, Daphna Joel, Isaac Mailijson, and Eytan Ruppin, "Evolution of Reinforcement

Learning in Uncertain Environments", Adaptive Behavior 10, no. 5 (2002), doi: 10.1177/ 10597123020101001

[2] Syed Naveed Hussain Shah, Ingo Schlupp, and Dean F. Hougen, "Nurturing promotes the evolution of reinforcement learning in changing environments regardless of instincts", Submitted to Adaptive Behavior, 27 pages, 2017.

[3] Brent E. Eskridge and Dean F. Hougen, "Nurturing promotes the evolution of learning in uncertain environments", Second Joint IEEE International Conference on Development and Learning / Epigenetics Robotics Conference, 6 pages, 2012.

[4] Bryan Hoke and Dean F. Hougen. "Nurturing Promotes the Evolution of Generalized Supervised Learning." Accepted to IEEE Congress on Evolutionary Computation. 8 pages, 2018.

[5] Steven A. Roberts and Dean F. Hougen, "Information and resource sharing in reinforcement learning agents dealing with risk", In Preparation, 7 pages, 2018.