Predicting Population Dynamics and Evolutionary Trajectories based on Performance Evaluations in Alife Simulations

Matthias Scheutz and Paul Schermerhorn
Artificial Intelligence and Robotics Laboratory
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556, USA
{mscheutz,pscherm1}@cse.nd.edu

ABSTRACT

Evolutionary investigations are often very expensive in terms of the required computational resources and many general questions regarding the utility of a feature $\mathcal{F}$ of an agent (e.g., in competitive environments) or the likelihood of $\mathcal{F}$ evolving (or not evolving) are therefore typically difficult, if not practically impossible to answer. We propose and demonstrate in extensive simulations a methodology that allows us to answer such questions in setups where good predictors of performance in a task $\mathcal{T}$ are available. These predictors evaluate the performance of an agent kind $\mathcal{A}$ in a task $\mathcal{T}*$, which can then transformed by including costs and additional factors to make predictions about the performance of $\mathcal{A}$ in $\mathcal{T}$.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*; I.6.0 [**Simulation and Modeling**]: General

General Terms

Experimentation, Performance

Keywords

A-Life, Adaptive Behavior, Agents, Evolution

1. INTRODUCTION

Evolutionary computation is a tool widely used in the "A-Life" and "Adaptive Behavior" communities to explore the space of possible designs of artificial creatures or biologically inspired agents, ranging from their physical make-up to their control systems ([2, 1, 5, 12]). Often, genetic algorithms are employed to demonstrate that a feature $\mathcal{F}$ can evolve in an agent type $\mathcal{A}$ ([3, 9]). The evolvability of $\mathcal{F}$ in $\mathcal{A}$ is then used to support two kinds of claims: (C1) $\mathcal{F}$ is beneficial to $\mathcal{A}$ (as it increases the fitness of $\mathcal{A}$) ([6, 4]), and (C2) $\mathcal{F}$ can and/or is likely to evolve in agents of type $\mathcal{A}$ ([7, 8, 13, 10]). The former is a claim about the utility of $\mathcal{F}$ for $\mathcal{A}$ (in a given task and environment), the latter about the possibility and/or likelihood of $\mathcal{F}$ evolving in agents of type $\mathcal{A}$.

While the utility of $\mathcal{F}$ for $\mathcal{A}$ or the possibility and/or likelihood of $\mathcal{F}$ evolving in $\mathcal{A}$ are usually the focus of optimization-oriented inquiries, there are classes of biologically inspired questions that are concerned with the relative utility of $\mathcal{F}$ in a given context (e.g., the dynamics of two competing populations of agents, where one kind has $\mathcal{F}$ and the other one does not) or the likelihood of $\mathcal{F}$ *not evolving* in $\mathcal{A}$ (e.g., because having $\mathcal{F}$ might be too costly for agents of type $\mathcal{A}$ relative to the gain in fitness they get based on $\mathcal{F}$). Although experimental results from a set of genetic algorithm runs can be used to support both types of claims (C1) and (C2), they typically cannot be used to support their negations: ($\neg$C1) that $\mathcal{F}$ is *not* beneficial to $\mathcal{A}$ and ($\neg$C2) that $\mathcal{F}$ *cannot* or is *not* likely to evolve in agents of type $\mathcal{A}$—for not having been able to evolve $\mathcal{F}$ in a particular set of genetic algorithm runs does not imply that $\mathcal{F}$ would not have evolved in other runs.

One way to address this problem would be to run genetic algorithm simulations for the full space of initial conditions relevant to $\mathcal{F}$. From the complete set of runs it would then be possible to determine how many times $\mathcal{F}$ evolved and under what circumstances it did not evolve. Obviously, this route is practically infeasible for even small search spaces.

In this paper, we propose an alternative method to support both positive and negative claims about the likely evolvability of $\mathcal{F}$ in $\mathcal{A}$ based on "unscaled and scaled performance spaces". The *unscaled performance space* is the result of systematic evaluations of the performance of $\mathcal{A}$ in a task $\mathcal{T}*$ varying $\mathcal{F}$ (and averaging over initial conditions), which can then be transformed into a *scaled performance space* that can be used to evaluate the "fitness" of $\mathcal{A}$ in a different task $\mathcal{T}$ (e.g., an evolutionary survival task) based on the "cost" associated with different variations of $\mathcal{F}$ and additional parameters specific to $\mathcal{T}$.

The rest of the paper is structured as follows. We start with a more formal discussion of how claims like (C1), (C2), and their negations can be supported by simulation experi-

GECCO'05, June 25–29, 2005, Washington, DC, USA.

ments and introduce the proposed method of using performance spaces to answer questions about the relative benefits of feature $\mathcal{F}$ to $\mathcal{A}$ or about its evolvability in agents of type $\mathcal{A}$. We then demonstrate this methodology using an object collection task, which requires agents to find and collect objects in an artificial environment. First, we show the results for the unscaled performance space, then we discuss how it can be transformed into a scaled performance space, which can be used to answer questions about (1) likely survivors in competitive multi-agent environments with agents differing with respect to $\mathcal{F}$, or (2) evolutionary trajectories between different points in the parameter space induced by $\mathcal{F}$.

2. METHODOLOGY

Claims regarding the evolvability of a given feature $\mathcal{F}$ (or set of features $\mathcal{F}$)[1] for a given task $\mathcal{T}$ and a given agent kind $\mathcal{A}$ can be difficult to evaluate. Typically such statements take the following logical form:

$$\Diamond evolves(\mathcal{F}, \mathcal{A}, \mathcal{T}) \tag{1}$$

where $evolves(\mathcal{F}, \mathcal{A}, \mathcal{T})$ means that $\mathcal{A}$ evolves $\mathcal{F}$ for task $\mathcal{T}$ and $\Diamond\phi$ means that there exists some set of initial conditions such that there exists an evolutionary trajectory to a state in which ϕ is true. To establish the validity of such an *existence claim*, it is sufficient to show that there exists an evolutionary trajectory for agent kind $\mathcal{A}$ from some set of initial conditions that results in agents with $\mathcal{F}$ for the given task $\mathcal{T}$. Typically the existence of initial conditions and trajectories from them leading to the evolution of the target feature are established by the outcomes of runs of genetic algorithms or similar evolutionary computational tools.

The demonstration of (1) by itself, however, does not say anything about its *likelihood* (i.e., whether the initial conditions and subsequent evolutionary trajectories are likely to obtain). To establish "likely evolution" a stronger argument is required. In the limit case, (1) can be thought of as the *modal dual* to what might be called the "inevitable evolution" of $\mathcal{F}$:

$$\Box evolves(\mathcal{F}, \mathcal{A}, \mathcal{T}) \tag{2}$$

where $\Box\phi$ means that every evolutionary trajectory from every set of initial conditions leads to a state in which ϕ is true. Since logical necessity implies certainty, (2) states that *$\mathcal{F}$ will evolve from any initial condition.*

Given its logical strength, (2) will be in general very difficult to establish (only for extremely limited domains will it be feasible to investigate all possible evolutionary trajectories). Yet, for practical purposes, a logically weaker probabilistic formulation is often sufficient:

$$P(evolves(\mathcal{F}, \mathcal{A}, \mathcal{T}) | D_{IC}) > \theta \tag{3}$$

where $P(p|q)$ is the probability of p given q, D_{IC} is the distribution of initial conditions IC, and θ is a certain threshold value in [0,1]. For θ close to 1, (3) states that $\mathcal{F}$ is *very likely to evolve* in agents $\mathcal{A}$ for task $\mathcal{T}$. But even the demonstration of (3) might be practically infeasible due to the sheer size of IC. Consequently, a different strategy might be necessary to provide convincing arguments for claims like (C2) or ($\neg$C2).

The direction we will pursue in this paper is to evaluate $\mathcal{A}$'s performance in a different, but related task $\mathcal{T}^*$ (in the same environment) for a set of variations of $\mathcal{F}$ and use the results of this evaluation to make informed inferences about $\mathcal{A}$'s performance in $\mathcal{T}$. The approach is based on the following three observations:

(O_1) Variations in (physical) features of agents (such as the speed of movement or the "sensory range" within which the agent's sensors can detect stimuli) are typically gradual and continuous.[2]

(O_2) Variations of these features typically have clearly established (physically determined) lower and upper boundaries (e.g., there are minimum and maximum speeds at which a given biped can walk).

(O_3) If $\mathcal{A}$'s reproductive success in an evolutionary task $\mathcal{T}$ can be predicted (in a statistical sense) based on its performance in a task $\mathcal{T}^*$, then $\mathcal{A}$'s performance for different variations of $\mathcal{F}$ in $\mathcal{T}^*$ will be a predictor for $\mathcal{F}$ in $\mathcal{T}$.

Observations (O_1) and (O_2) guarantee that the set of variations of $\mathcal{F}$ is bounded and fairly well-behaved within its boundaries $[\mathcal{F}_{low}, \mathcal{F}_{high}]$. While the exact nature of $\mathcal{A}$'s performance in $\mathcal{T}^*$ within the variations of $\mathcal{F}$ is unknown, it is possible to compute an approximation of the performance function that is sufficient for the step in observation (O_3). Specifically, the performance of $\mathcal{A}$ in $\mathcal{T}^*$ can be evaluated at each sample point in $[\mathcal{F}_{low}, \mathcal{F}_{high}]$ at a particular spatial sampling frequency ϕ (dependent on the variability of $\mathcal{F}$) averaging over a randomly drawn subset S_{IC} of initial conditions from D_{IC}—the result is the "(unscaled) performance space" $\mathcal{P}^{[\mathcal{F}_{low}, \mathcal{F}_{high}]}_{\mathcal{T}*, \mathcal{A}, \mathcal{F}, \phi, S_{IC}}$. Since the sample space will be a good approximation of the true space $\mathcal{P}_{\mathcal{T}, \mathcal{A}, \mathcal{F}, D_{IC}}$ (for sufficiently small ϕ dependent on the nature of $\mathcal{F}$), it can be used to predict $\mathcal{A}$'s performance in $\mathcal{T}$ based on (O_3) (i.e., on the fact that performance in $\mathcal{T}^*$ is a predictor for performance in $\mathcal{T}$).

It is now possible to make predictions about the utility of a particular value for $\mathcal{F}$ in line with (C1) and ($\neg$C1) from above. In particular, it is possible to predict the performance of two agent kinds $\mathcal{A}_{\mathcal{F}_i}$ and $\mathcal{A}_{\mathcal{F}_j}$ that differ with respect to $\mathcal{F}$ in direct competition in an environment by comparing their respective performances in $\mathcal{P}^{[\mathcal{F}_{low}, \mathcal{F}_{high}]}_{\mathcal{T}*, \mathcal{A}, \mathcal{F}, \phi, S_{IC}}$: the agent kind with the higher performance in $\mathcal{T}^*$ is likely to have higher performance in $\mathcal{T}$. More precisely, the *null hypothesis* stating that there is no performance difference between $\mathcal{A}_i$ (with feature $\mathcal{F}_i$) and $\mathcal{A}_j$ (with feature $\mathcal{F}_j$) will have to be rejected based on the significance of their performance difference as measured by a T-test, for example:

[1]Everything said about a single feature $\mathcal{F}$ also applies to a set of features, hence we will use "$\mathcal{F}$" to denote both a single feature as well as a set of features. Moreover, we will also use "$\mathcal{F}$" as a variable representing the range of variation of feature $\mathcal{F}$.

[2]Note that while this observation is typical of phenotypes, it is not in general true of genotypes as there may be many very different encodings of the same phenotypical features. Fortunately, our methodology is only concerned with phenotypes, since phenotypes and not genotypes are subject to performance evaluation and adaptation.

$$T\text{-}test(\mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}(\mathcal{F}_i) = \mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}(\mathcal{F}_j)) < 0.05$$

In a similar vein, the p-value of the significance test can be used to create an upper bound on the conditional probability that a feature $\mathcal{F}$ will or will not evolve in agents $\mathcal{A}$ for task $\mathcal{T}$, in line with (C2) and (¬C2) from above. For example, for (¬C2) the *null hypothesis* would be that the evolved agent's fitness is higher that that of the evolving agent:

$$P(evolves(\mathcal{F},\mathcal{A},\mathcal{T})|D_{IC}) < \quad (4)$$
$$T\text{-}test(\mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}(\mathcal{F}) > \mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}(\neg\mathcal{F})))$$

where $\neg\mathcal{F}$ indicates the absence of feature $\mathcal{F}$ in agent $\mathcal{A}_{\neg\mathcal{F}}$.

The exact numeric probabilities will usually depend on several factors, but most importantly on the "accuracy" of $\mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}$ and the degree to which performance in $\mathcal{T}$ is predicted by performance in $\mathcal{T}*$. For example, it will be often possible to improve the prediction by scaling the performance space $\mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}*,\mathcal{A},\mathcal{F},\phi,S_{IC}}$ via a "cost function" f that takes peculiarities of $\mathcal{T}$ into account that are not accounted for by $\mathcal{T}*$ alone.[3] The resultant space (obtained by applying f to all points in the performance space) is called a "scaled performance space".

In the following, we will illustrate the above methodology and demonstrate its utility for evolutionary investigations in artificial life settings. Specifically, we will use the performance of agents in an "object collection task" $\mathcal{T}*$ to predict their performance in (1) a generational survivability task $\mathcal{T}_1$, where agents $\mathcal{A}_{\mathcal{F}_i}$ (for different $\mathcal{F}_i$) need to survive in homogeneous and heterogeneous environments, and (2) an evolutionary adaptation task, where the question at hand concerns the likelihood of evolving feature $\mathcal{F}_{dest}$ starting from feature $\mathcal{F}_{orig}$ in a population of agents $\mathcal{A}_{\mathcal{F}_{orig}}$.

3. PREDICTING PERFORMANCE

Investigations of population dynamics or evolutionary trajectories are based on the performance of agents in a task $\mathcal{T}$, typically related to the survivability of the agents (e.g., $\mathcal{T}$ could be a foraging task, in which agents have to find the resources they need to survive and procreate). Hence, for the above methodology to apply, a task $\mathcal{T}*$ is needed such that the performance in $\mathcal{T}*$ generally predicts the performance in $\mathcal{T}$. Note that since $\mathcal{T}*$ will have to be evaluated for different variations of $\mathcal{F}$, it is important that $\mathcal{T}*$ be computationally tractable (and preferably inexpensive to allow for a fine-grained sampling ϕ of the whole interval $[\mathcal{F}_{low},\mathcal{F}_{high}]$).

One way to find such a $\mathcal{T}*$ is to find a measure that correlates well with overall performance in $\mathcal{T}$ (e.g., because it measures a subcomponent of $\mathcal{T}$) and then devise a simpler task $\mathcal{T}*$ that approximately measures it. For example, in a survival study where foraging for resources is critical for survival and procreation, the efficiency with which agents forage (e.g., as defined in terms of "energy consumption per time unit") is a measure of success. Consequently, any task that measures "foraging efficiency" well should also predict $\mathcal{T}$ well. In evolutionary studies using genetic algorithms, the employed evaluation function to determine the "fitness" of agents can often be directly used.

The exact relationship between the performance measure for $\mathcal{T}*$ and the performance measure for $\mathcal{T}$ must be clear in order for the scaled performance space to have predictive power. Often a simple task $\mathcal{T}*$ will not sufficiently capture additional factors in $\mathcal{T}$ for it to be effectively predictive. For example, the cost of movement at a particular speed in $\mathcal{T}$ might not be factored into $\mathcal{T}*$. Yet, this cost might be critical in determining in the overall performance in $\mathcal{T}$ (as shown in the next section). Hence, a transformation function (which often will be non-linear) needs to be formulated that scales performance in $\mathcal{T}*$ based on $\mathcal{F}$ so as to take these additional factors into account and establish a better prediction.

Finally, it is essential to select only tasks $\mathcal{T}*$ that allow for the full variation of $\mathcal{F}$ as otherwise the performance space cannot be determined (e.g., fitness functions directly taken from experiments with genetic algorithms might not explicitly incorporate $\mathcal{F}$). For example, if the goal of the study is to demonstrate that a population of agents will evolve a particular combination of speed of motion ($\mathcal{F}_1$) and sensory range ($\mathcal{F}_2$), the performance space will be represented by a two-dimensional matrix, where the number of rows and columns will be determined by the granularity of sampling for different speeds ϕ_1 and sensory ranges ϕ_2.

In the following we will consider two versions of a biologically inspired survival task in a continuous 2D environment, in which agents need to gather resources to survive and procreate. Agents have sensors that allow them to detect energy sources and have effectors that allow them to move through the environment. Different sensory ranges and different speeds are possible for different agents, which have to pay at each update cycle the cost associated with their sensory range and speed.

In this setup, the first kind of investigation, called "generational studies," is concerned with the dynamics of homogeneous and heterogeneous agent populations. Typical questions that arise in such studies are: (1) will a given agent kind (i.e., a kind with a given speed and sensory range) survive for a fixed number of cycles, given initial conditions, distribution and capacity of energy sources, and energy influx; or (2) will a particular agent kind be better than another agent kind as measured in terms of average number of survivors if put in competition within the same environment.

The second kind of investigation, called "evolutionary studies," is concerned with evolutionary trajectories from given conditions and whether particular traits or features of agents will evolve. Typical questions are: (1) what kinds of agents will evolve from a given agent kind; or (2) will a particular trait always evolve given an initial random distribution of traits (taken from a subset of traits not including the target trait).

We will demonstrate that the proposed methodology can be used answer questions like the above for both kinds of investigations by virtue of empirically determining the performance space for a *collection task*, in which multiple agents work together (although not cooperatively) to collect all objects in an environment. All simulation experiments described later were conducted using the SWAGES artificial life simulation environments.[4]

[3]We call this a "cost function" because in a biological setting there are typically costs in terms of required energy involved for variations in $\mathcal{F}$.

[4]SWAGES is a flexible agent-based artificial life simulation experimentation environment that consists of several

We will first introduce the collection task and then describe the setup in the generational and evolutionary studies, and their relation to the collection task, in more detail.

3.1 The Collection Task

The collection task consists of $|\mathcal{A}|$ agents of type $\mathcal{A}$ randomly placed in an environment and C items, also randomly placed.[5] Agents travel at a fixed speed s, and can sense items within sensory range r in a 360° radius. The object of the task is to collect as many items as possible within a given time frame.

The agents employed in the collection task are simple reactive agents that explore the environment searching for objects to collect. When an item is detected, the agent moves directly to it and collects it. When no item is detected, the agent performs a random walk, moving straight ahead for W cycles, at which point it makes a random turn. More formally, the agents implement the following rules:

- *Rule 1:* if no object is perceived and walk counter w is less than W, increment w and move straight ahead
- *Rule 2:* if no object is perceived and walk counter $w = W$, reset w to 0 and turn 1-45 degrees in either direction
- *Rule 3:* if at least one object is perceived, reset w to 0 and go directly toward the closest one
- *Rule 4:* if some object is within collection distance, collect it

These rules constitute the model for the "collection agents" used in the performance space experiments.

The performance space is a represented by a 10x20 matrix of performance evaluations. Each of the 200 points in performance space represents the performance of a particular agent configuration with respect to two features: $\mathcal{F}_1$=speed (s) and $\mathcal{F}_2$=sensory range (r) . Speed ranges from 1 ($\mathcal{F}_{1,low}$) to 10 ($\mathcal{F}_{1,high}$) in steps of one (ϕ_1), while sensory range goes from 25 ($\mathcal{F}_{2,low}$) to 500 ($\mathcal{F}_{2,high}$) in steps of 25 (ϕ_2). With the exception of these two parameters, the agent configurations are identical. We denote a particular agent configuration $\mathcal{A}$ with speed s and sensory range r by $\mathcal{A}_{s,r}$.

The performance of $\mathcal{A}_{s,r}$ in the collection task is a measure of the efficiency with which $\mathcal{A}_{s,r}$ can collect items. This measure "collection efficiency" can be very directly related to "foraging efficiency", the measure of long-term success in the generational and evolutionary studies, by virtue of including the cost associated with foraging (which is absent in the collection task). We assume the following cost model:

$$Cost_{\mathcal{A}_{s,r}} = Base_{\mathcal{A}_{s,r}} + Speed^{phys}_{\mathcal{A}_{s,r}} + Speed^{act}_{\mathcal{A}_{s,r}} + \left(\frac{r}{c}\right)^2 \quad (5)$$

where $Base_{\mathcal{A}_{s,r}} = 10$ is the base cost for agent agent per update cycle (which is the same for all s and r), $Speed^{phys}_{\mathcal{A}_{s,r}}$ is the cost of maintaining the physical components required to travel at speed s, and $Speed^{act}_{\mathcal{A}_{s,r}}$ is the cost of the agent's actual speed, either s^2 when the agent is moving or 0 when it collects an item. Finally, agents pay a quadratic cost based on their sensory range at every cycle scaled by $c = 25$ (which numerically maps sensory ranges into cost ranges).[6]

With the performance space and the cost function in hand, it is now possible to create the scaled performance space. The scaled performance space yields a measure of relative performance where the costs for movement and sensing are taken into account. Relative performance can be thought of as a measure of how many objects were collected for each unit of energy spent and can, therefore, be used make predictions about experiments, where energy efficient search for energy sources is critical.

3.2 Generational Studies

As a first step toward validating our approach, we use the relative performance space to predict the outcomes of biologically inspired generational studies, in which agents–as mentioned above–are required to forage through their environments to find food. Collecting food items is the only means agents have to replenish and build their energy stores; agents that perform the foraging task poorly simply die. Agents reproduce once they have acquired sufficient resources. The agent model for the generational studies is identical to the model for the collection studies, with the following addition:

- *Rule 5:* if $E >= ProcEner$, produce offspring

where E is the agent's energy level and $ProcEner$ is the energy threshold for procreation. Thus, when the agent has stored enough energy, it will always procreate and an identical copy of the agent will be placed in a random near-by location.

The goal of all agents in generational studies is to survive and procreate. The measure of performance is the number of agents alive at the end of the simulation run (i.e., after a fixed number of cycles). We predict that agent configurations with high relative performance on the collection task will perform well in the generational task, too. This is because these agents are better foragers for the cost than other agent configurations with lower relative performance–they forage more efficiently and, therefore, spend less energy per resource gathered than agents with other configurations. Note that performance on the generational task is already "relativized"—because cost is assessed as the simulation progresses, there is no need to take it into account again.

3.3 Evolutionary Studies

Evolutionary studies are identical to generational studies with two exceptions. First, evolutionary studies employ an evolutionary mechanism to change traits of agents, whereas in generational studies agent configurations are fixed. Second, evolutionary studies are typically longer than generational studies because of the large number of generations

components, most importantly (1) a simulation component "Simworld", which gives it the flexibility of designing agents that can vary greatly in complexity, from simple reactive agents, to highly complex cognitive agents, and (2) an experiment scheduler that can schedule simulation experiments in heterogeneous computing environments (without the need for a separate grid engine) and supervise their timely execution. SWAGES is freely available at http://www.nd.edu/~airolab/software/.

[5]Note that we use $\mathcal{A}$ both to denote the agent type as well as a set of agents of that type.

[6]These are conservative assumptions about the real increases of energy requirements based on increases in speed and sensory range.

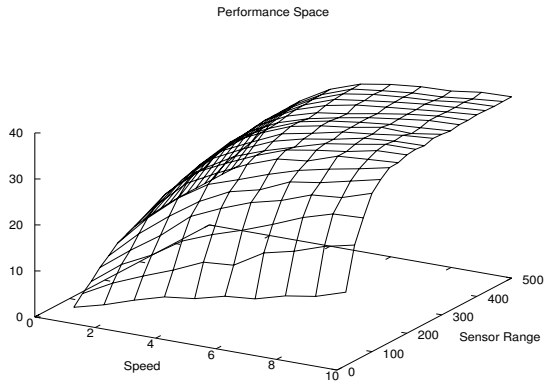

Figure 1: Performance results for the collection task, speed from 1 to 10, sensory range from 25 to 500.

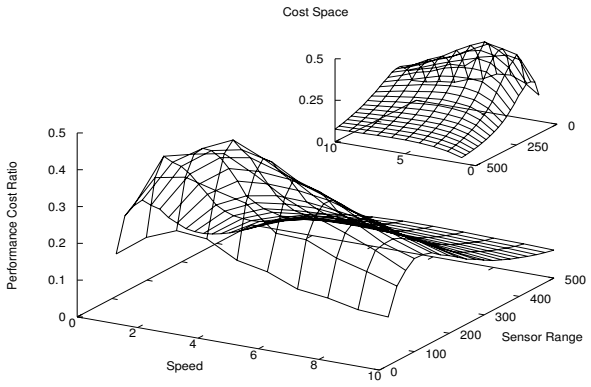

Figure 2: Relative performance on the collection task, speed from 1 to 10, sensory range from 25 to 500. The embedded graph depicts the same data with the x and y axes reversed for perspective.

required to make substantial progress along an evolutionary trajectory. The evolutionary mechanism employed below is simple mutation: offspring have a fixed probability of having both inherited features $\mathcal{F}_1$ (speed) and $\mathcal{F}_2$ (sensory range) modified. We assume a fixed probability for mutation throughout the simulation and two mutation operations for the two features: speed can be changed by ± 1, while sensory range can be changed by ± 25. We predict that agent configurations with low scaled performance figures in the collection task will tend to evolve toward configurations with higher scaled performance; agents with lower scaled performance will be less likely to survive and reproduce (as will be demonstrated in the generational studies), so agents that mutate in that direction are less likely to have successful offspring, whereas agents that mutate to configurations with higher relative performance on the collection task will be more likely to survive and reproduce.

4. COLLECTION STUDIES

The collection studies consist of 200 40-simulation experiments for values of s from 1 to 10 and r from 25 to 500. In each group, the same set of 40 initial conditions is used, and only agent parameters are varied (i.e., agent and food locations in experimental run 28 are initially the same across all experiments, allowing us to compare directly between experiments; the differences in outcomes are due only to differences in agent parameters). Each simulation begins with 5 agents ($|\mathcal{A}| = 5$) and 40 items ($|C| = 40$). Agents do not incur costs during the course of the simulation, and they cannot die. The performance measure selected is the number of items collected at the end of 500 simulation cycles.

Figure 1 presents the results of the collection experiments. Predictably, agents with high speed and sensory range collected the most items by the end of the simulation. There is a large region in performance space in which all combinations collect nearly all 40 items within the time allotted. The poorest performers are those that can perceive little and move through the environment slowly.

The space in Figure 2 depicts the scaled performance of each configuration (scaled by the cost function (5) from the previous section). Here, the main figure is from the same perspective as Figure 1, demonstrating graphically that the agents with the best scaled performance are those with medium to low speed and low sensory range. The configurations with the highest unscaled performance are among the worst scaled performers. The configuration with the best scaled performance is $s = 4, r = 100$, with a scaled performance value of 0.49; its unscaled performance ranking was 151 out of 200 (22.475 items collected on average). Conversely, the configuration with the best unscaled performance ($s = 10, r = 400$) collected an average of 39.5 items, but scores only a 0.11 scaled performance, ranking it 150 out of 200.

We turn next to the generational experiments, in which we expect the configurations around [4,100] to perform very well compared to other configurations.

5. GENERATIONAL STUDIES

As mentioned before, the results of the collection studies can be used to answer various kinds of questions in *generational studies*, since scaled performance in the collection task is a measure of foraging efficiency. More formally, we can define the average "collection efficiency" $CollEff_{s,r}$[7] of an agent kind $\mathcal{A}_{s,r}$ with speed s and sensory range r in the collection task $\mathcal{T}_c*$ as

$$CollEff_{s,r} = \mathcal{P}^{[\mathcal{F}_{low},\mathcal{F}_{high}]}_{\mathcal{T}_c*,\mathcal{A}_{s,r},\mathcal{F},\phi,S_{IC}} / (|\mathcal{A}_{s,r}| \times Cycles_{\mathcal{T}_c*}) \quad (6)$$

Given $CollEff_{s,r}$ from the collection task, we can get an estimate for the amount of energy consumed on average, the average "energy consumption" $EnerCons_{s,r}$, by an agent $\mathcal{A}_{s,r}$ between food sources in the foraging task based on the cost function (5): $EnerCons_{s,r} = Cost_{s,r}/CollEff_{s,r}$. Note that $EnerCons_{s,r}$ is a lower bound on the "food energy" ($FoodEner$), i.e., how much energy each food source should provide to the agent. If $FoodEner = EnerCons_{s,r}$ and food sources are generated at least at the frequency $CollEff_{s,r}$, then the environment should be able to sustain one agent.

[7]For legibility, we only include subscripts for speed and sensory range here.

Experiment		[4,100] Agents		[2,200] Agents	
		Mean	Conf.	Mean	Conf.
$FoodEner = 5628.48$	[4,100]	4.80	2.12		
	[2,200]			1.23	0.89
	[4,100] & [2,200]	1.68	1.20	0.00	0.00
$FoodEner = 12127.5$	[4,100]	5.20	3.06		
	[2,200]			4.40	1.67
	[4,100] & [2,200]	5.43	2.76	0.00	0.00

Table 1: Experimental results of generational studies. The top rows are the average performance and confidence intervals in $FoodEner = 5628.48, FoodGenRate = 0.09$ environments for homogeneous [4,100], homogeneous [2,200] and heterogeneous [4,100] & [2,200] experiments. The bottom rows are analogous results for $FoodEner = 12127.5, FoodGenRate = 0.085$ environments.

The "food generation rate" ($FoodGenRate$) for n agents must be at least $n \times CollEff_{s,r}$.

Since agents can reproduce in generational studies, they have to transfer some of their energy to their offspring and lose some as overhead of procreation. We can use the results of the collection studies to define values for the "initial energy" ($IE_{s,r}$) shared with the offspring, as well as the amount of energy an agent must have in order to reproduce (i.e., the "procreation energy threshold", or $ProcEner_{s,r}$). When an agent is initially created, it must have enough energy to survive until its first energy intake. Given the average amount of energy required for an agent to survive between energy intakes, $EnerCons_{S,R}$, this value can be used for $IE_{s,r}$, ensuring that, on average, offspring will survive until they find their first food source. Parents must also be left with enough energy to find another food source, so $ProcEner_{s,r}$ must be sufficiently high so as to leave the parent with at least $EnerCons_{s,r}$ units of energy. The overhead of reproduction is (arbitrarily) fixed at 25%, hence $ProcEner_{s,r} = EnerCons_{s,r} + 1.25 \cdot EnerCons_{s,r}$. When an agent's energy level reaches $ProcEner$ it creates offspring.

Each simulation run begins with a fixed number of food sources (40) randomly placed throughout the environment. The simulation is tailored to the agents we expect to survive in it. Thus, additional food sources are added with a probability of $FoodGenRate = CollEff_{s,r}$ per cycle (i.e., on average every $\frac{1}{FoodGenRate}$ cycles) in random locations. The amount of energy contained in each of these food sources is $FoodEner = EnerCons_{s,r}$. Agents start with energy level equal to $FoodEner$.

We first apply the procedure to the best agent in the collection studies. With $s = 4$ and $r = 100$, the $\mathcal{A}_{4,100}$ agents' cost per cycle is 46 (10 + 4 + 16 + 16, by equation (5)). These agents collected an average of 22.475 items in 500 cycles. Thus, the energy needed per food item ($FoodEner$) must be at least 5116.8 units. $FoodGenRate$ for ten agents must be at least $10 \cdot CollEff_{4,100} = 0.0899$. For the experiments below, we multiply $EnerCons_{4,100}$ by 1.1 (a margin of safety for the agents) to yield $FoodEner = 5628.48$ and round $FoodGenRate$ to 0.09. Given these values, we obtain $IE_{4,100} = 5628.48$ and $ProcEner_{4,100} = 12664.08$.

The second agent configuration we will consider is $s = 2$ and $r = 200$; this agent kind placed 57th in scaled performance, better than almost 75% of all configurations. Calculating the energy requirements of $\mathcal{A}_{2,200}$ agents in the same way, we find that $EnerCons_{2,200} = 11025$ and $CollEff_{2,200} = 0.00839$. These are the minimum values we would expect to need in order for $\mathcal{A}_{2,200}$ agents to survive, and again we add a 10% buffer, making $IE_{2,200} = 12127.5$ and $ProcEner_{2,200} = 27286.875$.

The first set of experiments were conducted in environments where the food energy $FoodEner = 5628.48$. Based on their scaled performances in the collection task, we expect that $\mathcal{A}_{4,100}$ agents will be able to survive in these conditions, while $\mathcal{A}_{2,200}$ agents, which require much more energy according to our calculations, will not do as well. Three experiments were conducted, one with 5 initial $\mathcal{A}_{4,100}$ agents, one with 5 initial $\mathcal{A}_{2,200}$ agents, and one which started with 5 of each. The upper portion of Table 1 presents the results of these three studies. As predicted, $\mathcal{A}_{4,100}$ agents were able to survive in this environment, but only 4.80 agents were alive on average at the end of 10,000 cycles. The $\mathcal{A}_{2,200}$ agent configuration was able to survive, with an average of 1.23 agents alive at the end of the experimental runs, however significance tests show the $\mathcal{A}_{4,100}$ agents' performance to be significantly better than the $\mathcal{A}_{2,200}$ agents'.

Furthermore, $\mathcal{A}_{4,100}$ agents had an advantage in heterogeneous environments where they competed against $\mathcal{A}_{2,200}$ agents. An average of 1.68 $\mathcal{A}_{4,100}$ agents survived at the end, whereas no $\mathcal{A}_{2,200}$ agents were alive in any simulation run. Although fewer $\mathcal{A}_{4,100}$ agents survived on average in the heterogeneous environments, the difference in survivability between the two agent kinds is statistically significant.

To be fair to $\mathcal{A}_{2,200}$ agents, we also conducted these tests in environments with $FoodEner = 12127.5$, rounding up from $EnerCons_{2,200}$, and $FoodGenRate = 0.85$. This should allow $\mathcal{A}_{2,200}$ agents to do well in heterogeneous environments. Yet, it should also be an advantage to $\mathcal{A}_{4,100}$ agents, so we predict that they will win again in heterogeneous environments. The bottom rows of Table 1 confirm this. $\mathcal{A}_{2,200}$ agents are able to survive when alone in the environment, with an average of 4.40 agents alive at the end of the simulation. However, $\mathcal{A}_{4,100}$ agents also boost their performance, with 5.20 surviving on average in homogeneous environments. Although more $\mathcal{A}_{4,100}$ agents survive than $\mathcal{A}_{2,200}$ agents, the difference is not significant. However, when placed in the same environment, $\mathcal{A}_{4,100}$ agents again dominate, with an average of 5.43 survivors, while $\mathcal{A}_{2,200}$ agents again failed to survive to the end of any experimental run. Again, this difference is statistically significant.

The results of the generational studies are encouraging. The predictions we made based on the collection studies were proved correct. This is of great interest, because the number of experiments performed to obtain the scaled per-

Exp. Run	Speed		Range		Agents Alive	Rank
	Mean	Conf.	Mean	Conf.		
2	4.00	0.00	75.0	0.00	48	2
3	4.00	0.00	100.0	0.00	41	1
14	6.00	0.00	100.0	0.00	21	12
16	3.94	0.09	75.0	0.00	32	2
21	6.05	0.11	150.0	0.00	19	23
24	6.00	0.00	204.17	6.12	31	40
29	4.00	0.00	72.40	2.24	23	2
30	4.20	0.17	100.00	0.00	31	1
40	5.00	0.00	125.00	0.00	14	8

Table 2: Average speed and range for surviving agents at the end of 100,000 simulation cycles.

formance space (200) is dramatically less than the number of combinations one would have to run to test each configuration against each other (199000). Using the scaled performance space, we can make reasonably accurate predictions without paying the cost of exploring the whole space of combinations.

We turn now to the evolutionary studies, to see whether predictions based on the scaled performance space are valid for the evolutionary task.

6. EVOLUTIONARY STUDIES

The results of the collection studies led us to predict that agents will evolve to the point $\mathcal{A}_{4,100}$ in the trait space. This is the highest point in scaled performance space for all configurations in the collection study (see Figure 2). The simulations for the evolutionary experiment are the same as those for the generational studies, except that mutation is added as an evolutionary mechanism. The values of $FoodEner$ and $FoodGenRate$ are arrived at the same way: $FoodEner = EnerCons_{s,r}$, and $FoodGenRate = n \cdot CollEff_{s,r}$, where n is the number of agents we want the environment to sustain.

Each experimental run begins with 5 agents with $s = 2$ and $r = 200$ (this is the same configuration tested in the generational studies). $FoodEner = 12127.5$, as in the second set of generational studies, because we want the environment to be able to sustain $\mathcal{A}_{2,200}$ agents. For the evolutionary studies, however, we set $FoodGenRate = 0.15$ because we want the environment to be able to sustain more agents as insurance against population crashes. Also, whereas the generational studies were carried on for 10,000 simulation cycles, the evolutionary study extends the simulations to 100,000 cycles to allow sufficient time for mutation to search the trait space.

The mutation rate M employed in these experiments is 0.01. Thus, agents have a 0.01 probability of being born with a different base speed than their parents, and a 0.01 probability of being born with a different sensory range than their parents. As mentioned before, mutation can increase or decrease speed and/or sensory range in discrete steps within the predetermined speed and sensory range limits, i.e., speed can mutate one point in either direction in the interval [1,10]. Range can mutate in increments of 25 in either direction in the interval [25,500]. Procreation is asexual, hence crossover (or any other evolutionary operators that presuppose two genotypes) is not applicable; the parent's parameters are passed directly to the offspring when no mutation is present.

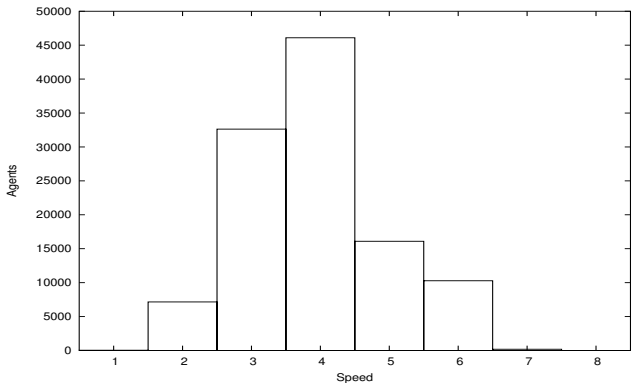

Figure 3: Number of agents with each speed throughout the full histories of simulation runs with surviving agents (speeds not depicted were not possessed by any agent).

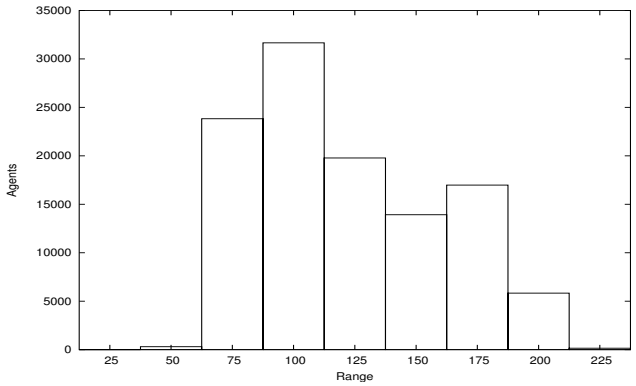

Figure 4: Number of agents with each sensory range throughout the full histories of simulation runs with surviving agents (ranges not depicted were not possessed by any agent).

The results of the evolutionary study validate our prediction. Table 2 presents a summary of simulation runs with surviving agents (in 31 of the 40 experimental runs agents failed to survive despite our increasing $FoodGenRate$ to 0.15). The overall weighted average speed for all survivors is 4.62, and the weighted average range is 107.29. The nearest point in the trait space to these averages is [5,100]. This is very close to the predicted final state (only one mutation off, and ranked third in the scaled performance space), and, in fact, there was no statistically significant difference between the performance of $\mathcal{A}_{4,100}$ agents on the collection task and the performance of $\mathcal{A}_{4,100}$ agents. However, there are some individual runs that are much further away. Experimental run 24 is nearest to [6,200] in the trait space, which is ranked 40 in scaled performance on the collection task. This is an improvement over the initial state (recall that $\mathcal{A}_{2,200}$ was 57), but not very close to the predicted final state. Figures 3 and 4 depict the frequency at which agents possessed each possible speed and sensory range value for all agents in simulations with survivors (i.e., not just the survivors). In both cases, the large majority of agents are within one step of the predicted values ($\mathcal{A}_{4,100}$). This confirms the general

trend; with more time, populations like experimental run 24 will continue to move toward the predicted values.

7. DISCUSSION AND CONCLUSION

Evolutionary investigations are often very expensive in terms of the required computational resources and thus are often limited to "existence proofs", i.e., the demonstration that a particular feature $\mathcal{F}$ can be evolved in a set of initial conditions. Many general questions regarding the utility of $\mathcal{F}$ (e.g., in competitive environments) or the likelihood of $\mathcal{F}$ evolving or not evolving are therefore typically difficult, if not practically infeasible to answer. We have proposed a methodology that allows us to answer such questions in setups where good predictors of task performance $\mathcal{T}$ are available. These predictors evaluate the performance of an agent kind $\mathcal{A}$ in a task $\mathcal{T}^*$, which can then be transformed by including costs and additional factors to make predictions about the performance of $\mathcal{A}$ in $\mathcal{T}$.

In the collection task $\mathcal{T}^*$ used for prediction of population dynamics and evolutionary trajectories, the performance evaluation amounts to the determination of the *apriori* probability $P_{\mathcal{A}_{s,r}}$ that an agent $\mathcal{A}_{s,r}$ with speed s and sensory range r will get an item in one simulation cycle. Thus, $P_{\mathcal{A}_{s,r}}$ can be seen to be a fitness measure of $\mathcal{A}_{s,r}$: the higher $P_{\mathcal{A}_{s,r}}$, the more items an agent of type $\mathcal{A}_{s,r}$ will be able to collect in a given time. Note, however, that $P_{\mathcal{A}_{s,r}}$ depends on the total number of agents participating in $\mathcal{T}^*$: everything else being equal, we get that $P^n_{\mathcal{A}_{s,r}} > P^{n+1}_{\mathcal{A}_{s,r}}$ for all $n = |\mathcal{A}_{s,r}|$. What warrants our inference from a single performance evaluation for a fixed group size to all group sizes in generational and evolutionary studies comparing heterogeneous agent environments, is the additional assumption that if $P^n_{\mathcal{A}_{s,r}} > P^n_{\mathcal{A}_{s',r'}}$, then $P^{n+1}_{\mathcal{A}_{s,r}} > P^{n+1}_{\mathcal{A}_{s',r'}}$ for all n, $\mathcal{A}_{s,r}$, and $\mathcal{A}_{s',r'}$. That is, the relative fitness of two agent kinds with regard to the performance evaluation does not change based on group size or environmental configurations. In the above cases, this principle is valid based on the rules that define a "collection agent": each agent follows a greedy foraging strategy that does not take the presence or strategies of other agents into account (i.e., agents of one kind do not discriminate among agent kinds, which otherwise might be major confounding factor that can significantly change agent performance [11], thereby making predictions based on single agent evaluations very difficult, if not impossible). If agents were to alter their behavior based on the number of other agents they can perceive, this principle is likely not to hold (e.g., suppose agents stop moving forever when too many other agents are around–in that case their probability of collecting an item goes to zero). Hence, it is critical for our approach to either establish the functional independence of agent functions (as is the case in our setup) or to show possible group size effects do not significantly influence fitness as captured by $P^n_{\mathcal{A}_{s,r}}$.

There are other limitations to the proposed method, the most obvious of which being cases where the best (known) predictor of task performance is *the task itself*. Less obvious limitations are imposed by interactions between factors (in addition to agent-agent interactions) that do not enter the performance evaluation, but play a crucial role in the evolutionary studies. For example, whether an offspring is placed in the vicinity of a parent or in a random location in the environment might not make a difference with respect to fitness in generational studies, but could potentially change the overall outcomes of evolutionary studies (as in one case parents will have to compete with their close offspring for resources, which happens only infrequently in the other). Finally, even the computation of the performance space for $\mathcal{T}^*$ might not be feasible if too many parameters have to be varied or a very high spatial frequency of sampling is required. Yet, in those cases it is likely that even regular evolutionary methods (e.g., genetic algorithms) will fail to produce reasonable results.

We believe that despite its limitations the proposed methodology can be of great utility to many investigations in the fields of artificial life and adaptive behavior, where the three assumptions of the methodology are often met (e.g., as in the case of the demonstrated generational and evolutionary studies). If nothing else, it can be viewed as an attempt to lay out a formal argument structure that can support general claims about the utility of features in dynamical interactions among agents or the likelihood of evolving (or failing to evolve) a feature.

8. REFERENCES

[1] A. D. Channon and R. I. Damper. Evolving novel behaviors via natural selection. In *Proc. Artificial Life VI*, pages 384–388, 1998.

[2] O. Holland and D. McFarland. *Artificial Ethology*. Oxford University Press, Oxford, 2001.

[3] M. Levin. The evolution of understanding: A genetic algorithm model of the evolution of communication. *BioSystems*, 35:167–178, 1995.

[4] B. MacLennan. Synthetic ethology: An approach to the study of communication. In *Artificial Life II: Proceedings of the Second Workshop on Artificial Life*, pages 631–658, 1991.

[5] A. Mark, D. Polani, and T. Uthmann. A framework for sensor evolution in a population of Braitenberg vehicle-like agents. In *Proc. Artificial Life VI*, 1998.

[6] J. H. Miller, C. T. Butts, and D. Rode. Communication and cooperation. *Journal of Economic Behavior and Organization*, 47:179–195, 2002.

[7] J. Noble. Cooperation, conflict and the evolution of communication. *Adaptive Behavior*, 7(3/4):349–370, 1999.

[8] M. Quinn. Evolving communication without dedicated communicatin channels. In *Proceedings of ECAL 2001*, pages 357–366, 2001.

[9] T. S. Ray. An approach to the synthesis of life. In *Artificial Life II*. Addison-Wesley, 1991.

[10] T. S. Ray. An evolutionary approach to synthetic biology: Zen and the art of creating life. In *Artificial Life*, volume 1, pages 195–226. MIT Press, 1994.

[11] M. Scheutz and P. Schermerhorn. The more radical, the better: Investigating the utility of aggression in the competition among different agent kinds. In *From Animals to Animats 8: Proceedings of Simulation of Adaptive Behavior 2004*. MIT Press, 2004.

[12] K. Sims. Evolving 3d morphology and behavior by competition. In *Proc. Artificial Life IV*, 1994.

[13] N. Zaera, D. Cliff, and J. Bruten. (Not) evolving collective behaviours in synthetic fish. In *Proc. SAB96*, 1996.