Identifying Reusable Early-Life Options

Aline Weber
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
aweber@inf.ufrgs.br

Charles P. Martin, Jim Torresen
Department of Informatics
University of Oslo
Oslo, Norway
{charlepm, jimtoer}@ifi.uio.no

Bruno C. da Silva
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
bsilva@inf.ufrgs.br

***Abstract*—We introduce a method for identifying short-duration reusable motor behaviors, which we call *early-life options*, that allow robots to perform well even in the very early stages of their lives. This is important when agents need to operate in environments where the use of poor-performing policies (such as the random policies with which they are typically initialized) may be catastrophic. Our method augments the original action set of the agent with specially-constructed behaviors that maximize performance over a possibly infinite family of related motor tasks. These are akin to primitive reflexes in infant mammals—agents born with our early-life options, even if acting randomly, are capable of producing rudimentary behaviors comparable to those acquired by agents that actively optimize a policy for hundreds of thousands of steps. We also introduce three metrics for identifying useful early-life options and show that they result in behaviors that maximize both the option's expected return while minimizing the risk that executing the option will result in extremely poor performance. We evaluate our technique on three simulated robots tasked with learning to walk under different battery consumption constraints and show that even random policies over early-life options are already sufficient to allow for the agent to perform similarly to agents trained for hundreds of thousands of steps.**

***Index Terms*—Machine Learning methods for robot development; Development of skills in biological systems and robots**

I. Introduction

Using Reinforcement Learning (RL) algorithms to solve high-dimensional control problems may require a number of samples that is prohibitively large. Solving Atari games, for instance, often requires an agent to interact with its environment for hundreds of millions of timesteps. This is in sharp contrast with the level of performance achieved by humans and other animals when interacting with new tasks. One of the reasons why RL algorithms cannot yet achieve these performance levels is that they solve each new problem *tabula rasa*, while humans come in with a multitude of prior knowledge—either innate or acquired throughout their lifetimes.

Developmental psychologists have studied the prior knowledge that humans often use when interacting with their environments, both in terms of visual biases [1] and in terms of developmental processes for acquiring reusable motor skills such as reaching or grasping [2]. From a computational perspective, previous works have investigated the different ways in which the performance of RL agents is hurt due to the lack prior knowledge—e.g., lack of innate visual biases and biases towards exploring objects [3], [4]. In the RL community, a common way of equipping agents with motor priors for accelerating learning is through the use of *options*, or temporally-extended actions [5]. Options are reusable behaviors defined in terms of primitive actions or other options. One of the motivating principles underlying this idea is that subproblems recur, so that options can be reused when solving a variety of related tasks throughout an agent's lifetime.

Most of the existing state-of-the-art methods for learning options focus on identifying (*during an agent's lifetime*) useful recurring behaviors that help it to acquire optimal policies more rapidly. Popular approaches include the creation of options for reaching states deemed to be important, such as subgoals [6], [7]. Other techniques directly optimize the parameters of an option's policy so that the agent can more efficiently solve novel tasks [8], [9]. **In this work, by contrast, we do not wish to identify options that help an agent to perform well throughout its entire lifetime. We, instead, want to identify options that allow robots to perform well in the *very early stages* of their lives**. This is important whenever agents may need to operate in environments where the use of poor-performing policies (such as the random policies with which they are typically initialized) may be catastrophic. We call these behaviors *early-life options* and consider them to be similar to primitive reflexes—such as the sucking reflex or the Moro reflex—in infant mammals. The Moro reflex [10] is a particularly relevant example to the setting we tackle: it is present at birth and causes an infant's legs and head to extend, while the arms jerk up, whenever the infant experiences sudden shifts in its head position. In human evolutionary history, this reflex may have helped infants to hold on to their mothers while being carried around. Importantly, this reflex is useful only in the very early stages of an infant's life and disappears after about six months.

We introduce a method for identifying short-duration reusable early-life options that allow robots to perform well in the very early stages of their lives. Our method augments the original action set of the agent with specially-constructed options akin to primitive reflexes in mammals, so that agents equipped with them are capable of achieving high performance on a possibly infinite family of related motor tasks—i.e., they are reusable across many tasks. We propose an offline

This work was partially supported by the Research Council of Norway (RCN); by the Norwegian Centre for International Cooperation in Education (SIU), part of the project Collaboration on Intelligent Machines (COINMAC), grant no. 261645; and by FAPERGS, grant no. 17/2551-000.

optimization process for generating, evaluating, and selecting candidate options that maximize specially-constructed performance metrics. We propose three metrics for evaluating the usefulness of candidate options and show that they identify behaviors that maximize both the expected return of an option while also minimizing the risk that executing it will result in extremely poor performance. We evaluate our technique on three simulated robots tasked with learning to walk under different battery consumption constraints and show that even random policies over early-life options are already sufficient to allow for the agent to perform similarly to agents that are trained for hundreds of thousands of steps.

II. Background

We consider a Reinforcement Learning (RL) agent interacting with its environment and model this problem as a Markov Decision Process (MDP). An MDP M is a tuple $(\mathcal{S}, \mathcal{A}, r, p, \gamma)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a function returning scalar rewards for executing action a in state s, p is a transition function specifying the probability $p(s'|s, a)$ of transitioning to s' after taking action a in state s, and $\gamma \in [0, 1)$ is a discount factor. A policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a map specifying the probability $\Pr(a|s)$ of selecting action a when in state s. The goal of an RL agent is to learn a policy that accumulates as much reward as possible. Let the reward received at time t be the random variable R_t and the cumulative reward (or *return*) from time t be the random variable $G_t \doteq \sum_{i=0}^{T-1} \gamma^i R_{t+i}$, where T is a time horizon. A value function $v_\pi(s)$ is defined as the expected returned achieved when following policy π and starting in state s: $v_\pi(s) \doteq \mathbb{E}[G_t|S_t = s]$. Solving an MDP M consists of finding a policy π^* that maximizes the agent's expected return.

We introduce a method for identifying reusable behaviors that help a robot to perform well in the early stages of its life. As previously discussed, a standard way of representing such behaviors is via the *Options framework* [5]. This framework describes a set of formalisms for learning using temporally-extended actions (or *options*), which are reusable policies defined in terms of primitive actions or other options. One of the motivating principles underlying this idea is that subproblems recur, so that options can be reused in a variety of tasks. An option o consists of three components: *(1)* a policy $\pi_o(s, a)$ describing the probability of taking action a while executing option o in state s; *(2)* an initiation set I_o specifying the states $s \in S$ in which the option can be initiated; and *(3)* a termination condition $\beta_o(s)$ specifying the probability of the option terminating at a state s. To keep our formalism simple we assume that options can be initiated from any state (i.e., $I_o = \mathcal{S}$) and that they last a short pre-defined number T of timesteps so that its termination condition $\beta_o = 1$ iff the option has been executed for T steps. This latter decision is justified by the observation that while learning with longer options may be more efficient, if an option set is not guaranteed to be ideal for a task (e.g., it does not allow for optimal policies to be represented exactly), then shorter options are more flexible and may result in better solutions [11]. Due to these assumptions, we henceforth refer to an option o simply by its policy π_o and leave its other components implicit.

III. Setting

We assume an RL agent that needs to solve not a single problem (task), but that may be presented with a *sequence* of tasks drawn from some task distribution. This is the setting typically tackled by methods dealing with learning options in multi-task problems [12]–[14]. Each task is modeled as an MDP, and we assume that the MDPs have dynamics and reward functions similar enough so that they can be considered variations of the same task. In Section V we expand on this point and introduce an infinite family of related MDPs corresponding to motor problems where robots need to learn to walk efficiently while operating under different power consumption constraints. Let Ψ be the set of possible tasks that an agent may need to solve. Each element of this space is an MDP which we assume can be compactly described by a vector τ of parameters. Learning to grasp a particular object, for instance, is a task that may be compactly characterized by parameters specifying the object's shape and weight. Assume, furthermore, that problems in Ψ occur in an agent's lifetime with probabilities given by some distribution P.

Similarly to how we defined the value $v_\pi(s)$ of a state s, let us define the value $v_{d_0,P}(\pi)$ of a policy π when evaluated over a distribution P of tasks and by considering the possible initial states from which it may be deployed—where initial states are drawn from a distribution d_0:

$$v_{d_0,P}(\pi) \doteq \int P(\tau) \sum_{s \in \mathcal{S}} d_0(s) v_\pi(s, \tau) d\tau, \tag{1}$$

where $v_\pi(s, \tau)$ is defined similarly to $v_\pi(s)$ but makes it explicit that the policy is being evaluated in a particular task τ. We omit the dependence on d_0 and P to simplify the notation and refer to the performance of a policy over a distribution of tasks and initial states simply as $v(\pi)$. Importantly, note that $v(\pi)$ is an *expected value*. In this paper, however, we will be concerned not only with optimizing the expected performance (return) of a policy or option, but with optimizing more sophisticated properties of its *distribution* of its possible returns. Let a context $C \doteq (\tau, s_0)$ be a random variable denoting a tuple containing a task τ drawn from P and an initial state s_0 drawn from d_0. Let $V(\pi)$ be the random variable denoting the possible returns obtained by executing π in a random context; it should be clear, then, that $v(\pi) = \mathbb{E}_{d_0,P}[V(\pi)]$. In the next sections we will evaluate not only single options, but *sets* of options. We abuse notation and extend the definition of the value of an option, $V(\pi_o)$, to the value of a set of options, $V(\pi_{o_1}, \ldots, \pi_{o_K})$. This is a random variable denoting the average of the corresponding options' returns:

$$V(\pi_{o_1}, \ldots, \pi_{o_K}) \doteq \frac{1}{K} \sum_{i=1}^{K} V(\pi_{o_i}). \tag{2}$$

IV. METHOD

Our goal is to identify a set $O^* = \{\pi_{o_1}, \ldots, \pi_{o_K}\}$ of options that can be used to augment $\mathcal{A}$, so that the resulting agent has access to behaviors that are akin to primitive reflexes in infant mammals. In particular, agents born with such options, even when acting randomly in the early stages of their lives, should be capable of producing rudimentary behaviors with performance comparable to that of agents that optimize a policy for hundreds of thousands of steps. We call these *early-life options* since their goal is to guarantee good performance in the early stages of an agent's life where the use of initial near-random policies may be catastrophic. One way of characterizing this set is by identifying a set of options whose expected return, when evaluated over a distribution of possible tasks and initial states, is maximal:

$$O^* \doteq \underset{\pi_{o_1},\ldots,\pi_{o_K}}{\arg\max} \quad \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}[V(\pi_{o_i})]. \tag{3}$$

This definition makes an important assumption: it evaluates a set of candidate options based on their individual performances. Usually, however, options are evaluated *not* by the return that they provide individually, but by the benefits they provide to an agent throughout its entire lifetime (e.g., [6], [9]). This is *not* our objective, however. We, by contrast, wish to identify options that help agents in the *early stages* of their lives, possibly when they are still operating under near-random initial policies. The criterion in Eq. 3 models this worst-case scenario precisely: it represents the average return achieved by an agent acting under a random policy over options. A set of options with high expected return even in this a case—i.e., when the agent is not learning and acts under an initial random policy—is considered to be a good set of early-life options.

A. Quantifying the Performance of Early-Life Options

Eq. 3 defines an optimal set of options by maximizing their expected returns. We are also interested, however, in defining more sophisticated option evaluation metrics that take into account not only the distribution's mean return but also properties of its tails in order to minimize the probability that an option may produce risky performances. We introduce three metrics to evaluate a candidate early-life option π_o based on more general properties of its return distribution:

1) the *Maximum-Mean* metric $\psi_\mu(\pi_o)$. This is the simplest way of evaluating an option, by directly measuring the expected value of its return distribution. It does not take into account return variance or the negative tails of its distribution (i.e., the risks associated with the option);
2) the *Negative Tail-Averse* metric $\psi_-(\pi_o)$. This metric takes into account not only the expected value of an option's return distribution, but also the area under the curve of its negative tail. It favors options with both high expected return and with a low probability of producing significantly poor (possibly catastrophic) returns;
3) the *Positively-Skewed* metric $\psi_+(\pi_o)$. This metric takes into account not only the expected value of an option's return distribution, but also the area under the curve to the right of the distribution's mean. It favors options with a high expected return, and that maximize the probability that they will produce behaviors of above-average quality—even if at the risk of sometimes producing behaviors with subpar performance.

We can now re-write Eq.3 in a more general form so that the value of each option is evaluated w.r.t. a selected metric ψ and not necessarily w.r.t. its expected return. The set of optimal options O^*_ψ w.r.t. a given metric ψ is, then

$$O^*_\psi \doteq \underset{\pi_{o_1},\ldots,\pi_{o_K}}{\arg\max} \quad \frac{1}{K}\sum_{i=1}^{K} \psi(V(\pi_{o_i})). \tag{4}$$

Note that Eq. 4 is equivalent to Eq. 3 if the *Maximum-Mean metric* ψ_μ is used. As previously discussed, this metric directly measures the expected value of an option's return distribution and is thus defined as

$$\psi_\mu(\pi_o) \doteq \mathbb{E}[V(\pi_o)]. \tag{5}$$

Given this definition, it should be easy to see that $O^*_{\psi_\mu}$ is simply a set of K options with highest expected return. Let us denote the mean and standard deviation of this option set as $\mu^* \doteq \mathbb{E}[V(O^*_{\psi_\mu})]$ and $\sigma^* \doteq (\text{Var}[V(O^*_{\psi_\mu})])^{\frac{1}{2}}$, respectively. Note, again, that these statistics refer to the set of early-life options selected solely to maximize their *expected returns*, but without taking into account any risks associated with executing the options. To account for this possibility, we introduce two risk-aware metrics for evaluating the performance of candidate early-life options, so as to quantify both how well they perform over a wide range of possible contexts (tasks and initial states) and the possible *risks* involved with executing them. A common way of incorporating the notion of risk when evaluating policies is via the Markowitz mean-variance model [15]. According to this model, one should prefer policies π that maximize $\mathbb{E}[V(\pi)] - \beta\text{Var}[V(\pi)]$, where $\beta \in \mathbb{R}$ regulates the penalty on return variability. This criterion imposes a trade-off that penalizes expected return in favor of policies with lower variance; it does not care, however, whether the variance is equally caused by above-average and below-average returns, or whether, e.g., most of the variance results from extremely positive (above-average) returns. In this latter case, it should be intuitively clear that return variance is desirable and should not be penalized. To more carefully model the different ways in which return variability may positively or negatively affect the desirability of a candidate option, we introduce two novel metrics for evaluating their performances:

$$\psi_-(\pi_o) = \mathbb{E}[V(\pi_o)] - k\text{Pr}\big(V(\pi_o) < (\mu^* - \alpha\sigma^*)\big), \tag{6}$$

which we call the *Negative Tail-Averse* metric; and

$$\psi_+(\pi_o) = \mathbb{E}[V(\pi_o)] + k\text{Pr}(V(\pi_o) > \mu^*), \tag{7}$$

which we call the *Positively-Skewed* metric. The Negative Tail-Averse metric ψ_- (Eq. 6) takes into account both the mean return of an option and the probability that its execution will result in returns that are α standard deviations below the mean

of $O^*_{\psi_\mu}$. Intuitively, it first characterizes the negative tail of the return distribution of the options set constructed greedily solely based on options' mean returns (via the ψ_μ criterion). Then, it trades-off between achieving high expected return while minimizing the probability that the returns of the option may fall in that tail. This metric favors early-life options that are similar in nature to the ones identified by ψ_μ (in terms of large expected returns) but that also are risk-aware—they would only get selected by Eq. 4 if they are unlikely to produce extremely poor performances. The Positively-Skewed metric ψ_+ (Eq. 7), by contrast, favors early-life options that have both a large expected return and whose returns tend (with high probability) to be situated above the mean of $O^*_{\psi_\mu}$. Intuitively, it cares about the mean return of an option and also about maximizing the probability that its execution will, *most of the time*, produce better-than-expected returns—even if at the risk of sometimes (with low probability) producing behaviors with subpar performances. These are risk-seeking early-life options since they favor behaviors that tend to generate extremely positive returns while accepting the risk of a possibly longer negative tail. This metric results in options with a return distribution that is positively-skewed—thus its name. The preference for actions with positively-skewed returns has been extensively studied in Prospect Theory. Evidence exists, for instance, that when losses are costlier than gains, investors tend to favor stocks with positively-skewed returns [16].

B. Constructing Early-Life Option Sets

We approximate the solution to Eq. 4 using a three-step procedure: **(a)** generation of a set containing N candidate early-life options; **(b)** approximation of each candidate option's return distributions by evaluating its return over Z different random contexts; and **(c)** selecting the top K highest-ranking options according to a given metric ψ. The generation of candidate options is done by sampling N possible random contexts $C_i = (\tau, s_0)$ in which the agent could have to perform in the early stages of its life. We place the agent in such a context and record a short trajectory of optimal actions drawn from an optimal policy for τ. The action trajectory is then used to construct a candidate option's policy π_o capable of (approximately) reproducing the behavior observed in the trajectory. This can be achieved via imitation learning algorithms or standard supervised learning techniques. The process of approximating a candidate option's return distribution is simpler: for each candidate option π_o, we generate a large number Z of possible random contexts and execute the option in each one. We then use the resulting Z return observations of π_o to construct an approximation of the return distribution of the option; this can be achieved, e.g., by using kernel density estimation techniques over the returns. Finally, the process of selecting the top K best options according to a metric ψ requires only that we use the estimated return distribution of each option π_o to compute its corresponding metric $\psi(\pi_o)$. The K highest-ranking candidate options w.r.t. ψ can then be used to define the set O^*_ψ. For more details, see Algorithm 1.

Algorithm 1 Construction of Early-Life Option Sets

(a) Generate Candidate Options
1. Draw a small set W of tasks $\{\tau_1, \ldots, \tau_M\}$ from P
2. Compute a near-optimal policy $\pi^*_{\tau_i}$ for each task in W
for i from $1 \ldots, N$ **do**
- Sample a random context $C_i = (\tau, s_0)$
  · (τ is uniformly drawn from W; s_0 is drawn from d_0)
- Execute π^*_τ for T steps, starting in s_0
- Record the resulting action trajectory $h_i = (a_1, \ldots, a_T)$
- Define an option π_{o_i} that reproduces the behavior in h_i
end for
3. Return the set of candidate options $\{\pi_{o_1}, \ldots, \pi_{o_N}\}$

(b) Estimate Return Distribution of Candidate Options
for each candidate option π_{o_i} **do**
for j from $1 \ldots, Z$ **do**
- Sample a random context $C_j = (\tau, s_0)$
  · (τ drawn uniformly from W; s_0 is drawn from d_0)
- Execute π_{o_i} for T steps, starting in s_0
- Record the return R_j achieved by π_{o_i} in T steps
end for
- Let $R(i) = \{R_1, \ldots, R_Z\}$ be the returns of π_{o_i}
- Use $R(i)$ to approximate the return distribution of π_{o_i}
end for
- Return the approximate return distribution of each option

(c) Select Top K Candidate Early-Life Options
1. Let ψ be the option evaluation metric of interest
2. Compute $\psi(\pi_{o_i})$ for each candidate option π_{o_i}
3. Return the set O^*_ψ with the K highest-ranking options

V. Experiments

In this section, we show that it is possible to identify K reusable early-life options that generalize across many tasks, and that allow for an agent to perform well in the early stages of their lives, even when no learning is taking place. We evaluate our method in three simulated robots selected due to their sensitivity to poor-performing initial policies. Walker is a planar biped robot consisting of two legs and a torso and with 6 actuated joints. Half-Cheetah is a planar biped robot composed of two legs, a torso, and 6 actuated joints. Finally, Ant is a quadruped robot with 13 rigid links, including four legs and a torso, along with 8 actuated joints. The state space of all robots includes information such as its current pose; their reward functions incentivize proper walking while consuming as little energy as possible. We train all agents using the Proximal Policy Optimization (PPO) algorithm [17]. In our experiments, we define the family of tasks Ψ that each robot may face by defining an infinite number of MDP variations; these were obtained by modifying the reward function of each robot by varying the electricity costs of running each actuated joint along the continuous range of $[-2.4, -1.4]$. Higher costs keep the robot from moving efficiently and lower costs often create unstable walking movements since the robot has no incentive to pursue parsimonious movements. We define the task distribution P to be uniform over the range of possible electricity costs and the initial state distribution d_0 to be the one that results from initializing the agent in a standard pose and running a random policy over primitive actions for a

random amount t of steps, where $t \sim U[0, 100]$. In all experiments, we generated $N = 600$ candidate early-life options and estimated their return distributions by evaluating each option in $Z = 300$ possible contexts. Option policies were open-loop sequences lasting $T = 200$ steps and directly mimicking each sampled action trajectory h_i. Return distributions in our analyses were obtained via kernel density estimation over the set of Z returns collected for each option. All experiments evaluate the setting where option sets are of size $K = 5$.

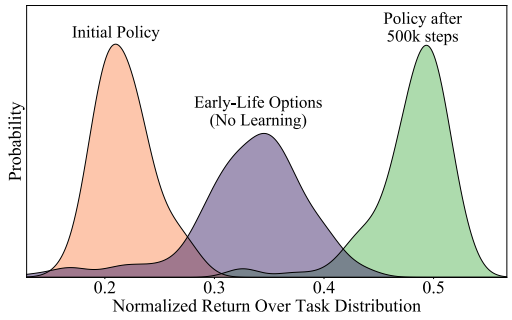

Fig. 1. [Ant Robot] Return distribution achieved with early-life options vs. two learning agents at different moments in their lifetimes.

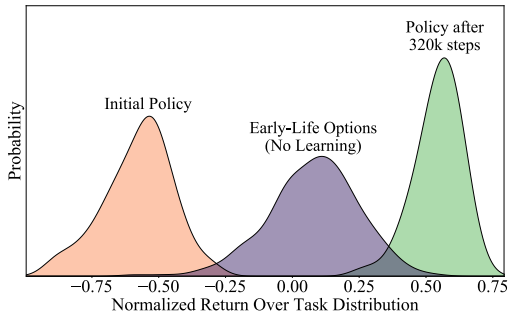

Fig. 2. [Cheetah Robot] Return distribution achieved with early-life options vs. two learning agents at different moments in their lifetimes.

Our first experiment aims at demonstrating that our method is capable of learning behaviors akin to primitive reflexes in infant mammals: our "infant agent", born with early-life options, is capable of directly producing rudimentary behaviors with performances comparable to those acquired by learning agents optimizing a policy for hundreds of thousands of steps. Figures 1 and 2 depict the distribution of returns achieved by different agents when evaluated over $Z = 300$ random contexts, where each context specifies a random task (a random setting for the electricity costs) and a random initial state. Fig. 1 shows the performance achieved by our Ant agent, when equipped with early-life options identified via the Maximum-Mean metric ψ_μ and evaluated during the hardest period of its lifetime: immediately after it is initialized, when it has not yet learned a policy and when its behavior is still essentially a random policy over primitive actions and options. We compare the distribution of returns achieved by our agent with the distributions achieved by two other agents: a learning agent operating under its initial random policy over primitive actions; and a learning agent acting under a policy acquired after 500k training steps. The mean performance of our agent (which is *not learning*, but merely selecting early-options at random) is comparable to that of an agent trained with PPO for around 200k steps. Fig. 2 presents a similar analysis but for the Half-Cheetah robot. We again observe that our optimized early-life options allow the agent—even when selecting uniformly at random from the options set—to perform similarly to an agent trained with PPO for around 190k steps. As expected, early-life options did allow for the agents to perform well (even before any learning has taken place) in the very early stages of their lives. Our second experiment analyzes the properties of different metrics for evaluating early-life options: the Negative Tail-Averse metric ψ_- and the Positively-Skewed metric ψ_+. Fig. 3 compares the return distribution achieved by the best option w.r.t. ψ_μ and w.r.t. ψ_-. It highlights, in particular, the improvement (decrease) in the area under the curve (AUC) of the distribution's left tail that results from using ψ_-. This confirms that ψ_- is capable of identifying options with both high mean return and with a smaller negative tail—thereby producing behaviors that minimize the probability of extremely poor performances. Detailed numerical results regarding improvements to the negative tail's AUC are in Table I. Interestingly, the use of a metric that trades-off mean return and negative tail minimization often resulted in options with a *higher* mean return, compared to that achieved by greedily constructing options based on ψ_μ. Figure 4 presents a similar analysis but regarding the use of the Positively-Skewed metric ψ_+ for selecting options. As previously discussed, ψ_+ cares not only about the mean return of an option, but also about maximizing the probability that its execution will, *most of the time*, produce better-than-expected returns. This figure highlights the improvement (increase) in the area under the curve (AUC) to the right of the distribution's mean. It confirms that the use of this metric is capable of identifying options with

TABLE I
MEAN RETURN & NEGATIVE TAIL'S AUC UNDER ψ_μ AND ψ_-.

	Return under ψ_μ		**Return under ψ_-**		
	Mean	AUC	Mean	AUC	AUC Improvement
Ant	0.328	0.133	**0.332** ($k = 0.05$)	**0.097**	**27.0%**
Cheetah	0.082	0.143	**0.079** ($k = 0.75$)	**0.127**	**11.1%**
Walker	0.122	0.243	**0.115** ($k = 0.75$)	**0.192**	**20.9%**

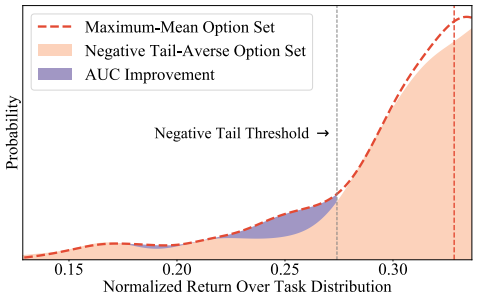

Fig. 3. Negative tail's AUC improvement due to the use of the ψ_- metric.

TABLE II
MEAN RETURN & ABOVE-THE-MEAN AUC UNDER ψ_μ AND ψ_+.

	Return under ψ_μ		**Return under ψ_+**		
	Mean	AUC	Mean	AUC	AUC Improvement
Ant	0.328	0.533	**0.334** ($k = 0.05$)	**0.597**	**12.0%**
Cheetah	0.082	0.550	**0.082** ($k = 0.05$)	**0.550**	0.0%
Walker	0.122	0.563	**0.125** ($k = 0.10$)	**0.595**	**5.7%**

both high mean return and that favors behaviors that tend to generate positive returns while accepting the risk of a possibly longer negative tail. Detailed numerical results of the above-the-mean AUC improvement resulting from ψ_+ are in Table II. Once again, the use of a risk-aware metric often resulted in options with *higher* mean return, compared to that achieved by greedily constructing options based on ψ_μ.

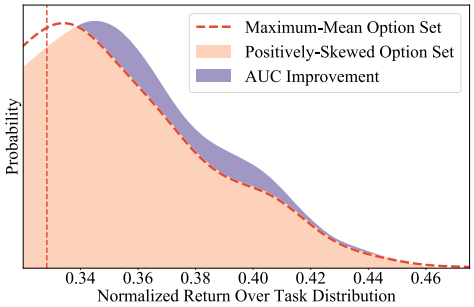

Fig. 4. Above-the-mean AUC improvement due to the use of the ψ_+ metric.

Finally, we study the distribution of values generated by one of our metrics (ψ_+). Fig. 5 shows that very few of the candidate options have large metric values: only 0.67% of all options have performance within 10% of top option's performance. This is surprising since all option policies were generated by sampling directly from optimal policies for a given task (Algorithm 1). This observation implies that it is unlikely that such options, if selected randomly from the set of candidates, would directly generalize over a wide range of contexts, and further reinforces the need for a careful optimization process for identifying efficient early-life options.

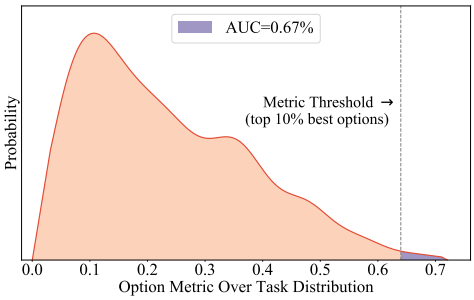

Fig. 5. Distribution of the ψ_+ metric values over candidate options.

VI. CONCLUSION

We have introduced a method for identifying short-duration reusable motor behaviors—*early-life options*—that allow robots to perform well in the very early stages of their lives. Most of the existing work in the area of option generation focuses on identifying options that help an agent throughout its entire lifetime [6]–[9]. We, by contrast, are motivated by the observation that many infant mammals have primitive reflexes that are essential to guarantee their safety only in the early stages of their lives, but that disappear shortly. We introduced a method capable of generating options of this type by optimizing different performance metrics that take into account both an option's mean return and the potential risks that its execution may cause. We evaluated our technique on three simulated robots operating under different battery consumption constraints and shown that random policies over learned early-life options are already sufficient to produce performances similar to those of policies trained for hundreds of thousands of steps. As future work, we intend to study the generation of close-loop early-life options, which may allow for longer movements that better generalize across tasks.

REFERENCES

[1] E. S. Spelke, "Principles of object perception," *Cognitive Science*, vol. 14, no. 1, 1990.

[2] N. Berthier and R. Keen, "Development of reaching in infancy," *Experimental Brain Research*, vol. 169, pp. 507–518, 2006.

[3] F. Doshi-Velez and Z. Ghahramani, "A comparison of human and agent reinforcement learning in partially observable domains," in *Proceedings of the 33th Annual Meeting of the Cognitive Science Society*, 2011.

[4] R. Dubey, P. Agrawal, D. Pathak, T. Griffiths, and A. Efros, "Investigating human priors for playing video games," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1349–1357.

[5] R. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, pp. 181–211, 1999.

[6] M. C. Machado, M. G. Bellemare, and M. H. Bowling, "A laplacian framework for option discovery in reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[7] A. McGovern and A. Barto, "Automatic discovery of subgoals in reinforcement learning using diverse density," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[8] J. Harb, P. Bacon, M. Klissarov, and D. Precup, "When waiting is not an option: Learning options with a deliberation cost," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[9] P. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[10] L. Berk, *Child Development*, ser. Child Development. Allyn & Bacon/Pearson, 2009.

[11] A. Harutyunyan, P. Vrancx, P. Bacon, D. Precup, and A. Nowé, "Learning with options that terminate off-policy," *CoRR*, 2017.

[12] J. Kober, A. Wilhelm, E. Oztop, and J. Peters, "Reinforcement learning to adjust parametrized motor primitives to new situations," *Autonomous Robots*, 2012.

[13] B. da Silva, G. Konidaris, and A. Barto, "Learning parameterized skills," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[14] F. Stulp, G. Raiola, A. Hoarau, S. Ivaldi, and O. Sigaud, "Learning compact parameterized skills with a single regression," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2013.

[15] H. M. Markowitz, "Portfolio selection: Efficient diversification of investment," *The Journal of Finance*, vol. 15, 12 1959.

[16] A. Kumar, M. Motahari, and R. Taffler, "Skewness preference and market anomalies," Social Science Research Network, Tech. Rep., 2018.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, 2017.