Learning to Walk Autonomously via Reset-Free Quality-Diversity

Bryan Lim∗
Adaptive and Intelligent Robotics Lab
Imperial College London, U.K.
bryan.lim16@imperial.ac.uk

Alexander Reichenbach∗
Adaptive and Intelligent Robotics Lab
Imperial College London, U.K.
alexander.reichenbach20@imperial.ac.uk

Antoine Cully
Adaptive and Intelligent Robotics Lab
Imperial College London, U.K.
a.cully@imperial.ac.uk

ABSTRACT
Quality-Diversity (QD) algorithms can discover large and complex behavioural repertoires consisting of both diverse and high-performing skills. However, the generation of behavioural repertoires has mainly been limited to simulation environments instead of real-world learning. This is because existing QD algorithms need large numbers of evaluations as well as episodic resets, which require manual human supervision and intervention. This paper proposes Reset-Free QD (RF-QD) as a step towards autonomous learning for robotics in open-ended environments. We build on Dynamics-Aware QD (DA-QD) and introduce a behaviour selection policy that leverages the diversity of the imagined repertoire and environmental information to intelligently select of behaviours that can act as automatic resets. We demonstrate this through a task of learning to walk within defined training zones with obstacles. Our experiments show that we can learn repertoires of locomotion controllers autonomously without manual resets and with high sample efficiency in spite of harsh safety constraints. Finally, using an ablation of different target objectives, we show that it is important for RF-QD to have diverse types solutions available for the behaviour selection policy over solutions optimised with a specific objective. Videos and code available at https://sites.google.com/view/rf-qd.

CCS CONCEPTS
• Computational Methodologies → Evolutionary Robotics.

KEYWORDS
Quality-Diversity, Reset-free, Safety-aware Learning, Robotics

ACM Reference Format:
Bryan Lim, Alexander Reichenbach, and Antoine Cully. 2022. Learning to Walk Autonomously via Reset-Free Quality-Diversity. In Genetic and Evolutionary Computation Conference (GECCO ’22), July 9–13, 2022, Boston, MA, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512290.3528715

1 INTRODUCTION
Despite the recent popularity of Quality-Diversity (QD) algorithms [8, 33], these algorithms have been limited to domains in which evaluations can be performed in simulation. This is because QD algorithms need to perform evaluations in the order of millions and where the outcomes are not safety critical or dangerous. Examples of these application domains include robotics [2, 7, 9], video games [11, 14] and aerodynamics [13]. In the field of robotics, physics simulators [4, 20, 35] are commonly used and QD algorithms depend heavily on these to obtain abundant amounts of data and evaluations to learn behavioural repertoires of robots. However, building fast and accurate physics simulators to model the complex dynamics of robots and the wide variety of potential environments is difficult. Furthermore, even with extensive modelling of different scenarios, there is still the difficult problem of sim-to-real transfer [1, 21, 38]. To realize the potential of learning and QD algorithms for robotics and to have the real-world impact we want them to have, we need algorithms which can effectively learn repertoires of skills autonomously and adapt directly in the real-world.

In this work, as a step towards using QD algorithms in the real-world, we address two key issues that arise when attempting to learn behavioural repertoires in the real-world in an autonomous manner: resets and safety. We then aim to maximise the sample efficiency of learning behavioural repertoires while considering these constraints. We specifically focus on reset-free learning of robotic locomotion skills to highlight these issues and our approach to solving them.

An often overlooked requirement of QD algorithms when used for Reinforcement Learning (RL) is the episodic setting they function in. This requires the environment to be set to a fixed initial state at the start of every episodic trial as the behavioural descriptor is measured as a function of the trajectory of states from this initial state to the final state in the episode. In real-world settings, this would correspond to humans manually resetting the robot and environment after every episode. This is an impractical and
expensive solution considering the number of evaluations that con-
ventional QD algorithms require are in the order of millions. With
the considerable amount of human supervision and instrumenta-
tion required for resets, this defeats the purpose of QD algorithms
to autonomously learn complex skills.

Another key challenge of learning in the real-world is safety. Ac-
tions taken must not be dangerous to the robot and the environ-
ment. For learning locomotion skills, this corresponds to avoiding
collisions with objects in the environment during learning. Achiev-
ing this while performing QD would require both the capability to
predict the outcome of the execution of a new behaviour as well as
an understanding of its implications with respect to its safety.

Finally, we want to learn skills efficiently with efficiency be-
ing measured by the number of evaluations taken. In the real-
world, this also directly corresponds to the learning time needed.
The goal is to intelligently select behaviours that would help im-
prove the behavioural repertoire while minimizing unnecessary
non-informative trials.

We introduce Reset-Free Quality Diversity (RF-QD) as a frame-
work for the real-world execution of QD algorithms (see Figure 1).
In a nutshell, RF-QD is a Dynamics-Aware Quality-Diversity (DA-
QD) algorithm combined with a Behaviour Selection Policy to select
only safe and valuable behaviours for evaluation in the (potentially
dangerous) real-world. We demonstrate an algorithm which au-
tonomously acquires a diverse repertoire of locomotion skills on a
hexapod robot in safety-constrained environments.

2 RELATED WORK

2.1 Quality-Diversity and Behavioural
Repetoire Learning in Robotics

Quality-Diversity (QD) optimization is a class of algorithms that
aims to generate a collection of both diverse and high-performing
solutions [8, 33]. In the context of robotics, each solution can for
instance, be a parametric policy which determines sequences of
actions to execute (i.e. motor commands), resulting in a behaviour.
A behaviour can then be represented by a numerical vector referred
to as the Behavioural Descriptor (BD). The BD is a low-dimensional
representation of the trajectory of states the policy visited and is
usually defined manually depending on the tasks. However, the
BD can also be learned in an unsupervised manner [5, 15, 30]. The
choice of the BD is important as it determines the novelty of a
solution and will be used to drive the exploration to cover the BD
space [23].

Conceptually, QD extends the single novelty-seeking objective
introduced in the Novelty Search algorithm [23] with another mea-
sure of quality. MAP-Elites [26] and Novelty Search with Local
Competition [24] are two commonly used and well-known QD al-
gorithms. Cully and Demiris [8] suggested that most QD algorithms
can be represented using a common framework consisting of two
key components; the archive and selector. Variants of QD algorithms
develop around these components, all building on the QD loop of se-
lection, variation, evaluation and (tentative) addition to the archive.
The archive is used to store the highest performing solutions for
each niche. Instead of a uniform grid used to discretize the BD space,
methods like CVT-MAP Elites [36] or Sliding-Boundary MAP-Elites
[10] modify this to make the archive more flexible. More recently,
there has also been a body of work focused on using more complex
selectors and efficient optimizers such as evolutionary strategies
[3, 12] and policy gradients [29, 32].

In the field of robotics, it was shown that the final archive of
solutions discovered by QD algorithms can be used as a behavioural
repertoire to perform downstream tasks [9]. For example, each so-
lution is a controller which is represented by a set of parametric
functions which controls the robot’s joints. Coupled with Bayesian
optimization, the controllers generated in the archive can be used
to quickly adapt to unforeseen mechanical damage, help with sim-
to-real transfer and solve longer-horizon tasks using planning al-
gorithms [2, 7].

2.2 Model-based Quality-Diversity

One of the main bottlenecks of Quality-Diversity (QD) algorithms
is the sample efficiency. QD algorithms typically require on the
order of millions of evaluations and rely on parallel computation of
these evaluations. This single factor alone usually make them
unsuitable to be used directly in the real-world.

A line of work that attempts to address this problem, now re-
ferred to as model-based quality-diversity methods, is through the
use of models. Surrogate-Assisted Illumination (SAIL) [13] first
introduced the use of surrogate models for QD algorithms. SAIL
integrates surrogate models, in the form of Gaussian Process (GP)
models, to approximate the objective function and reduce the num-
ber of evaluations for the computationally expensive application of
aerodynamic design. Another algorithm called M-QD [18] later
follows up on this idea and used neural network models that map
the parameter space to the behaviour and fitness space as a surro-
gate model. They demonstrate this on robotic pushing and placing
tasks.

Dynamics-Aware Quality-Diversity (DA-QD) [25] is another ap-
proach which instead uses learnt dynamics models as a surrogate
model. DA-QD introduced the concept of the imagined repertoire
which allows QD to be performed fully in imagination using the
learnt dynamics models. The dynamics models are trained incre-
mentally and online as data is collected through evaluations. DA-QD
showed a significant (~20 time) increase in sample-efficiency. Our
work uses the DA-QD framework to heavily reduce the number of
evaluations needed making it feasible to be considered for a
real-world application. With RF-QD, we extend DA-QD to make
better use of the imagined repertoire to select behaviours more
intelligently and in a sequential manner. While DA-QD only uses
variation operators for optimization in imagination, we also further
study the effect of optimizing for different objectives in imagination.

2.3 Reset-free Learning

Reset-free Learning has mainly been studied in gradient-based
Deep Reinforcement Learning (DRL) where the episodic setting is
also usually a prerequisite of the Markov Decision Process (MDP)
formulation of the problem. One approach taken to enable real-
world RL is to automate resets using other manually scripted robots
to reset objects and the environment to the initial state distribution
required [28]. While this works well for resetting manipulation
tasks in which the workspace is relatively limited, this approach is
difficult to apply for learning locomotion behaviours.
We build on the DA-QD framework proposed by Lim et al. [25]. We work explicitly learning policies for tasks in a predefined distribution which extends the conventional QD framework [8, 33] discussed briefly summarizing DA-QD here and refer the reader to the full paper for further details. DA-QD is a model-based QD algorithm to learn dexterous manipulation policies. Both these probabilistic models are used. This forward model can be represented as

\[ p(\tilde{s}_{t+1} | \tilde{s}_t, \tilde{a}_t) \]

The learnt dynamics model is a forward model with 

\[ \tilde{p}(s_{t+1} | s_t, a_t) \]

Both these models are used. This forward model can be represented as

\[ p(\tilde{s}_{t+1} | \tilde{s}_t, \tilde{a}_t) \]

The dynamics model is a forward dynamics model and is represented by a neural network parameterized by \( \theta \). To capture both aleatoric and epistemic uncertainties, an ensemble of probabilistic models are used. This forward model can be represented as

\[ \tilde{p}(s_{t+1} | s_t, a_t) \]

The dynamics model is a forward dynamics model and is represented by a neural network parameterized by \( \theta \). To capture both aleatoric and epistemic uncertainties, an ensemble of probabilistic models are used. This forward model can be represented as

\[ \tilde{p}(s_{t+1} | s_t, a_t) \]

Figure 2: RF-QD performs QD in imagination (as in DA-QD) and uses a more intelligent behaviour selection policy to keep the robot in the safe regions of its environment while maximizing the value gained by every real-world evaluation.

Most similar to our work and another promising method is to use a multi-task RL approach [16, 17]. The key idea behind this approach is to use a scheduler and different tasks present in the multi-task setup as resets for each other. Ha et al. [17] showed this in the context of learning simple locomotion policies while Gupta et al. [16] demonstrated this approach on more extensive multi-task setting to learn dexterous manipulation policies. Both these works explicitly learn policies for tasks in a predefined distribution of tasks. Each policy is optimized individually using an off-the-shelf deep RL algorithm and separate instances of networks and replay buffers. Our work instead concurrently learns a repertoire of diverse policies using QD algorithms and leverage the diversity of the behavioural repertoire as resets. In Multi-task MAP-Elites [27], it is also showed that the behaviour space can also be viewed and formulated as a task-space where each cell is a task.

In the context of QD algorithms and behavioural repertoire learning, the Reset-Free Trial and Error (RTE) [2] algorithm has also aimed to address the reset problem. RTE demonstrated this for adaptation using the behavioural repertoire as a prior for Gaussian Process models. This is a different setting from the work we present in this paper as the behavioural repertoire generation process itself in RTE is performed fully in simulation using resets. The reset-free in RTE refers to the reset-free adaptation when performing simulation to-real transfer or reset-free adaptation to mechanical damage. In our work, we aim to learn the behavioural repertoire itself in a reset-free manner.

Here, the disagreement between predictions of all models in the ensemble captures the epistemic uncertainty, i.e. it indicates the uncertainty of the prediction due to a lack of samples. The overall model disagreement \( \mu_d \) can be calculated as the expected difference between any two models in the ensemble \( f_\theta \) for one state-action pair, averaged over all time step predictions in one rolled-out trajectory (i.e. one evaluated behaviour) of length \( T \) [19]:

\[
\mu_d = \frac{1}{T} \sum_{t=0}^{T} \text{disag}(s_t, a_t)
\]

State transition data is collected and stored in a replay buffer \( \mathcal{B} \) as evaluations of robot behaviour are performed in the environment. The model is trained in a self-supervised manner to maximise log-likelihood of the transitions sampled from replay buffer and is optimized via back-propagation.

The dynamics model \( \tilde{p}(s_{t+1} | s_t, a_t) \) can be called recursively to evaluate policies in what is referred to as an imagined roll-out. The expected fitness and BD can be obtained from this imagined roll-out as both these quantities measured are a function of the state trajectory. DA-QD introduced the concept of an imagined repertoire \( \tilde{A} \) to organise and maintain solutions that have been evaluated in imagination using the dynamics model. The imagined repertoire \( \tilde{A} \) uses the same addition conditions as the repertoire \( A \). The imagined repertoire \( \tilde{A} \) only allows solutions that have been evaluated in imagination that are expected to be novel or better performing than existing solutions to be considered for evaluation. This is where the sample-efficiency of this method is derived from. Additionally, this allows QD to be performed fully in imagination. This means that the selection and mutation of the QD algorithm can be continuously performed from the imagined archive for any desired number of imagined generations without any samples or evaluations on the real system.

Finally, with the introduction of the imagined repertoire \( \tilde{A} \), this necessitates selection of solutions from \( \tilde{A} \) to be evaluated. As the original DA-QD algorithm does not consider the reset-free sequential evaluation setting, the authors select all the solutions that have been added to the imagined archive to be evaluated in parallel. Our work extends DA-QD and proposes a more intelligent method to select and manage solutions in the imagined archive given the

3 BACKGROUND: DYNAMICS-AWARE QD
We briefly summarize DA-QD here and refer the reader to the full paper [25] for further details. DA-QD is a model-based QD algorithm which extends the conventional QD framework [8, 33] discussed in section 2.1 with three key components: a dynamics model, an imagined repertoire and selector from the imagined repertoire.

The learnt dynamics model is a forward dynamics model and is represented by a neural network parameterized by \( \theta \). To capture both aleatoric and epistemic uncertainties, an ensemble of probabilistic models are used. This forward model can be represented as

\[ \tilde{p}(s_{t+1} | s_t, a_t) \]
reset-free setting and safety constraints from the environment. Additionally, DA-QD also does not explicitly use the resulting model-disagreement. In this paper, the model-disagreement is used both as a heuristic to select behaviours to execute more intelligently (Sec. 4.1) and as an optimization objective in imagination (Sec. 5.4).

4 METHODS

We present Reset-Free Quality-Diversity (RF-QD) as a method to enable the application of QD’s behavioral repertoire learning in non-episodic real-world environments (see Algorithm 1). We treat the robot as an actor in its environment that performs a constant search for new and improved behaviours and storing these in the archive. For this, we extend the classical QD loop by two steps. Firstly, we build on the pre-evaluation of any new behaviour ‘in imagination’ by a dynamics model (DA-QD). Secondly, we introduce a behaviour selection policy, that modulates the robot’s search for novel and high-performing behaviours as to comply with the safety constraints given by the environment (see Figure 2).

In the following, we first elaborate on the core of our method: the behaviour selection policy. Then, we detail its main components: the safety evaluation, safety constraints, the prioritisation metrics and the recovery policy.

Algorithm 1 Reset-free Quality-Diversity (RF-QD)

Input: archive \( \mathcal{A} = \emptyset \), candidate buffer \( \mathcal{C} = \emptyset \), dynamics model \( \hat{p}_\theta \), experience replay buffer \( \mathcal{B} \)

\[ \mathcal{A}, \mathcal{B} \leftarrow \text{random_archiveInitialization}() \]
\[ \hat{p}_\theta \leftarrow \text{train_model}(\mathcal{B}) \]

for max_iterations do
if \( \mathcal{C} = \emptyset \) then
\( \mathcal{A} \leftarrow \text{generate_candidates}(\mathcal{A}, \hat{p}_\theta) \rightarrow \text{Using imagination} \)
\( \mathcal{C} \leftarrow \mathcal{A}, \mathcal{A} \)
\( s \leftarrow \text{get_robot_state()} \rightarrow \text{Get state to evaluate safety} \)
if \( s \) is safe then
\( \mathcal{C}_\text{safe} \leftarrow \text{apply_safety_constraint}(\mathcal{C}) \)
\( x \leftarrow \text{argmax( prioritise_candidates(\mathcal{C}_\text{safe}) )} \)
\( b_x, f_x \leftarrow \text{execute_behaviour}(x) \)
\( \mathcal{A} \leftarrow \text{add_to_archive}(x, b_x, f_x) \)
else
\( \mathcal{A} \leftarrow \text{add_to_archive}(x, \text{recovery_policy}(\mathcal{A})) \)
\( \mathcal{B} \leftarrow \text{add_to_replay_buffer}(b_x) \)
\( \hat{p}_\theta \leftarrow \text{train_model}(\mathcal{B}) \)

4.1 Reset-free Behaviour Selection

To be able to stay safe while acting in its environment, we introduce a behaviour selection policy to modulate the robot’s actions in the real world. This behaviour selection ensures that every new behaviour will only be performed if it is expected to be safe for the robot. At every step, new behaviours are selected from a candidate buffer \( \mathcal{C} \). The candidate buffer \( \mathcal{C} \) is regularly filled with new policies from the imagined repertoire \( \mathcal{A} \) that are not already present in the repertoire \( \mathcal{A} \). \( \mathcal{A} \) is a component introduced in DA-QD that maintains solutions that were evaluated in imagination using the dynamics model \( \hat{p}_\theta \). Based on the robot’s current state in the environment, our policy then selects a subset of candidate behaviours \( \mathcal{C}_\text{safe} \) that have a low risk of violating the safety constraints given by the environment. Out of these, the candidate behaviour that has the highest projected prioritization score will then be evaluated in the real world. In the following sections, the core components will be described in more detail.

4.2 Safety Evaluation

In this paper, we assume knowledge of the environment layout, represented by ‘safety regions’ (see Figure 1), that indicate the region of dangerous states \( \Omega \). In practice, this information could as well be obtained using Simultaneous Localisation and Mapping (SLAM) methods with an on-board camera.

Derived from the robot’s state \( s \), we define the exploration parameter \( \epsilon(s) \), which indicates the relative degree of safety in the current state. It is calculated as the smallest distance between \( s \) and \( \Omega \) and normalised by the maximum encountered distance value (see Equation 2). While inside the safe region i.e. \( s \notin \Omega \rightarrow \epsilon(s) > 0 \), the robot must choose any potential solution to be evaluated in the real world that is predicted to keep \( \epsilon(s) > 0 \), i.e. does not enter any unsafe state. To lower the risk of damage to the robot, an offset \( \beta \) can be added in the computation of \( \epsilon(s) \) as an increased threshold for the minimum distance towards the border of the region of unsafe states within the state space.

\begin{equation}
\epsilon(s) = \frac{\text{dist}(s, \Omega) - \beta}{\max \text{dist}(s, \Omega) - \beta}
\end{equation}

From the dynamics model, we can obtain the predicted next state \( s' \) after the execution of a candidate behaviour and compute \( \epsilon(s') \). \( s' \) corresponds to the state \( s_{T+1} \) after \( T \) timesteps, where \( T \) is the length of one behaviour. Generally, we seek the robot to stay as close as possible to the safest point(s) in the environment, i.e. maintain maximal distance to the region of dangerous states \( \epsilon(s) \approx 1 \).

4.3 Safety Constraints

For every behaviour selection performed by our policy, we first employ a safety constraint to determine the safe subset \( \mathcal{C}_\text{safe} \) of all available candidate behaviours with respect to the robot’s current state. We can use different constraints depending on our knowledge of the environment and the intended risk aversion of our exploration. In the experiment section below, we evaluate the following constraints, all of which are based on the predicted robot state \( s' \) after the execution of each imagined behaviour (given the current robot state \( s \)):

- As a minimal constraint we consider only candidate behaviours with \( \epsilon(s') > 0 \) to ensure we never execute a behaviour that was already expected to be unsafe.
- Alternatively, a contextual constraint carries weight only if the current robot state is near the border of the region of unsafe states \( \epsilon(s) \approx 0 \), but enables free exploration if it is far away...
from potential danger ($\epsilon(s) \approx 1$):

$$\epsilon(s') > \epsilon(s) \cdot (1 - \epsilon(s))$$ \hspace{1cm} (3)

- If we have access to the gradient of the epsilon function, the direction of maximal improvement of safety with respect to the next state can be computed as $\nabla_s \epsilon(s)$. The gradient-minimal constraint considers only solutions moving in the general direction of the gradient. Based on the dot product of the unit vectors of the gradient of the epsilon function ($\nabla_s \epsilon(s)$) and the projected movement in state space ($s' - s$), we formulate a lower bound for deviation from the direction of the gradient as:

$$s' - s \geq \frac{\nabla_s \epsilon(s)}{||\nabla_s \epsilon(s)||} \cdot (1 - \epsilon(s))$$ \hspace{1cm} (4)

Geometrically, this is equal to a maximum deviation of 90° in 2D space as visualised in Figure 3 (green semicircle).

- Again, we can modify this into a more strict gradient-contextual constraint by using the value of epsilon at the current state of the robot to modulate the constraint. This way, the constraint is more relaxed towards the centre of the region of safe states but only accepts small deviations from the direction of the safety gradient close to the border of the region of unsafe states:

$$s' - s \geq \frac{\nabla_s \epsilon(s)}{||\nabla_s \epsilon(s)||} \cdot (1 - \epsilon(s))$$ \hspace{1cm} (5)

Geometrically, this is equal to a deviation from the gradient proportional to $\epsilon(s)$ (see yellow region in Figure 3 for $\epsilon(s) = 0.5$).

- Finally, safety can also be enforced not by a hard constraint, but as a component of the prioritization measures. This can especially be useful as a supplement to the gradient-free constraints in complex environments.

4.4 Prioritization Metrics

After the safe subset of candidate behaviours $C_{safe}$ has been selected based on the safety constraint, the remaining candidates are ranked according to a prioritization measure as the second step in behaviour selection. This is intended to give priority to the real-world evaluation of candidate behaviours which have the highest value for the overall QD algorithm performance, as real-world samples are expensive to collect. Finally, the candidate with the highest prioritization score is selected. The composition of prioritization measures can be adapted depending on the task at hand. We can either use a single prioritization measure or a (weighted) sum of multiple values. In this work, we have evaluated the following measures:

- Firstly, the robot’s safety can be considered again as a prioritization measure through the dynamic exploration parameter $\epsilon(s')$ as outlined above. Generally, this approach will be used in combination with another metric to enable the behaviour policy to tolerate a possible safety violation in favor of a higher score.

- Another key measure to score a candidate behaviour is the dynamics model disagreement. The dynamics model used in DAQD consists of an ensemble of models to capture the epistemic uncertainty via disagreement between the predictions of the models (see Section 3 and Equation 1). The epistemic uncertainty can also be interpreted and formalised as an information theoretic measure of the expected information gain [31, 34]. Maximising the model-disagreement has been used as a self-supervised intrinsic reward for exploration in Deep RL literature [31, 34].

4.5 Recovery Policy

As a final safeguard to keep the robot in the safe region of the environment, we introduce a recovery policy to return the robot to safety if it ever violates any of the environment’s safety constraints. These constraints can be derived from the environment in various ways, e.g. as a minimum distance to obstacles represented by ‘safety regions’ as in this work. Should the robot leave the safe region, the discovery of new behaviours will be halted and a greedy behaviour selection policy will be employed over the archive of behaviours that were already evaluated in the environment instead of the buffer of candidate behaviours. Here, we pick the single behaviour that is projected to effect the greatest improvement in safety.

5 EXPERIMENTS

We evaluate our method with an 18 DoF hexapod robot on an adapted version of the omni-directional locomotion task [9]. In this task, the robot learns behaviours to walk in every direction from an initial position. For the controllers, we evolve parameters of a sinusoidal control signal that is sent to each motor. This sinusoidal
signal acts as a structural prior towards periodic movement for locomotion. As we focus on a reset-free setting, all evaluations of new behaviours have to be done sequentially and cannot be parallelised. All simulations are performed in RobotDART building on the Dynamics Animation and Robotics Tookit (DART) simulator [22]. To simulate a practical number of trials that would be performed in the real-world experiment, the number of evaluations performed in any single run of the algorithm are limited to 10,000.

5.1 Baseline comparison

Firstly, we evaluate the general capability of the RF-QD method. For this, we compare against “vanilla” QD and DA-QD [25] as baselines. RF-QD and both baselines use the Iso-dd [37] variation operator. We use a simple flat environment with a circular region of safety with radius $r = 2.0m$. Figure 4 shows example trajectories of the baselines compared to RF-QD. The baselines’ random selection of behaviours causes the robot to trail off deeply into the dangerous region, while RF-QD performs its exploration almost entirely within the safe region. The depicted RF-QD run leaves the safe region once, but then deploys the recovery policy (blue line) to return to safety.

As the baseline methods are not made for a reset-free environment, for all further comparisons we perform manual resets to the starting position if the robot leaves the safe region by more than 50 cm. This is similar to what is done when performing QD on a real-world robot today. For the baseline comparisons, RF-QD was run with a gradient-contextual safety constraint and encouraging maximal novelty through the prioritization strategy. This configuration has proven powerful in our evaluation of different constraints and prioritization measures. Table 1 quantifies the safety of the three algorithms averaged over 10 replications of each. We can see, that RF-QD achieves almost perfect safety - never once requiring a safety reset as described above and only rarely taking a single step outside the safe region.

Additionally, RF-QD slightly outperforms its direct baseline DA-QD in terms of both QD-score and coverage as shown in Figure 5. While the distance to vanilla QD is due to DA-QD’s increased sample efficiency, RF-QD’s behaviour selection policy does not sacrifice performance for safety, but even improves performance by its candidate prioritization strategy (i.e. novelty in this case).

5.2 Comparison of Policy Configurations

Additionally, we evaluated the various configurations of the Behaviour Selection Policy as introduced in Section 4.1. Figure 6 shows an overview over the different combinations of safety constraints and prioritization measures. Here, the policy configurations are evaluated by performance (represented by their final coverage) and safety (represented by the number of recovery steps), both from runs of 10,000 steps over 10 replications. In short, Figure 6 shows strong separation between the relatively unsafe minimal and contextual constraints (both gradient-free) and all remaining constraints. The strongest performance is exhibited by variants combining the novelty or disagreement maximising prioritization measures with a gradient contextual constraint. Out of the naive gradient-free constraints, which must be used if there is no single ’safest’ direction of movement (as e.g. in more complex environments such as the one following in Section 5.3), only the soft constraints achieves comparable safety scores and performances as the gradient-based configurations. Which exact configuration should be chosen will however always depend on the exact task at hand.

5.3 Robustness to environment complexity

To evaluate RF-QD’s performance in increasingly complex environments, we exchange the previous circular environment for a closed 4x4m room with a number of column-shaped obstacles. Figure 7 shows examples of such environments including RF-QD’s trajectories in them (top row). We can observe that the robot acting under
RF-QD keeps its distance from the obstacles, while building archives of behaviours (middle row) that are radically less affected by the environment complexity than those created by DA-QD (bottom row).

In these complex environments, we employed RF-QD with a safety-focused configuration. This uses a minimal (hard) safety constraint combined with two equally weighed prioritization measures to select behaviours that maximise safety (through $\epsilon$) and have low model disagreement. As a benchmark for QD performance, we again add a version of DA-QD that uses safety resets, now triggered on any collision with an obstacle. We also keep a ‘naive’ version of DA-QD, that is not reset upon collision (same as RF-QD). These algorithms were compared in rooms with 0 to 15 obstacles (see Figure 8). While in an empty room, all algorithms perform similarly well, the naive DA-QD variant quickly drops in performance with a growing number of obstacles through a large number of collisions (which render the corresponding evaluations invalid). At the same time, RF-QD manages to fully keep up with the upper baseline of DA-QD (using safety resets). While a more performance-focused prioritization strategy (i.e. novelty as in Section 5.1) for RF-QD might have increased QD-scores slightly, this would have sacrificed the safety of the robot in more challenging environments.

### 5.4 Effect of objectives in imagination

We also study the effect of the type of solutions available in the candidate buffer that the behaviour selection policy chooses from. To study this, we investigate the influence of different optimisation objectives for the generation of the candidate buffer during
We perform an ablation of these three different objectives with solutions with the desired metric being selected. This results in a higher probability for the target objective. This observation corresponds to our initial hypothesis where targeted optimization results in low disagreement overall trends are always observed: The minimising disagreement objective (green) always results in low disagreement. The same observation applied to individuals being selected by the behaviour selection policy regardless of the prioritization metrics used by the behaviour selection policy. This is an interesting result as one could expect the variants with aligned prioritization measures and optimization objectives to perform better. We hypothesize that the buffer of candidate solutions being generated by targeted objectives become too specialised while the objective-agnostic Iso-DD can generate a diverse buffer of solutions to choose from. This is not such a surprising observation as Multi-Emitter MAP-Elites [6] had previously also shown that when using simultaneously multiple emitter types, the random emitter (based on Iso-dd) remains the most fruitful through the entire process compared to other objective-driven emitters.

6 DISCUSSION
In this paper, we have presented RF-QD, a method to learn behavioural repertoires autonomously without resets in realistic environments. We demonstrate how an intelligent behaviour selection policy can be used with QD in imagination to learn safely and efficiently. We first test RF-QD to learn while remaining within a designated area and show that the behaviour selection policy is necessary to prevent the need for resets and to stay within the safe training area. We then show how RF-QD can also operate in more complex environments with many obstacles and minimal room for error. Our results also show that we can acquire full repertoires despite increasing environment complexity while the performance of DA-QD and Vanilla QD baselines deteriorate with the increase in complexity. Lastly, we conduct an ablation to investigate the effect of the composition of this candidate buffer on the performance of RF-QD. We demonstrate that using targeted optimization objectives when performing QD in imagination can bias the distribution of solutions presented to the behaviour selection policy. Our results show that it is important to keep the diverse types of solutions in the candidate buffer over just specialised solutions biased towards a single metric.

For future work, we also hope to show RF-QD learning directly on a real world system, with no dependence on simulators. Additionally, this paper only considers safety and danger in the form of obstacles. We leave other forms dangerous scenarios for future work.

ACKNOWLEDGMENTS
This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/V006673/1 project REcoVER. We would like to thank the members of the Adaptive and Intelligent Robotics Lab for their very valuable comments.

REFERENCES
[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving rubik’s cube with a robot hand. arXiv preprint arXiv:1910.07113 (2019).
Learning to Walk Autonomously via Reset-Free Quality-Diversity

GECCO ’22, July 9–13, 2022, Boston, MA, USA

[2] Konstantinos Chatzilygeroudis, Vassilis Vassiliadis, and Jean-Baptiste Mouret. 2018. Reset-free trial-and-error learning for robot damage recovery. Robotics and Autonomous Systems 100 (2018), 236–250.

[3] Cédric Colas, Vashisht Madhavan, Joost Huizinga, and Jeff Clune. 2020. Scaling map-elites to deep neuromorphology. In Proceedings of the 2020 Genetic and Evolutionary Computation Conference. 67–75.

[4] Erwin Coumans and Yufei Bai. 2016–2020. PyBullet, a Python module for physics simulation for games, robotics and machine learning. http://pybullet.org.

[5] Antoine Cully. 2019. Autonomous skill discovery with quality-diversity and unsupervised descriptors. In Proceedings of the Genetic and Evolutionary Computation Conference. 81–89.

[6] Antoine Cully. 2020. Multi-Emitter MAP-Elites: Improving quality, diversity and convergence speed with heterogeneous sets of emitters. arXiv preprint arXiv:2007.03552 (2020).

[7] Antoine Cully and Jean-Baptiste Mouret. 2013. Behavioral repertoire learning in robotics. In Proceedings of the 15th annual conference on Genetic and evolutionary computation. 175–182.

[8] Matthew C Fontaine, Scott Lee, Lisa B Soros, Fernando de Mesentier Silva, Julian Togelius, and Amy K Hoover. 2019. Mapping heartbeat stone decks through map-elites with sliding boundaries. In Proceedings of The Genetic and Evolutionary Computation Conference. 161–169.

[9] Matthew C Fontaine, Julian Togelius, Stefanos Nikolaidis, and Amy K Hoover. 2020. Covariance matrix adaptation for the rapid generation of behavior space. In Proceedings of the 2020 genetic and evolutionary computation conference. 94–102.

[10] Adam Gaier, Alexander Asterosev, and Jean-Baptiste Mouret. 2018. Data-efficient design exploration through surrogate-assisted illumination. Evolutionary computation 26, 3 (2018), 381–410.

[11] Daniele Gravina, Ahmed Khalifa, Antonios Liapis, Julian Togelius, and Georgios N Konstantinos Chatzilygeroudis, Vassilis Vassiliades, and Jean-Baptiste Mouret. 2018. Discovering the elite hyper- volume by leveraging interspecies correlation. In Proceedings of the Genetic and Evolutionary Computation Conference. 149–156.

[12] Deepak Pathak, Dhuraj Gandhi, and Abhinav Gupta. 2019. Self-supervised exploration via disagreement. In International conference on machine learning. PMLR, 5062–5071.

[13] Thomas Pierrot, Valentin Macé, Geoffrey Cideron, Karim Beguir, Antoine Cully, Olivier Sigaud, and Nicolas Perrin. 2021. Diversity Policy Gradient for Sample Efficient Quality-Diversity Optimization. (2021).

[14] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. 2016. Quality diversity: A new frontier for evolutionary computation. Frontiers in Robotics and AI 3 (2016), 40.

[15] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. 2020. Planning to explore via self-supervised world models. In International Conference on Machine Learning. PMLR, 8583–8592.

[16] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 5026–5033.

[17] Vassilis Vassiliadis, Konstantinos Chatzilygeroudis, and Jean-Baptiste Mouret. 2017. Using centroidal voronoi tessellations to scale up the multidimensional archive of phenotypic elites algorithm. IEEE Transactions on Evolutionary Computation 22, 4 (2017), 623–630.

[18] Wenshuai Zhao, Jorge Peña Queralt, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 737–744.