UNDISCOUNTED POTENTIAL-BASED REWARD SHAPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning usually trains with a discount factor $\gamma < 1$ for stability, even when the task's objective is the undiscounted return. In this paper, we compare two forms of a frequently used reward shaping method, which adds a shaping term to immediate rewards to speed up learning. Potential-based reward shaping (PBRS) provides a framework which guarantees that the shaping term does not change the discounted optimization objective. However, when training uses $\gamma < 1$, the undiscounted objective does not necessarily coincide with the discounted objective. This means that PBRS cannot shape rewards towards an undiscounted optimum when the discounted objective differs from the undiscounted one. To provide stable training in these cases, we examine undiscounted potential-based reward shaping (PBRS_1), which drops the discount factor from the shaping term but still uses discounted targets when training reinforcement learning agents. Though PBRS_1 drops the guarantee to be objective-preserving, we show that when the potential function corresponds to the optimal state value function, the PBRS_1-shaped rewards correspond to the action advantage function of the undiscounted objective and reproduce an optimal policy. As in practice we do not have optimal values, we further present an error bound based on the estimation bias of the state value function. In our experiments, PBRS_1 beats PBRS and the correction term shaping with the same learned potential in several settings.

1 INTRODUCTION

In reinforcement learning, the discount factor γ is used to reduce the importance of future rewards. A lower discount weights a reward k steps ahead by γ^k, so rewards far beyond the effective horizon $1/(1-\gamma)$ contribute little to the expected returns and estimates average over less randomness and have lower variance (Jiang et al., 2015; Amit et al., 2020; Schulman et al., 2016). However, the γ also changes the objective the agent needs to optimize. Thus, in the important case where we want to optimize an undiscounted return, i.e., $\gamma = 1$, the agent might converge against a different optimum. In other words, if we do not want to mostly ignore costs beyond the effective horizon, we need to train the agent towards the undiscounted objective. For example, in MountainCar (Moore, 1990) a car has to rock itself up a hill and receives a reward of -1 per step until it reaches the top. Thus, its score is the negative number of required steps (capped at -200), a plain undiscounted sum. Though the goal lies about a hundred steps ahead, far beyond the horizon of 20 steps at $\gamma = 0.95$, a learner trained there barely sees the goal and learns poorly (Fig. 1a). We want to optimise the undiscounted return because this preserves the original optimisation task, and we want a training discount $\gamma_{\text{training}} < 1$ because a short horizon stabilizes training. Whether we optimise the original goal at a given training discount is decided by the *Blackwell limit*: an episodic task with finitely many states and actions has a Blackwell limit, a discount $\gamma_{\text{blackwell}} < 1$ above which every policy that is optimal for the discounted return is also optimal for the undiscounted return (Blackwell, 1962; Grand-Clément & Petrik, 2023). With a training discount $\gamma < \gamma_{\text{blackwell}}$ training will optimise a different objective. For example, add a decoy into MountainCar: a second exit on the left hill that the car reaches sooner than the true goal, while the flag pays a bonus larger than the extra steps it takes. The undiscounted optimum still drives to the true goal, but at every small enough discount the discounted optimum takes the decoy. The limit is a property of the task itself and is unknown before the task is solved, so a training discount chosen for stability cannot be known in advance to lie above it. A general approach to shift the reward structure of an MDP to offer more immediate rewards and, thus, shifting long term rewards inside the effective horizon is reward shaping. Reward shaping adds shaping terms to true rewards that mark the progress. Done freely, this might change the objective. The classic failure is an agent that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

loops through a rewarded region forever instead of finishing the task. Potential-based reward shaping (PBRS) is the standard safe form (Ng et al., 1999): choose a *potential* Φ — one number per state, scoring how good the state is — and replace the reward r of every transition $s \to s'$ by

$$\tilde{r} \;=\; r + \gamma_{\text{shaping}}\Phi(s') - \Phi(s), \qquad \gamma_{\text{shaping}} = \gamma \qquad (\text{PBRS}).$$

In PBRS, $\gamma_{\text{shaping}} = \gamma$, i.e., objective, training and reward shaping assume the same discounting. However, in cases where we choose γ based on training stability, but need to optimize an undiscounted objective, choosing an individual value for γ_{shaping} becomes a feasible option.

Ng et al. (1999) proved that PBRS does not change the optimal policy, for every Φ, and that it is the only shaping form built on state potentials with that guarantee. The theorem states that PBRS does not change the optimal policy of the discounted MDP. PBRS cannot move the optimum, so its guarantee can provably not fix a training discount set below the Blackwell limit.

In this paper, we study the following variant of PBRS: drop the γ from the shaping term,

$$\tilde{r} \;=\; r + \Phi(s') - \Phi(s) \qquad (\text{PBRS}_1),$$

while training still uses a $\gamma < 1$. We call this variant PBRS_1 and it allows to choose a γ value suitable for stable training while directing reward shaping towards an undiscounted objective. Under weak assumptions no PBRS shaping reproduces what PBRS_1 does at a training discount below 1 (§3). The case in between, a shaping discount $\gamma_{\text{shaping}} < 1$, but larger than the training discount γ, is analysed in Ng's thesis, but he excludes the undiscounted case with $\gamma_{\text{shaping}} = 1$ from the analysis (Ng, 2003, §2). That endpoint remained an open case which is relevant as many tasks actually require undiscounted objectives. In addition, Ng et al. (1999) already remarked that the undiscounted form might be worth trying "judiciously and with care". The potential commonly used when doing PBRS shaping is a learned value estimate $\hat{V}$ (Ng et al., 1999; Adamczyk et al., 2025). For this potential, we argue that breaking policy invariance is worth its price: the bias PBRS_1 introduces steers the learner towards the undiscounted optimum.

The entire difference between PBRS and PBRS_1 is one line of algebra:

$$\underbrace{\Phi(s') - \Phi(s)}_{\text{PBRS}_1} \;=\; \underbrace{\big[\gamma\Phi(s') - \Phi(s)\big]}_{\text{difference term} = \text{PBRS}} \;+\; \underbrace{(1-\gamma)\,\Phi(s')}_{\text{level term}}. \tag{1}$$

We call $(1-\gamma)\Phi(s')$ the *level term*. Adding such a correction on purpose — train at a low discount, add a reward term that points the objective at a longer horizon — is an idea with its own literature (Efroni et al., 2018; Tessler & Mannor, 2020; Cheng et al., 2021), described in §2.

We show that PBRS_1 guides training to an optimal policy for an undiscounted objective under the assumption that the potential function $\Phi(s)$ corresponds to the optimal state value function $V_1^*(s)$.

In practice we often only have an estimate of V_1^*. Thus, we present an error bound based on the bias between estimated and true state values. If Φ is within ε of V_1^* at every state, a greedy policy of the shaped problem gives up at most 2ε of total reward per step, at every training discount and with no factor $1/(1-\gamma)$ (Theorem 3). This makes the selection of γ less sensitive. Choosing γ is about variance and stability alone, and a too low value no longer makes the learner converge to a distorted objective, as long as Φ is close enough to V_1^*. Furthermore, on every transition but the last, PBRS_1, unlike either term alone, is independent of the level of Φ (§3).

To explain the impact of varying reward structures on PBRS and PBRS_1, we examine the following setting: delay a task's reward to the end of the episode and take the reward accumulated so far as the potential. Then, PBRS_1 rebuilds the dense task exactly (Theorem 4), and PBRS provably cannot (Theorem 1). We call a task *sparse* when it pays the episode total once, at the end. Correspondingly, a task is called *dense* when it is not sparse. Read in reverse, Theorem 4 says that every dense task is the PBRS_1-shaped version of its own sparse form. Turning a sparse reward back into a dense one is therefore an instance of PBRS_1 shaping, and we conjecture that it is practised as such without being recognised.

PBRS is widely used where Φ is handcrafted — a distance to a goal, a progress measure — and equally where Φ is learned: a value estimate carried over from an earlier or pretrained agent, the learner's own bootstrapped value (Adamczyk et al., 2025), a heuristic value fitted to offline data (Cheng et al., 2021). Our proposal is for that learned case: when the potential is a learned value

estimate, PBRS_1 is worth trying in place of PBRS, and in our empirical evaluation, it is the better of the two forms more often than not.

Contributions. We provide a theoretical and an empirical motivation to use PBRS_1 instead of PBRS or the level term of the reward-tweaking literature. First, with $\Phi = V_1^*$ the optimum of the PBRS_1-shaped problem is the original undiscounted optimum at every training discount $\gamma < 1$ (Theorem 2). Second, under a good choice of Φ, PBRS_1 is more robust than PBRS against a bad choice of γ (Remark 3). Third, under weak assumptions no PBRS shaping can imitate a PBRS_1 shaping (Theorem 1). Fourth, we show that PBRS_1 is a PBRS term plus the correction term of the reward-tweaking literature at an undiscounted target (§2). Fifth, PBRS_1 turns a sparse MDP into a dense one and back (Theorem 4). In our empirical evaluation, PBRS_1 scores above PBRS and the level term and is more robust to the choice of γ in several settings (§4). Code to replicate our results is in the supplementary material.

2 Related work

Reward shaping. Ng et al. (1999) introduced the safe form of adding a term to the reward to speed up learning, proving that the potential-based term $\gamma\Phi(s') - \Phi(s)$ leaves optimal policies unchanged. PBRS has been the standard answer to safe shaping since. In practice the potential Φ is often a learned estimate of the value function (Adamczyk et al., 2025). Another shaping related to what we propose is reward centering (Naik et al., 2024): it subtracts a scalar level to stabilise training.

Undiscounted PBRS. This line studies the same variant we do: shape with $\Phi(s') - \Phi(s)$ while training at $\gamma < 1$, and treat the deviation from the safe form as an error to be tolerated. Ng et al. (1999) proved the invariance of PBRS and mentioned the undiscounted variant only in passing, as possibly easier for an expert to specify and worth trying with care, without analysis. Ng's thesis (Ng, 2003) goes further: its Appendix 3.B trains at a discount below the shaping discount and bounds the loss, for a shaping discount strictly below 1, and it proposes the procedure of §4.1, a learned value estimate as potential with a reduced training discount. Grześ & Kudenko (2009) analyse it: they derive its bias, prove that it can change optimal policies, and give three reasons for accepting it anyway: the bias is small for $\gamma \approx 1$, a slightly suboptimal policy learned fast is often worth more than an optimal one learned slowly, and the undiscounted form is robust against exploration cycles. Later work keeps this reading (Grześ, 2010; 2017). In deep RL, Jeon et al. (2023) set the shaping discount to 1 for humanoid locomotion because training is more stable, noting the guarantee "is technically sacrificed". Palmer et al. (2025), shaping with the value functions of earlier best-response iterations, set it to 1 while training at 0.99 and leave evaluating the choice as future work. The closest prior treatments of the choice between PBRS and PBRS_1 are the single-discount tabular comparison of Grześ & Kudenko (2009); Grześ (2010) and Palmer et al. (2025), who name the open question outright. For the undiscounted form the residual is read as a defect, never as pointing anywhere. What it lacks is an account of *which* Φ makes the residual beneficial, of how that depends on the training discount, and a connection to the correction-term methods below, which Efroni et al. (2018) draw only for Ng's thesis at a shaping discount below 1.

Correction terms for a longer-horizon optimum. A separate line addresses the gap between the training discount and the undiscounted objective: training at $\gamma < 1$ is statistically convenient but optimises the wrong objective. Its fix is to keep the low training discount and add a reward term that moves the optimum back out to a longer horizon. The added terms have the form $(\gamma_{\text{hi}} - \gamma_{\text{lo}})\, h(s')$ for a heuristic h, where γ_{lo} is the discount actually used in training and γ_{hi} is the longer-horizon discount whose optimum the method targets. Within this line, κ-PI, a policy-iteration method (Efroni et al., 2018), mixes cheap short-horizon updates with a correction so that the fixpoint is the optimum of the longer horizon. HuRL, heuristic-guided reinforcement learning (Cheng et al., 2021), shapes with a heuristic h and a shortened horizon and proves the bias vanishes when h is the optimal value of the horizon they target. Reward tweaking (Tessler & Mannor, 2020) learns a surrogate reward whose discounted optimum is the undiscounted one. The heuristic in this line is frequently a *learned* value estimate, the same object the PBRS construction uses as Φ (Cheng et al., 2021; Tomar et al., 2020). So the two lines share the construction — which is why a single learned estimate $\hat{V}$ can serve as the Φ of all three shaped rewards our experiments compare (the level term, PBRS and PBRS_1 shapings of §4.1). Cheng et al. (2021) also print the algebra for discounted pairs: HuRL trains on

the blended reward $r + (1 - \lambda)\gamma h(s')$ at the shortened discount $\lambda\gamma$, and their §4.2 notes that this surrogate is equivalent "up to a PBRS transformation" to the PBRS-shaped reward $r + \gamma h(s') - h(s)$ at that shortened discount. What none of these works state is the case $\gamma_{\text{hi}} = 1$ of that algebra: written as a shaping rule with one γ-free potential, their correction with the target horizon pushed to the undiscounted one, *plus* a PBRS term, is precisely the undiscounted variant that the line above tolerates as an error. That identification is one of this paper's contributions: the undiscounted line treats the level term as an error to tolerate, the correction line adds it on purpose with $\gamma_{\text{hi}} = 1$, and their sum is PBRS$_1$. Simply raising γ is not a fix, for the variance reasons of §1 (Fedus et al., 2020). To the best of our knowledge our contributions are new. *The comparison:* no prior work compares PBRS and PBRS$_1$ across training discounts and environments, the prior treatments above covering one discount or one domain each. *The identity:* Cheng et al. (2021) and Efroni et al. (2018, Prop. 11) print the correction-plus-PBRS algebra only at discounted pairs, and no work reads its case $\gamma_{\text{hi}} = 1$ as the variant Ng et al. (1999) left open. GAE (Schulman et al., 2016, §4) comes closest: at $\gamma = 1$ its estimator is the λ-discounted sum of PBRS$_1$-shaped rewards with the critic as Φ, used inside the policy gradient without separating the shaping discount from γ. *The fixpoint and perturbation theorems:* the shaping and correction lines prove theirs at a discounted target (Remark 2), Tessler & Mannor (2020, Thm. 1) reach the undiscounted optimum with the level term alone, not a potential-based shaping, and give neither a $\tilde{Q}^*$ identity nor a bound, and the existing bounds (Ng, 2003; Cheng et al., 2021) hold only for shaping discounts strictly below 1. *The reconstruction and separation theorems:* at $\gamma = 1$ the first is the return-decomposition identity (Arjona-Medina et al., 2019; Chan et al., 2024). We found no statement that it stays a PBRS$_1$ shaping at every $\gamma < 1$, and no proof that PBRS shaping cannot create PBRS$_1$ shaping.

3 Theoretical analysis of PBRS$_1$

Setup. We work in an episodic Markov decision process (MDP). In the MDP the agent observes a state s, takes an action, receives a reward r and moves to a state s'. An episode ends either by *termination*, the task's own end condition, or by *truncation*, a time limit. The objective of the task is the total undiscounted reward of an episode. An agent is trained to learn a policy that maximises this objective, and the learner's training discount is $\gamma < 1$. On the transition that ends the episode we set $\Phi(s') = 0$, a convention we call *gating*. Gating is not a device of ours: policy invariance in an episodic task requires the potential to vanish at the episode end, as Ng et al. (1999) require for episodic tasks and Grześ (2017) sets out in detail. Throughout we assume the *proper episodic setting*, which consists of four standard conditions. The state and action spaces are finite. Under every policy the episode ends with probability 1 and in finite expected time. The episode end is a function of the state, and in particular the step counter is part of the state whenever the end is a time limit. Rewards and potentials are bounded. Under these conditions the undiscounted values below exist and are finite (Bertsekas & Tsitsiklis, 1991).

Notation. $V^*_\gamma(s)$ is the optimal γ-discounted value of state s. $V^*_1(s)$ is the optimal *undiscounted* value of state s, that is, the total reward the best policy collects from s. $Q^*_1(s, a)$ is the optimal undiscounted state-action value function: the total reward the best policy collects from s when action a is taken first. $A^*_1(s, a) = Q^*_1(s, a) - V^*_1(s)$ is the undiscounted advantage, which is zero for a best action and negative otherwise. A *greedy* policy takes the action with the highest action value. Shaping replaces the reward the agent trains on by $\tilde{r}$, and $\tilde{Q}^*_\gamma(s, a)$ and $\tilde{V}^*_\gamma(s)$ are the optimal values of the shaped reward at training discount γ. $V^\pi_1(s)$ is the undiscounted value of the policy π at s. Where the state is clear from context, we drop the argument and write, e.g., V for $V(s)$ and Φ for $\Phi(s)$. The four shapings of this paper are the 2×2 over the two terms of Eq. (1):

	level term off	level term on
difference term off	vanilla: $\tilde{r} = r$	level term: $\tilde{r} = r + (1 - \gamma)\Phi(s')$
difference term on	PBRS: $\tilde{r} = r + \gamma\Phi(s') - \Phi(s)$	PBRS$_1$: $\tilde{r} = r + \Phi(s') - \Phi(s)$

Overview. Proposition 1 states Eq. (1) formally. Theorem 1 separates PBRS$_1$ from every PBRS shaping. Theorems 2 and 3 take a (learned) value function as potential, and Theorem 4 the accumulated reward. Proofs are in Appendix A.

Proposition 1 (Decomposition). *For every Φ, every γ and every transition (s, s'), including a transition into a terminal state, where $\Phi(s') = 0$,*

$$\underbrace{\Phi(s') - \Phi(s)}_{\text{PBRS}_1} \;=\; \underbrace{\big[\gamma\Phi(s') - \Phi(s)\big]}_{\text{PBRS}} \;+\; \underbrace{(1-\gamma)\,\Phi(s')}_{\text{level term}}.$$

Hence PBRS_1 *and the* level term *shaping differ by a* PBRS *term and have the same optimal policy set at every γ.*

Remark 1 (What transfers, and what does not). *Proposition 1 writes* PBRS_1 *as a* PBRS *term plus a level term, and each of the two literatures covers one of them. The* PBRS *term transfers exactly: it leaves the optimal policy set unchanged, so the* level term *shaping and* PBRS_1 *share their optimal policies at every γ. The level term transfers only in form: it is the correction literature's $(\gamma_{\text{hi}} - \gamma_{\text{lo}})\,h(s')$, which moves the optimum from the training discount γ_{lo} to a longer horizon γ_{hi}, at $\gamma_{\text{hi}} = 1$, $h = \Phi$. But that literature's bounds are stated for $\gamma_{\text{hi}} < 1$ and carry factors $1/(1-\gamma_{\text{hi}})$ that blow up at 1 (Efroni et al., 2018; Cheng et al., 2021). At $\gamma_{\text{hi}} = 1$ and $h = \Phi$ that literature's condition on the heuristic h becomes a condition on the potential Φ. The one fixpoint result at $\gamma_{\text{hi}} = 1$, Tessler & Mannor (2020, Thm. 1), is stated for the level term alone (taken under π^*); Theorems 2 and 3 prove the case $\gamma_{\text{hi}} = 1$ for the* PBRS*-form shaping.*

One consequence of the identity deserves a name, *level independence*: add a constant k to Φ, and on every non-final transition the PBRS_1 shaped reward is unchanged, while the PBRS shaped reward shifts by $-(1-\gamma)k$ and the level term shaped reward by $+(1-\gamma)k$. Neither component has the property. Their combination creates it (§5 gives an applied case).

PBRS_1 **is not a re-tuned** PBRS**.** Before analysing PBRS_1 we settle whether PBRS_1 is a method distinct from PBRS. Whenever the potential carries information the rewards do not, no PBRS shaping matches PBRS_1's shaped returns, even up to a constant. Theorem 1 shows this rigorously. First we define an informative potential.

Definition 1 (Informative potential). Φ is *informative* if the rewards alone do not already determine it: two reachable trajectories start in the same state and produce the same rewards, yet accumulate different potential,

$$s_0(\tau_1) = s_0(\tau_2), \qquad r_t(\tau_1) = r_t(\tau_2)\ \ \forall t, \qquad S_\gamma(\tau_1) \neq S_\gamma(\tau_2),$$

where $S_\gamma(\tau) = \sum_{t=1}^{T-1} \gamma^{t-1}\Phi(s_t)$ is the discounted potential collected strictly between start and end.

Theorem 1 (PBRS shaping cannot create PBRS_1 shaping). *Let $\gamma < 1$ and let Φ be informative. In words: no* PBRS *shaping, with any potential and any shaping discount, reproduces the* PBRS_1*-shaped return of every trajectory, even up to a constant. Formally, there is no potential Ψ (a function of the state), discount $\tilde{\gamma} \in [0, 1]$ and constant κ such that*

$$\sum_{t=0}^{T-1} \tilde{\gamma}^t \big[r_t + \tilde{\gamma}\,\Psi_{t+1} - \Psi_t\big] + \kappa \;=\; \sum_{t=0}^{T-1} \gamma^t \big[r_t + \Phi(s_{t+1}) - \Phi(s_t)\big]$$

holds on every reachable trajectory, where Ψ_t is Ψ evaluated at time t and both potentials are gated.

The theorem separates shaped *returns*, not optima: two shaped rewards can differ on every trajectory and still be optimised by the same policy. What it rules out is reconstruction.

Potential Φ_V: a (learned) value function. Because PBRS_1 gives up policy invariance, the first question is which optimum the shaping installs instead. The next theorem answers it for Φ being a value function: feed PBRS_1 the undiscounted optimal value V_1^* and the training discount drops out of the answer — at every $\gamma < 1$ the shaped optimum is the policy that maximises total undiscounted reward.

Theorem 2 (Fixpoint, discount-independence). *Let $\Phi = V_1^*$, the undiscounted optimal value, and shape with* PBRS_1*. Then for every training discount $\gamma \in [0, 1)$ the shaped optimal action-values satisfy $\tilde{Q}_\gamma^* = A_1^*$ (the undiscounted advantage) and $\tilde{V}_\gamma^* \equiv 0$: the greedy policy is undiscounted-optimal at every such γ.*

In words: under the shaping of Theorem 2 the learner has nothing to do but pick the best next action, and the learner trains at any training discount and solves the undiscounted problem.

Remark 2 (The closest statements in the two literatures)**.** *Both lines name an optimal value as the ideal potential: Cheng et al. (2021) prove that their bias vanishes when the heuristic is the optimal value of the horizon they target, and Corollary 2 of Ng et al. (1999), at the matched discount, gives* $\tilde{Q}^*_\gamma = Q^*_\gamma - \Phi$*, so* $\Phi = V^*_\gamma$ *yields the* γ*-discounted advantage. Tessler & Mannor (2020, Thm. 1) prove the policy-level statement at* $\gamma_{\text{hi}} = 1$ *for the level-term shaping alone,* $r + (1-\gamma)P^{\pi^*}V^*_1$*. Theorem 2 proves it for the* PBRS_1 *shaping, whose extra* $\gamma\Phi(s') - \Phi(s)$ *is Ng-invariant and so leaves that optimum in place, but shifts the values:* $\tilde{Q}^*_\gamma = A^*_1$ *and* $\tilde{V}^*_\gamma \equiv 0$*, where the level term alone gives* $\tilde{Q}^*_\gamma = Q^*_1$*.*

Remark 3 (Discount robustness)**.** *One* stationary, γ-free *potential serves every training discount at once, whereas the ideal potential of* PBRS*,* V^*_γ*, moves with the training discount. We expect this freedom to weaken rather than vanish when* Φ *is a learned estimate of* V^*_1*: the object* PBRS_1 *aims at is then* γ*-free and mis-specified by estimation error alone, while under* PBRS *the same error is compounded by a target that moves with the training discount. This is an intuition, not a theorem, but testable: a flatter* γ*-response for* PBRS_1 *than for* PBRS *under a learned potential (§4.1, Figs. 1 and 5).*

In practice we have only an estimate of V^*_1. The next theorem bounds the cost of an ε-wrong potential, uniformly in the training discount.

Theorem 3 (Perturbation, finite episodes)**.** *Let* $\Phi = V^*_1 + \eta$ *with* $\sup_s |\eta(s)| \le \varepsilon$ *(at a gated end both* Φ *and* V^*_1 *contribute* 0*), shape with* PBRS_1 *and train at any* $\gamma \in [0, 1)$*. Then the shaped optimal action-values satisfy* $\|\tilde{Q}^*_\gamma + \Phi - Q^*_1\|_\infty \le \varepsilon$*, and any greedy policy* $\tilde{\pi}$ *of* $\tilde{Q}^*_\gamma$ *satisfies, for every start state* s*,*

$$V^*_1(s) - V^{\tilde{\pi}}_1(s) \;\le\; 2\varepsilon\, \mathbb{E}^{\tilde{\pi}}_s[T],$$

where T *is the number of transitions before the episode ends under* $\tilde{\pi}$ *(finite in expectation in the proper episodic setting).*

Remark 4 (How Φ is obtained in practice)**.** V^*_1 *is not available during training. The* Φ_V *experiments therefore use the potential a practitioner already has. One* vanilla *agent is trained beforehand at* $\gamma = 0.99$*, its value estimate is frozen, and that one function is the* Φ *of the three shaped rewards at every training discount and for every seed (Appendix B). This potential is the value of the pretrained policy and not of the optimal one, it is discounted at* 0.99 *and not undiscounted, and it carries estimation error. Theorem 3 absorbs all three departures from* V^*_1 *into one* ε*, uniformly in* γ*. Section 4 measures what remains, and Appendix F varies the source policy on purpose. Relative to* PBRS *the cost is zero: both use the same pretrained estimate (Ng, 2003).*

Remark 5 (Relation to Ng's mismatched-discount bound)**.** *Theorem 3 is the episodic, shaping-discount-*1 *counterpart of Theorem 5 in Appendix 3.B of Ng (2003) (§2). The proof follows his with one step replaced: his value-to-policy step carries a factor* $2\gamma/(1-\gamma)$ *at the shaping discount, infinite at* 1*. We use instead the identity that writes the gap between two policies' undiscounted returns as a sum of advantages along an episode, so the horizon factor becomes the expected episode length. That shaping discount* 1 *admits a bound at all is what is new.*

Potential Φ_{cum}: the accumulated reward. Whether a task pays its reward at the end or as it is earned is a routine reward-design choice: terminal-only reward is the hard case for credit assignment, and the standard remedy pays intermediate rewards summing to the same episode total. We ask whether that remedy is a shaping, and of which family. Given an episodic base MDP with per-step reward r, its *delayed-reward* MDP (*lump* MDP for short) has states (s, c), where c is the reward accumulated so far: $c_0 = 0$ and $c_{t+1} = c_t + r_t$. Every transition pays 0 except the last (terminated or truncated), which pays the whole episode return at once.

Theorem 4 (Exact reconstruction)**.** *Let* $s_0, r_0, s_1, \ldots, r_{T-1}, s_T$ *be any trajectory of the base MDP, ending at* s_T *by termination or truncation. Run* PBRS_1 *on the delayed-reward MDP with* $\Phi(s, c) = c$*, gated at the trajectory's end. Then the shaped delayed reward satisfies* $\tilde{r}_t = r_t$ *for every* $0 \le t \le T-1$*: the shaped delayed-reward MDP and the dense MDP produce identical reward sequences on every trajectory.*

Remark 6 (What the reconstruction shows)**.** *The shaped learner sees every reward at the step it was earned.* PBRS *cannot do this: with any potential it leaves the optimal policies of the delayed task where they are, and Theorem 1 shows that it cannot reproduce the dense rewards.*

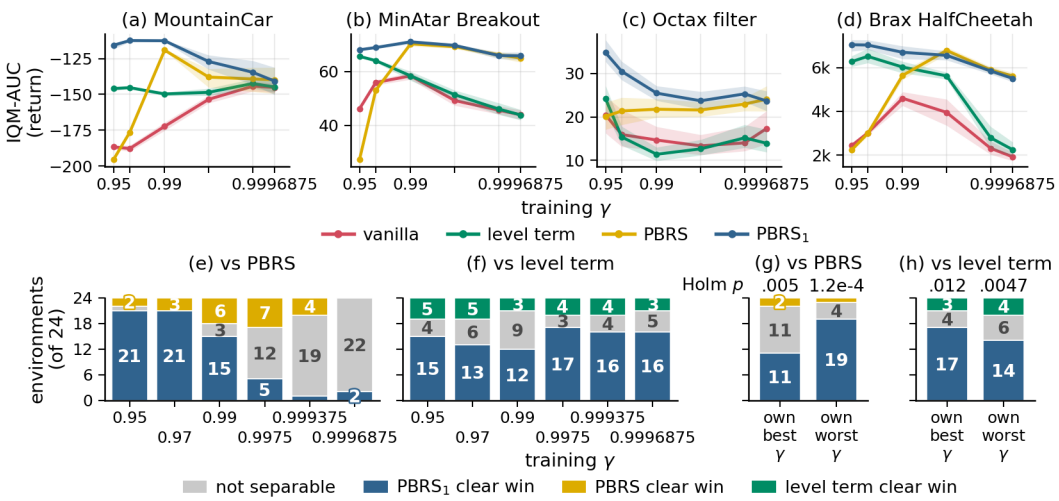

Figure 1: Results under the learned-value potential Φ_V. **(a–d)** Response to the training discount in four of the 24 environments: (a) MountainCar, (b–d) one environment from each other suite. All 24 are in Fig. 5. **(e, f)** At each γ, the environments (of 24) in which PBRS_1 is a clear win (blue), the compared shaping is (its own colour), or the two are not separable (grey): (e) against PBRS, (f) against the level term. **(g, h)** The same counts with each shaping at its own best γ and at its own worst γ, with the Holm-corrected p of the paired Wilcoxon test over the 24 per-environment differences above each bar: (g) against PBRS, (h) against the level term. Full tables: Appendix I. This figure with a best five-point moving-window score: Fig. 2. Learning curves of all 288 cells: Appendix E.

With Theorem 1, Theorem 4 answers which family the conversion belongs to, at every $\gamma < 1$. At $\gamma = 1$ the conversion is the return-decomposition identity of the credit-assignment literature (Arjona-Medina et al., 2019; Chan et al., 2024). What the theorems add is that it survives discounting only on the PBRS_1 side. §5 draws the wider reading: PBRS_1 shaping is far more prevalent than it looks, because every task that pays intermediate rewards and trains at $\gamma < 1$ is doing PBRS_1 shaping of its delayed counterpart, not PBRS shaping.

4 Experiments

Questions. We ask two questions under the learned-value potential Φ_V. *(1) Can* PBRS_1 *beat* PBRS *and the level term?* We compare each shaping at its own best γ, so PBRS_1 cannot win merely by acting like a higher discount. *(2) Is* PBRS_1 *more robust to the choice of* γ *than* PBRS *and the level term?* We compare each at its own worst γ. Under the accumulated-reward potential Φ_{cum} we ask: *(3) Is paying rewards earlier in a sparse task, which by Theorem 4 is* PBRS_1 *shaping, good practice, or should one apply the policy-invariant* PBRS *instead?*

Protocol. The four shapings of the 2×2 grid of §3, vanilla included, run on 24 environments. Classic control and MinAtar (Young & Tian, 2019) use PQN, a value-based learner (Gallici et al., 2025), while Brax (Freeman et al., 2021) and Octax (Radji et al., 2025) use PPO. We run two potentials, and the two answer different questions. The first potential, Φ_V, is the potential practitioners shape with under PBRS: the frozen vanilla value estimate of Remark 4. This estimate is neither optimal nor undiscounted, so the outcome of the comparison is open, and every comparison under Φ_V is conditional on that one estimate (Appendix B). An environment enters the suite only if the agent that supplies Φ_V learns in that environment, because otherwise Φ_V carries no information (Appendix H). The second potential, Φ_{cum}, is the reward accumulated so far on the delayed-reward version of each task, the potential of Theorem 4. The Φ_{cum} runs confirm that the PBRS_1 shaping practitioners already perform, paying dense rewards instead of a terminal one, beats the PBRS shaping of the same task. All four shapings see the accumulated reward as one extra input feature, and the baseline under Φ_{cum} is vanilla on the delayed task. Every environment runs under both potentials at $\gamma \in \{0.95, 0.97, 0.99, 0.9975, 0.999375, 0.9996875\}$, the usual range around 0.99. Performance

is measured as the undiscounted return, and the best (worst) training discount of a shaping is the one maximising (minimising) it. A *cell* is one combination of environment, training discount and potential together with its four shapings, and the campaign has 288 cells. Within a cell the four shapings see the same number of environment steps, the same hyperparameters and the same frozen potential, and each seed's random streams and initial network weights are bit-identical across shapings and training discounts (common random numbers). Brax and Octax use 50 seeds per shaping and the other environments use 100. No hyperparameter is tuned per cell, and PBRS, PBRS_1 and the level term are given the same amount of compute. The two potentials are gated differently. Under Φ_{cum} we gate as in the Setup of §3, zeroing Φ at the trajectory end whether the episode terminates or is truncated: Theorem 4 needs exactly this, because the lump payout and the gate then trigger on the same event. Under Φ_V the gate follows each learner's own value target, so that $\Phi(s')$ and the bootstrap treat every episode end alike: on ten environments (Octax, Breakout- and SpaceInvaders-MinAtar) the learner ends the episode at the time limit and Φ is zeroed there. The other fourteen (classic control, Brax, Asterix- and Freeway-MinAtar) do not treat the time limit as terminal in the value target and zero Φ at terminations only (Appendix C). Within a cell the gate is the same for all four shapings, so the 2×2 grid and Proposition 1 are untouched.

We expect PBRS_1 to help most on tasks where the horizon is long compared to $1/(1-\gamma)$ at the discounts a learner can afford. Where a large γ is affordable, PBRS and PBRS_1 coincide as $\gamma \to 1$.

Scoring. Each seed is scored by the mean of its undiscounted raw episode return, the objective of every task here, over the evaluation points of training (the area under the learning curve). The metric is discount-free, so tuning γ per shaping compares methods, not tasks. The seeds of one shaping in one cell are pooled by the interquartile mean (Agarwal et al., 2021). Within each environment and potential, the 24 scores of the four shapings at the six discounts are min–max rescaled over those 24: a within-environment ranking, not an effect size, which defines each shaping's best and worst γ and enters the tests below. Within an environment, PBRS_1 is a *clear win* against another shaping when it scores higher and the two cells' per-seed scores separate at the 5% level (two-sided Mann–Whitney), a *clear loss* when it scores lower and they separate, *not separable* otherwise. These counts are descriptive and uncorrected, and a cell's seeds are trained together in one or two vectorised processes, so no single cell is treated as evidence. The claims are tested across environments: one two-sided Wilcoxon signed-rank test on the 24 paired differences of normalised score per claim, Holm-corrected over the eight claims (two comparisons, two compared shapings, two potentials), which are never pooled. All numbers (Appendices I and J) are regenerated by the replication bundle in the supplementary material. Overall, PBRS_1 scores higher than both of its component shapings in both comparisons under both potentials, and all eight paired tests are significant after correction (Fig. 1g–h; Table 3).

4.1 Learned-value potential

The clearest comparison is the worst-γ score: at its own worst γ PBRS_1 is clearly above PBRS in 19 environments against one clear loss (Fig. 1g), and it has the highest worst-γ score of the four shapings in 18 of 24 environments. At the standard discount it is a clear win against PBRS in 15 environments (6 clear losses). The tuned-γ comparison is the narrowest: PBRS_1 at its own best γ still scores above PBRS at its own best γ in 21 of 24 environments, but only 11 are clear wins and 11 do not separate. The direction survives choosing each shaping's γ on half of the seeds and scoring on the other half (Appendix I). So the gain is not simply a higher effective discount: the best training discount stays below 1, and PBRS_1 wins there. This also explains why it does not speak against our claims that PBRS_1 and PBRS barely separate at the highest discounts (Fig. 1e), where the two must coincide as $\gamma \to 1$: the optimum usually does not lie at the top of our grid, which is the best training discount of PBRS_1 in only 3 and of PBRS in only 3 of the 24 environments. PBRS_1 does not beat PBRS everywhere. We conjecture that Φ is then not close enough to V_1^*, the failure Theorem 3 bounds. That distance is not observable, because V_1^* is unknown, so we do not test the conjecture directly. Against the level term alone, PBRS_1 is a clear win in 12 environments, a clear loss in 3 at $\gamma = 0.99$, and 17 against 3 with each shaping at its own best γ. At each shaping's own worst γ it is clearly above the level term in 14 environments against 4 clear losses. Both tested differences, at the tuned and at the worst γ, are significant after correction (Table 3), so the gain needs the full PBRS_1 shaping and not the level it adds alone.

4.2 Accumulated-reward potential: the dense task as PBRS_1 of the sparse one

Under Φ_{cum} each task is transformed into a sparse one by paying the episode's total reward once, at its end. PBRS_1 with the accumulated reward as potential turns it back into the dense task (Theorem 4), so this comparison asks whether paying dense rewards is better than applying PBRS to the sparse ones. At the standard $\gamma = 0.99$ the dense task is clearly better than PBRS in 21 of the 24 environments and clearly worse in one. With each shaping at its own tuned γ it is clearly better in 12 against 3, and at each shaping's own worst γ in 23 against none. Both tested differences, at the tuned and at the worst γ, are significant after correction (Table 3 in the appendix), and the curves of all 24 environments are in Fig. 6 in the appendix.

5 Discussion

Undiscounted shaping is applied in practice often without a significant theoretical justification. The undiscounted form surfaces as a local fix, justified by convenience or stability, not by its effect on the optimum. HOP-CPP (Blake & Schubert, 2026) declines the form $\gamma\Phi(s') - \Phi(s)$ because its large-magnitude potential would need $\gamma > 0.999$ to keep the shaping term from dominating the task reward, and shapes undiscounted instead. Widely used simulators ship the form unremarked as a task reward: the locomotion tasks of PyBullet (Coumans & Bai, 2016–2021) and Isaac Gym (Makoviychuk et al., 2021) reward progress $\Phi(s') - \Phi(s)$ toward a target under training at $\gamma = 0.99$, and PyBullet's source reads its alive bonus as the leak $(1 - \gamma)\Phi$ of a potential, the level term. These reports are consistent with our grid (Fig. 3c, Appendix D). They are observations, not a prevalence claim.

What the reconstruction theorem says about dense rewards. Theorems 1 and 4 admit a reading beyond the Φ_{cum} experiments. Converting a delayed reward into per-step rewards that decompose it exactly is common, and that conversion sits inside the PBRS_1 family and outside the PBRS family. So whenever such a dense reward trains better than its delayed version under discounted training, that is, by Theorem 4, an instance of PBRS_1 beating the unshaped baseline through a shaping the invariant form provably cannot reproduce (Theorem 1). Even where the delayed task itself is of little practical interest, the dense-versus-delayed comparison illustrates how PBRS_1 modifies the reward structure of sparse tasks.

Limitations. *The guarantee is gone.* PBRS_1 forfeits PBRS's policy invariance and pays for it where the level term misinforms. So a wrong Φ costs PBRS_1 more than PBRS. Some environments sit outside the Setup of §3, and the gating of Φ differs between stacks. We remark on both in Appendix C. We make a recommendation inside the shaping family, not a state-of-the-art claim. The grid compares four shapings of one learner at matched budget, matched hyperparameters and no per-cell tuning, so what it establishes is which shaping form to prefer once shaping has been chosen. The unshaped arm is a control of that design rather than a tuned agent, and no method outside the family is run at all. The recommendation we draw is addressed to practitioners already shaping with PBRS or the level term — try PBRS_1 — and is not a claim that PBRS_1 is the best available way to learn these tasks.

6 Conclusion

Training at $\gamma < 1$ lowers variance but can change the task: below the Blackwell limit, the discounted optimum need not maximise the original, undiscounted return. To repair the objective, the correction-term literature adds a level term that moves the optimum towards a longer horizon. PBRS is used to speed up learning, but it keeps the discounted optimum and so cannot repair the objective. PBRS_1 is exactly the sum of PBRS and the level term (Proposition 1). Practitioners already drop the γ from the shaping term because it works better, but they have no account of why (§5). Our theory gives that account: with $\Phi = V_1^*$, PBRS_1 recovers the undiscounted optimum at every training discount, an error ε in Φ costs at most 2ε per step, and with the accumulated reward as potential PBRS_1 turns a delayed reward into the dense one, which PBRS cannot (§3). Empirically, PBRS_1 scores above PBRS and the level term in a wide range of settings.

Disclosure of generative AI use

We used generative AI assistants for research ideation and execution: they helped develop the conceptual framework, propose and refine hypotheses, formulate the mathematical claims and find and write their proofs, design the methodology and experiments, implement the training, analysis and plotting code, and interpret the results. We also used them for retrieval and discovery of related work and its summary, to draft sections of the paper, and to edit the manuscript for readability. We did not use them to generate synthetic data or for translation. Dataset cleaning and qualitative or thematic data analysis do not apply to this work, whose data come from simulated environments and are analysed quantitatively. All AI-assisted work was reviewed by the authors: proofs were checked, the central identities are verified numerically in code included in the supplementary material, and suggested literature was read before being cited. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Reproducibility statement

The assumptions of every theorem are stated in the Setup of §3, and complete proofs, with a numerical check of each identity, are in Appendix A. The protocol and scoring are described in §4. Appendix B lists every hyperparameter of the three training stacks (Table 1), how both potentials are constructed, the evaluation protocol and the compute used. The environment suite and its selection are in Appendix H (Table 2), the gating conventions in Appendix C, the statistical tests in Appendix I, and the per-cell results behind every count in Appendix J. The replication bundle in the supplementary material contains the trainers of all three stacks, the shared shaping module and its tests, the frozen potentials of the 24-environment grid, the raw per-seed results of every reported run, and the scripts that regenerate the paper's numbers, tables and figures.

References

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V. Kulkarni. Bootstrapped reward shaping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Ron Amit, Ron Meir, and Kamil Ciosek. Discount factor as a regularizer in reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 269–278, 2020.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, 2019.

Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2): 719–726, 1962.

Johann Blake and Matthias Schubert. Hyperbolic maps and objective preserving rewards for learning coverage paths in unknown environments. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, Lecture Notes in Computer Science, pp. 236–252. Springer, 2026. doi: 10.1007/978-3-032-37664-0_14.

Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in reinforcement learning from human feedback. In *International Conference on Machine Learning*, 2024.

Ching-An Cheng, Andrey Kolobov, and Adith Swaminathan. Heuristic-guided reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2021.

Sam Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 433–440, 2012.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Beyond the one-step greedy approach in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 1387–1396, 2018.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. In *International Conference on Learning Representations (ICLR)*, 2020.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax – a differentiable physics engine for large scale rigid body simulation. *arXiv preprint arXiv:2106.13281*, 2021.

Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob N. Foerster, and Mario Martin. Simplifying deep temporal difference learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Julien Grand-Clément and Marek Petrik. Reducing Blackwell and average optimality to discounted MDPs via the Blackwell discount factor. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Marek Grześ. *Improving Exploration in Reinforcement Learning through Domain Knowledge and Parameter Analysis*. PhD thesis, University of York, 2010. Section 5.7.

Marek Grześ. Reward shaping in episodic reinforcement learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 565–573, 2017.

Marek Grześ and Daniel Kudenko. Theoretical and empirical analysis of reward shaping in reinforcement learning. In *International Conference on Machine Learning and Applications (ICMLA)*, pp. 337–344, 2009.

Se Hwan Jeon, Steve Heim, Charles Khazoom, and Sangbae Kim. Benchmarking potential based rewards for learning humanoid locomotion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1181–1189, 2015.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Andrew William Moore. *Efficient Memory-based Learning for Robot Control*. PhD thesis, University of Cambridge, 1990. Technical Report UCAM-CL-TR-209.

Abhishek Naik, Yi Wan, Manan Tomar, and Richard S. Sutton. Reward centering. In *Reinforcement Learning Conference (RLC)*, 2024.

Andrew Y. Ng. *Shaping and Policy Search in Reinforcement Learning*. PhD thesis, University of California, Berkeley, 2003.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287, 1999.

Gregory Palmer, Luke Swaby, Daniel J. B. Harrold, Matthew Stewart, Alex Hiles, Chris Willis, Ian Miles, and Sara Farmer. An empirical game-theoretic analysis of autonomous cyber-defence agents. *arXiv preprint arXiv:2501.19206*, 2025.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning*, 2018.

Waris Radji, Thomas Michel, and Hector Piteau. Octax: Accelerated CHIP-8 arcade environments for reinforcement learning in JAX. *arXiv preprint arXiv:2510.01764*, 2025.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Chen Tessler and Shie Mannor. Reward tweaking: Maximizing the total reward while planning for short horizons. *arXiv preprint arXiv:2002.03327*, 2020.

Manan Tomar, Yonathan Efroni, and Mohammad Ghavamzadeh. Multi-step greedy reinforcement learning algorithms. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, 2020.

Eric Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

Kenny Young and Tian Tian. MinAtar: An Atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.