The Goal-Directed Frame for General Agents

Geraud Nangue Tasse, Steven James, Benjamin Rosman
{geraud.nanguetasse1,steven.james,benjamin.rosman1}@wits.ac.za

University of the Witwatersrand, South Africa

Abstract

Reinforcement learning is often framed around episodic, discounted, or average scalar rewards. While useful, these views miss a core principle: *generally intelligent agents, whether biological or artificial, are fundamentally goal-directed*. They continually select and pursue evolving goals, a process not captured well by fixed episodes or simple discount factors. Building on insights from neuroscience, we argue that goals and termination should be agent-defined, not externally imposed. To this end, we propose reframing the agent as learning a General Value Function where the goals to be achieved and their corresponding rewards are not environment given, but rather agent-selected to both maximise standard environment rewards while capturing the optimal expected return for reaching any goal. Importantly, they can be learned *model-free* from a single stream of continual experience through a minimal modification to the standard reinforcement learning loop that encodes the *goal-directed prior* in the agent. This better reflects real-world learning as continual, goal-driven interaction, leading to emergent world models and drawing a closer connection between reinforcement learning, planning and control. Hence, this framework naturally unifies existing reinforcement learning paradigms and provides a powerful foundation for designing more general, adaptable agents.

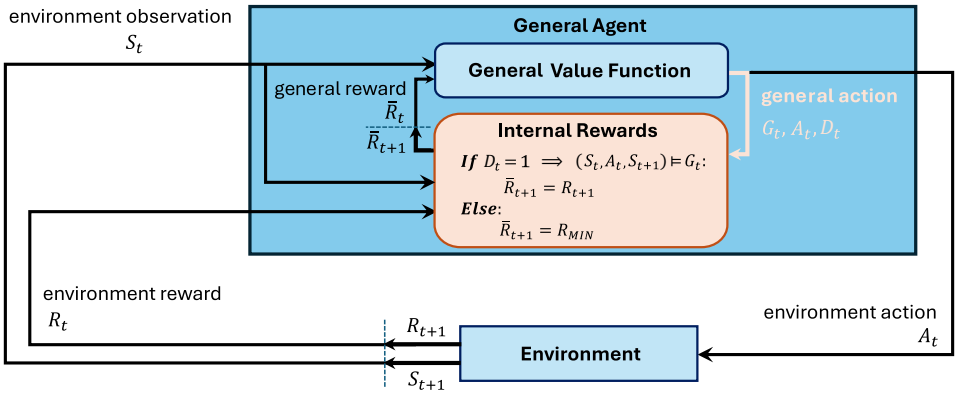

Figure 1: Expanding the standard RL loop (Sutton et al., 1998) to highlight the goal-directed frame. At each timestep t, the agent observes S_t, decides which goal to achieve G_t, decides which environment action to take A_t, decides whether to terminate D_t, transitions to S_{t+1}, and recieves the environment reward R_{t+1} if termination implies goal satisfaction, else self-inflicts a penalty R_{MIN}.

1 Introduction

Reinforcement learning (RL) has traditionally been framed through a handful of well-established paradigms: episodic RL, where tasks are divided into discrete episodes; discounted RL, which biases near-term rewards; and average-reward RL, which assumes continual interaction and optimises

steady-state performance (Sutton et al., 1998). Each of these is tailored to particular theoretical conveniences which have proven useful in specific contexts. But collectively, they obscure a more fundamental perspective: converging evidence from neuroscience, psychology, and robotics suggests that *goal-directedness* is a deeper organising principle (Daw & Shohamy, 2008; Stachenfeld et al., 2017; Taniguchi et al., 2018; Reed et al., 2022; Singh et al., 2022).

Recordings in rodents show that grid cells in the medial entorhinal cortex and place cells in the hippocampus encode predictive maps of likely future states, effectively binding spatial inference to latent reward expectations (Hafting et al., 2005; Stachenfeld et al., 2017). Functional-imaging work indicates that prefrontal and striatal circuits dynamically select and update behavioural goals rather than passively discount rewards (Daw & Shohamy, 2008). Even in early sensorimotor development, infants—and robots trained with "goal babbling" (Rolf et al., 2010)—explore by first sampling internal goals and only then discovering actions to achieve them (Biro & Leslie, 2007; Baranes & Oudeyer, 2013). Hence, whether navigating an environment, solving a puzzle, or manipulating objects, agents identify and work towards goals that shift over time (Molinaro & Collins, 2023; Chu et al., 2024). This goal-directed behaviour is not cleanly captured by episodic boundaries, discounting heuristics or even average rewards. Rather, it reflects a continual process of selecting and achieving goals in service of reward maximisation, which is especially important for small agents (with bounded memory or compute) acting in big worlds (Javed & Sutton, 2024).

This framing also aligns with discrete planning and continuous control, where tasks are typically defined in terms of reaching goal states (Bertsekas & Tsitsiklis, 1991). Both classic and recent RL ideas also hint at machinery for such flexibility. The *successor representation* decouples dynamics from rewards and can be viewed as a predictive map over arbitrary cumulative objectives (Dayan, 1993; Barreto et al., 2017). This can then be combined with hierarchical RL (HRL) approaches like the *options framework*, which formalises temporally-extended skills, exposing termination and initiation conditions that resemble goal predicates (Sutton et al., 1999; Konidaris & Barto, 2009; Barreto et al., 2019). In the general case, General Value Functions (GVFs) explicitly parameterise value by a goal description (Sutton et al., 2011), which can then be approximated using *universal value function approximators* (UVFAs), enabling generalisation across tasks (Schaul et al., 2015; Ma et al., 2020). To improve sample efficiency, one can then use *hindsight experience replay* (HER) which exploits realised but unintended goals to bootstrap sparse-reward learning (Andrychowicz et al., 2017). More recently, contrastive objectives have been shown to learn goal-conditioned policies directly from raw trajectories (Eysenbach et al., 2022), which has been used to demonstrate that exploration conditioned on a *single* given goal can drive the emergence of rich skill hierarchies (Liu et al., 2024).

However, most RL paradigms lack a formalism that places agent-chosen goals at the centre of learning. This gap stems from a common practice: **we, as system designers, define the specific goal the agent must achieve from each state and evaluate its performance by whether it achieves them**. Even general HRL approaches, such as the option-critic architecture (Bacon et al., 2017), are not necessarily goal-directed unless additional goal rewards or regularisation are given by the system designers (Harb et al., 2018; Luo et al., 2023). This external specification has led to a conceptual divide between goal reachability, as framed in planning, and reward maximisation, as framed in RL. But from the agent's perspective, there is no such distinction. The only objective is to maximise cumulative reward, and any notion of a goal must serve that end. It follows, then, that **goals should not be fixed externally but selected by the agent itself—constructed, refined, and pursued dynamically as tools for acquiring reward**.

We draw these threads together by proposing an agent-centric frame for *General Value Functions* as the foundational knowledge representation for RL. Instead of the goals per step, corresponding goal-conditioned rewards, and eventual termination being environment-given as in prior works, we propose making them agent-selected. Crucially, they can be learned *model-free* from a single stream of continual experience by adding only a goal variable and a done action to the standard RL loop (Figure 1). This minimal change instantiates the "goal-directed prior" suggested by cognitive neuroscience while remaining compatible with the big-world hypothesis. Hence, this reframing of GVFs unifies and generalises existing paradigms while offering greater flexibility and conceptual clarity.

2 Reinforcement Learning as a Goal-Directed Continual Process

Classical RL paradigms have led to major advances (Singh et al., 2022; Hafner et al., 2023; Chang et al., 2024), but they remain ill-suited for the complexities of real-world decision-making. Episodic formulations impose artificial environment boundaries, discounted approaches artificially bias agents toward short-term gains, and average-reward methods often disregard explicit goal-directed behaviour. These paradigms also typically assume that goals are provided externally, defined through environmental features or handcrafted rewards. Yet in practice, intelligent agents operate continually, selecting and refining their own goals in service of long-term reward acquisition.

Formally, the agent-environment interaction is often modelled as a Markov decision process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R \rangle$, where $\mathcal{S}$ is the set of *states*, $\mathcal{A}$ the set of *actions*, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ the *transition function*, and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ the *reward function*.[1] An RL agent can then be defined using the standard general value function formalism (Sutton et al., 2011; Schlegel et al., 2021). A GVF question is a tuple $q = \langle \pi, C, \gamma \rangle$, consisting of a target policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, a cumulant $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and a continuation function $\gamma : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. Its answer is a value function

$$V^q(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \left(\prod_{k=0}^{t-1} \gamma(S_k, A_k, S_{k+1}) \right) C(S_t, A_t, S_{t+1}) \mid S_0 = s \right],$$

with the empty product equal to one. For example, standard discounted RL is a single question $q = \langle \pi^*, R, \gamma \rangle$: the policy π^* is agent-learned to maximise V^q, the cummulant is simply the environment reward, and the continuation $\gamma \in [0, 1)$ is designer-chosen. Similarly, successor representations (SRs) define a bundle of feature-labeled GVF questions $q_i = \langle \pi^*, \phi_i, \gamma \rangle$, where $R(s, a, s') = \sum_i w_i \phi_i(s, a, s')$ is assumed linear in terms of designer-chosen transition features ϕ.

While this formulation underpins many RL methods, it lacks one critical element: an explicit and flexible representation of goals.

But what is the *goal* of an agent? The space of possible goals is broad (Liu et al., 2022; Colas et al., 2022; Davidson & Gureckis, 2024). A goal may be a state, a set of states, a region in a factorised representation, or a distribution over outcomes (Li et al., 2006; Konidaris et al., 2015; 2018; James et al., 2022; Marom & Rosman, 2024). This conceptual ambiguity necessitates a general-purpose knowledge representation capable of supporting goal pursuit under uncertainty, abstraction, and continual change. In general, let $\mathcal{G}$ be the set of goals and let $P_\models : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{G} \to [0, 1]$ be the *goal satisfaction function*, where $P_\models(g|s, a, s')$ represents the probability that the transition $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ satisfies the goal $g \in \mathcal{G}$, denoted as $(s, a, s') \models g$. For example, goals could be states ($\mathcal{G} \in 2^{\mathcal{S}}$) or subsets of states ($\mathcal{G} \subseteq 2^{\mathcal{S}}$), where $(s, a, s') \models g$ if $s = g$ or $s' \in g$ respectively.

Unlike classical frameworks such as stochastic shortest-path (SSP) problems (Bertsekas & Tsitsiklis, 1991), classical planning (Ghallab et al., 2004), or traditional goal-conditioned RL (GCRL) (Schaul et al., 2015), we argue that goals are not fixed entities defined by the environment. Instead, they are internal constructs that are chosen, shaped, and revised by the agent as it learns about the world. This view reframes the RL problem: not as one of maximising episodic rewards over a finite horizon or discounted and average rewards over an infinite horizon, but of continually selecting and pursuing useful goals to acquire reward.

3 Agent-Centric General Value Functions as a Unifying Framework

If we are to view RL as a continual, goal-directed process, we require a formal representation capable of encompassing the inherent flexibility and complexity of goals. However, traditional RL frameworks typically define goals explicitly via the environment, rather than allowing agent-driven goal identification and pursuit. Here, we propose the adoption of agent-centric general value functions (ACGVFs) as the canonical, unifying representation addressing precisely this limitation.

[1] Notice that this standard reward structure may be too general. We believe that general agents require different inductive biases to further constrain it, such as requiring sparse rewards or that total rewards tend to $-\infty$ in the absence of termination.

Let the agent maintain an *internal goal space* $\mathcal{G}$ and a Boolean *done action* $\mathcal{D} = \{0, 1\}$. We can then define a *general policy* $\pi : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \times \mathcal{D} \to [0, 1]$ as the probability that the agent selects in each environment state $s \in \mathcal{S}$: (i) a goal $g \in \mathcal{G}$ to reach, (ii) an environment action $a \in \mathcal{A}$ to take, and (iii) a done action $d \in \mathcal{D}$ to stop ($d = 1$) or continue ($d = 0$) accumulating rewards.[2] When the agent executes a in the environment and transitions to the next state $s' \in \mathcal{S}$, it then receives an *internal reward*

$$\bar{R}(s, g, a, d, s') \coloneqq \begin{cases} R(s, a, s'), & \text{if } d = 1 \implies (s, a, s') \models g, \\ R_{\text{MIN}}, & \text{otherwise,} \end{cases} \tag{1}$$

$$= \left[(1 - d) + dP_{\models}(g|s, a, s')\right] R(s, a, s') + d\left[1 - P_{\models}(g|s, a, s')\right] R_{\text{MIN}} \tag{2}$$

where R_{MIN} is the lower bound reward to penalise any policy that *incorrectly* terminates ($d = 1$) when the goal is not met ($(s, a, s') \not\models g$).[3] These agent-defined goals, termination, and reward function then induces an agent-centric GVF. Precisely, the ACGVF induced by π and $\bar{R}$ is defined as

$$Q^{\pi}(s, g, a, d) \coloneqq \mathbb{E}_{\pi}\left[\bar{R}(S_0, G_0, A_0, D_0, S_1) + (1 - D_0)V^{\pi}(S_1, G_0) \;\middle|\; S_0{=}s, G_0{=}g, A_0{=}a, D_0{=}d\right],$$

where $S_t \in \mathcal{S}$, $G_t \in \mathcal{G}$, $A_t \in \mathcal{A}$, and $D_t \in \mathcal{D}$ are random variables at time-step t induced by executing the policy π in the environment, and

$$V^{\pi}(s, g) \coloneqq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\left(\prod_{k=0}^{t-1}(1 - D_k)\right)\bar{R}(S_t, G_0, A_t, D_t, S_{t+1}) \;\middle|\; S_0{=}s, G_0{=}g\right]$$

is the expected return of executing π. Hence ACGVFs satisfy the usual goal-conditioned Bellman equations. For example, for a finite MDP, we have:

$$Q^{\pi}(s, g, a, d) = \sum_{s' \in S} P(s'|s, a)\left[\bar{R}(s, g, a, d, s') + (1 - d)\sum_{a', d'} \pi(g, a', d'|s')Q^{\pi}(s', g, a', d')\right].$$

The *optimal ACGVF* is then $Q^* \coloneqq Q^{\pi^*}$, where $\pi^*(g, a, d|s) \coloneqq \text{argmax}_{\pi} Q^{\pi}(s, g, a, d)$ is the *optimal general policy*. It is also the unique fixed-point of the usual goal-conditioned Bellman optimality operator:

$$Q^*(s, g, a, d) = \sum_{s'} P(s'|s, a)\left[\bar{R}(s, g, a, d, s') + (1 - d)\max_{a', d'} Q^*(s', g, a', d')\right].$$

Since this GVF is agent-centric, acting optimally in the *world* therefore reduces to the maximisation

$$\pi^*(s) \in \operatorname*{argmax}_{g, a, d} Q^*(s, g, a, d),$$

capturing goal, termination and environment action selection in one expression (analogous to Figure 1). Such agent-centric GVFs are also similar to World Value Functions (WVFs) when termination is environment-defined and the goal-space is defined as terminal states (Nangue Tasse et al., 2022). Interestingly, evidence from prior works leveraging them (Nangue Tasse et al., 2025) shows that for every goal g reachable from s, the policy $\pi^*_g(s) = \text{argmax}_{a,d} Q^*(s, g, a, d)$ reaches g with maximal probability and value $V^*(s, g) = \max_{a,d} Q^*(s, g, a, d)$. In the literature, agents with such a general ability to reach goals are said to possess *mastery* (Veeriah et al., 2018).

[2] The *done action* plays a similar role as the discount factor (which represents continuation probability) or the termination condition in the options framework. It is important for settings where environment rewards are only given during termination, or when total rewards tend to $-\infty$ in the absence of termination.

[3] While Equations 1 and 2 are only equivalent when goal satisfaction is deterministic (when $P_{\models} \in \{0, 1\}$), we treat Equation 2 as the more general definition when goal satisfaction is allowed to be stochastic. The penalty R_{MIN} can also be defined in practice similarly to prior works (Nangue Tasse et al., 2020) or even learned (Nangue Tasse et al., 2026).

Proper versus improper general policies. Analogously to SSP theory (Bertsekas & Tsitsiklis, 1991; Rosenberg et al., 2020), a general policy is *proper* if it terminates (selects $d = 1$) with probability 1 in finite time; otherwise it is *improper*. Any improper policy is then assumed to be unbounded-from-below—for example when all rewards are negative. This setting also mirrors that of classical planning and control, where policies that do not reach goals are invalid. This then guarantees the existence of a deterministic optimal *proper* policy using standard SSP theory and justifies the contraction used by dynamic-programming updates. It also implies that policies can be learned using classic TD-learning or policy-gradient algorithms without any changes (Sutton et al., 1998).

Why optimality still matters in big worlds. While we have so far discussed optimal GVFs, a resource-bounded agent operating in an open-ended world must keep *exploring* both actions and goals; full optimality is rarely attainable (Javed & Sutton, 2024; Kumar et al., 2024). GVFs nevertheless embed an unambiguous target: *mastery*. At every instant, the agent is explicitly trying to reach whichever goal it has selected, and the limiting performance it strives for is one in which every reachable goal can be achieved reliably. This explicit goal-directedness is absent from episodic, discounted or average-reward RL, yet it is central to behaviour observed in animals and humans.

Hence, agent-centric GVFs are more than a conceptual tool: they naturally encompass existing RL paradigms. Episodic RL arises when goals are terminal states. Discounted RL corresponds to implicitly defined horizon-sensitive goals. Average-reward RL emerges when goals reflect steady-state distributions. GVFs also connect naturally to planning and optimal control: stochastic shortest-path problems map directly to GVFs, and finite-horizon control problems become special cases where trajectory termination is defined via goal satisfaction.

4 Relationship to Goal-Conditioned Reinforcement Learning

Traditional GCRL (Schaul et al., 2015; Liu et al., 2022) can be defined as a goal-labeled family of GVF questions $q_g = \langle \pi_g, R_g, \gamma_g \rangle$ where $g \sim \mu(\cdot|s)$ is the goal to achieve and $\mu : \mathcal{S} \times \mathcal{G} \to [0, 1]$ is a goal-selection (question-selection) function. For example, in common sparse-reward GCRL, $R_g(s, a, s') = P_{\models}(g|s, a, s')$ and $\gamma_g(s, a, s') = \gamma\,[1 - P_{\models}(g|s, a, s')]$. GVF questions can also be seen as a generalisation of the options framework, where each option is goal-labeled, μ is the policy over options that also defines the initiation set, π_g is the intra-option policy, and $1 - \gamma$ represents termination. Thus GCRL, UVFAs, options, and many skill-discovery methods differ mainly in how they choose or learn the question components $\mu, \pi_g, R_g, \gamma_g$.

However, these components are still usually environment-defined, hand-designed, or learned from a separate designer-specified objective. GCRL typically learns π_g for goals sampled from a given μ, with R_g and γ_g hand-defined. Options typically learn or use intra-option policies and terminations, but initiation sets, rewards, or discovery objectives are again external or separate. Unsupervised RL and skill discovery can be interpreted as useful mechanisms for approximating good questions, but their notion of "good" is still designer-specified. The missing object is the *agent-centric question*: which goals, rewards, and terminations should be chosen by an agent whose only external grounding is the environment reward? Agent-centric GVFs provide an ideal target: the optimal questions are those induced by the optimal general policy that maximises environment reward while maintaining mastery. Table 1 summarises the differences between the frameworks.

A general policy $\pi(g, a, d|s)$ induces the folowing question components:

$$\mu(g|s) = \sum_{a,d} \pi(g, a, d|s), \quad \pi_g(a|s) = \sum_{d} \pi(g, a, d|s), \quad \gamma_g(s, a) = \sum_{d} \pi(g, a, d|s)(1 - d), \text{ and}$$

$$R_g(s, a, s') = \big[\gamma_g(s, a) + \gamma_{g,\models}(s, a, s')\big] R(s, a, s') + \big[(1 - \gamma_g(s, a)) - \gamma_{g,\models}(s, a, s')\big] R_{\text{MIN}}.$$

where

$$\gamma_{g,\models}(s, a, s') = (1 - \gamma_g(s, a)) P_{\models}(g|s, a, s').$$

Framework	Labels	Goal selector	Policy	Cumulant	Continuation
RL	None	None	π^*	R	γ External
SRs/SFs	Finite features i	None	π^*	ϕ_i External	γ External
GCRL/UVFAs	Arbitrary goals g	μ External	π^*_g	R_g External	γ_g External
WVFs	Terminal states g	μ^*	π^*_g	R^*_g	γ External
Agent-centric GVFs	Arbitrary goals g	μ^*	π^*_g	R^*_g	γ^*_g

Table 1: Different RL frameworks as GVF question families over an MDP. *External* means typically chosen by the environment, or system designer, or a separate designer-specified objective rather than by maximising the environment reward without additional information.

Hence, an optimal general policy π^* induces the optimal GVF questions $q^*_g := \langle \pi^*_g, R^*_g, \gamma^*_g \rangle$ and optimal goal-selection function μ^*. Their unique optimal GVF answers is then the corresponding optimal agent-centric GVF Q^*. This also highlights an interesting relationship between GVF questions and answers: Optimal GVF questions can be obtained from optimal GVF answers via policy improvement, while optimal GVF answers can be obtained from optimal GVF questions via policy evaluation, and both can be obtained from any initialisation via policy iteration (since agent-centric GVFs satisfy the Bellman equations).

Finally, an interesting question is what prevents unreachable or trivial goals. Since $\mu^*(s) \in \operatorname{argmax}_g \max_{a,d} Q^*(s, g, a, d)$, unreachable goals cannot be selected since they cannot obtain high return. This is because terminating without satisfaction incurs R_{MIN}, while never terminating is improper with unbounded-below return. However, trivial goals are not artificially forbidden. As with trivial actions, they are selected only when following them maximises environment rewards (since ACGVFs condition future expected values on the currrent selected goal). Crucially, this does not remove mastery. The goal-conditioned greedy policy $\pi^*_g(s) \in \operatorname{argmax}_{a,d} Q^*(s, g, a, d)$ reaches g with maximal probability, since not reaching it incurs R_{MIN} or unbounded-below returns. The goal-selector μ^* only decides which mastered goal is most valuable to pursue now. This decoupling is essential for building general agents capable of autonomous goal discovery, flexible replanning, and long-horizon reasoning (Nangue Tasse et al., 2024).

5 But What if My Problem is Different?

While we have argued for framing reinforcement learning as an inherently continual goal-directed process, practical considerations or established workflows may lead researchers or practitioners to adopt more specialised or traditional RL paradigms. Fortunately, using agent-centric GVFs as a foundational framework does not require abandoning the convenience or effectiveness of these conventional methods. In this section, we revisit some of the most common formulations and show how each one can be expressed or recovered inside the agent-centric GVF loop (Figure 1).

5.1 Episodic RL

If the problem domain naturally decomposes into clearly defined episodes, such as discrete tasks or bounded interactions, ACGVFs handle this seamlessly. Episodic RL requires that all states satisfying $s \models g$ for some $g \in \mathcal{G}_{terminal}$ are terminal and all policies must eventually reach $\mathcal{G}_{terminal}$ with probability 1 in finite expected time. For example, $\mathcal{G}_{terminal}$ can be defined as all states reachable after a fixed horizon T (assuming each timestep t is encoded in the state s). This ensures that undiscounted ($\gamma = 1$) total rewards are bounded and optimal policies that maximise it exist.

Episodic RL arises as a special case of ACGVFs by simply defining the internal goal space as: (i) the same as the episodic goal space $\mathcal{G} = \mathcal{G}_{terminal}$; or (ii) $\mathcal{G} = \{g_{terminal}\}$ where $g_{terminal}$ is a dummy variable such that $s \models g_{terminal}$ if $s \models g$ for some $g \in \mathcal{G}_{terminal}$. Then choosing $d = 1$ when $s \models g$ and $d = 0$ when $s \not\models g$ for all $g \in \mathcal{G}$, so the episode boundary arises from the agent's action rather than an external reset. The usual episodic return is then recovered by simply maximising over

the ACGVF: $\max_{g,a,d} Q^{\pi}(s, g, a, d)$. Hence, when $\mathcal{G} = \mathcal{G}_{terminal}$, an agent trained on an episodic task such as "unlock-door" can not only solve the given task, but also generalise to new tasks such as "close-door" and "half-open-door" without retraining.

5.2 Stochastic Shortest Path RL

This setting generalises the episodic setting by relaxing the requirement that all policies must reach the absorbing goal space $\mathcal{G}_{terminal}$, and by also relaxing typical ergodicity assumptions, such as the requirement that any state is reachable from any other state. Instead, it only requires that there exists at least one policy that always reaches $\mathcal{G}_{terminal}$, and policies that do not reach $\mathcal{G}_{terminal}$ have value functions that are unbounded from below. It can also be seen as a generalisation of discrete planning and continuous control, where policies that do not terminate are possible but not optimal (such as motion planning for drones and robotic mobile manipulation).

Similarly to the episodic setting, SSPs are also a special case of agent-centric GVFs. In this setting, the designer specifies an absorbing goal set $\mathcal{G}$ in advance. Agent-centric GVFs simply elevate that set to a *first-class action*: the agent chooses $g \in \mathcal{G}$ at every step and when to terminate $b \in B$. Optimal mastery ensures that the classical shortest-path policy for the current reward function can not only be recovered by maximising over the agent-centric GVF, but can also be inferred zero-shot for new reward functions provided they only differ in $\mathcal{G}$ (Nangue Tasse et al., 2022).

5.3 Discounted RL and the Implicit Horizon

In discounted RL, the discount factor implicitly prioritises near-term rewards, often reflecting uncertainty about the future or computational considerations. This is the most popular RL setting to date that has shown great practical success (Mnih et al., 2015; Guo et al., 2025). However, it requires that the environment discount factor be carefully chosen in $[0, 1)$. This choice of γ defines an implicit goal space $\mathcal{G}$, since any discounted MDP can be converted to an SSP (Bertsekas, 1987; Sutton et al., 1998). This ensures that the total discounted rewards are always bounded, and optimal policies that maximise it exist.

Agent-centric GVFs can capture this paradigm by simply defining the internal goal space as the state space (or as a single dummy goal state that is always satisfied), and then choosing $d = 1$ with probability $1 - \gamma$. Alternatively, for popular discounted tasks where $\mathcal{G}$ is already explicitely defined by the system designer and the rewards are non-zero only in $\mathcal{G}$, ACGVFs offer a simple generalisation by simply choosing $d = \mathbb{1}(s \models g)$ for all $g \in \mathcal{G}$—similarly to the episodic and SSP examples. This also highlights an interesting relationship with successor representations (SRs) (Dayan, 1993), which have been shown to encode similar spatial representations as the grid cells found in mammalian brains. When combined with goal-specific rewards, SRs give rise to goal-directed representations similar to hippocampal place cells (Stachenfeld et al., 2017). SRs predict the future discounted *state occupancies* $\mathbb{P}_{\pi}(s, g)$ of a given regular policy $\pi(s, a)$ by simply estimating the goal-conditioned values with goal space $\mathcal{G} = \mathcal{S}$ and binary cummulants $\mathbb{1}[s = g]$. Using the same binary cumulants $\bar{R}(s, g, a, d, s') = \mathbb{1}[s = g]$, optimal ACGVFs instead predict the future discounted state occupancies $\mathbb{P}_{\pi^*}(s, g)$ under the optimal policies for each goal. This has recently been shown to also lead to similar grid-cell spatial representations as regular SRs (Frasco et al., 2025). Similarly, more recent occupancy variants such as successor features appear as special cases obtained by appropriate cumulant choices (Barreto et al., 2017; Borsa et al., 2018).

5.4 World Models and the "Big-World" Regime

Recent progress in learning world models such as Alpha-Go (Fu, 2016) and Dreamer-v3 (Hafner et al., 2023), which are explicit or implicit representations of environment dynamics used for planning and simulation, aligns naturally with ACGVFs. For example, when $\mathcal{G} = \mathcal{S}$ the optimal GVF is isomorphic to the transition kernel and can be extracted for planning (Nangue Tasse et al., 2022). Such an agent can then answer counterfactual queries such as "*Could I reach a shelter before night-*

fall?" by maximising over candidate goals (g_{ore}) and checking if predicted values are close to the lower bound rewards—since an optimal agent would choose to terminate ($d = 1$) if the goal is unreachable.

In general, Richens et al. (2025) recently showed that general agents that are capable of satisfying a variety of given goals in their environment necessarily encode a world model. Because agent-centric GVFs embed environment dynamics within goal-conditioned value functions, they effectively incorporate world models without requiring separate model learning. Agents employing such GVFs thus gain intrinsic world-modeling capabilities, enhancing interpretability, sample efficiency, and planning potential with no additional overhead.

In summary, far from being a niche abstraction, general value functions subsume the most widely used RL formalisms *and* scale naturally to large, partially observed (given appropriate history representations) and arbitrary goal-specified domains. Consequently, framing agents as goal-directed by using GVFs opens a direct path toward general, autonomous and interpretable agents, which is also an important property for AI safety.

6 Conclusion

Reinforcement learning has long been segmented into a patchwork of paradigms—episodic, discounted, average-reward—each optimised for a narrow class of problems and justified by convenience rather than conceptual coherence. In this paper, we have argued that such formulations are not only limiting, but also misaligned with the true nature of intelligent behaviour. In the real world, agents do not act in neatly packaged episodes, nor do they blindly maximise infinite sums of rewards. Instead, they pursue goals that are dynamically selected, hierarchically structured, and continually refined based on experience.

Agent-centric GVFs offer a compelling and principled foundation for RL grounded in this perspective. By coupling every value estimate to an explicit goal variable and an agent–controlled done action, ACGVFs subsume episodic, discounted, average-reward, stochastic-shortest-path, and even classical planning settings as boundary cases of a single, coherent objective: mastery of reachable goals. This reframing is more than theoretical unification. It delivers three practical dividends:

1. **Generality through mastery.** Optimising an ACGVF yields, in one object, the policies required to reach *any* achievable goal. Agents therefore gain zero-shot transfer, rapid curriculum construction, and a principled notion of skill reuse without handcrafted rewards or option structures.
2. **World model integration.** When the goal space equals the state space, the optimal ACGVF is isomorphic to the transition kernel, enabling imagination rollouts, model-based planning, and counterfactual queries "for free". No separate dynamics learner is needed.
3. **Interpretability and safety.** Because goals are first-class, an observer (or verifier) can query the ACGVF to diagnose what the agent *would* do for a specified objective or impose goal-level constraints—tools that are largely inaccessible in conventional value-based RL.

The implications of this shift are significant. Adopting ACGVFs reframes the RL objective—not as a function of arbitrary horizon discounting or average accumulation, but as the efficient acquisition of reward through goal-directed interaction. This aligns reinforcement learning more closely with fields such as planning, optimal control, and neuroscience, all of which share a common language of goals, intentions, and task success. It also opens new avenues for generalization, compositionality, and abstraction—properties that are essential for scalable, robust intelligence.

Crucially, our claim here is not that adopting ACGVFs will make existing algorithms obsolete. Rather, we suggest that ACGVFs provide a deeper foundation upon which those algorithms can be interpreted, unified, and extended. They offer a lens through which both theoretical and empirical insights may be better understood, and a framework that supports the design of truly general-purpose agents. Hence, if our goal is to understand and build generally intelligent systems, then perhaps it is time to place the goal-directed frame embodied by ACGVFs at the centre of our formalism.

References

M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pp. 5048–5058, 2017.

P. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Adrien Baranes and Pierre-Yves Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

A. Barreto, W. Dabney, R. Munos, J. Hunt, T. Schaul, H. van Hasselt, and D. Silver. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pp. 4055–4065, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Shibl Mourad, David Silver, Doina Precup, et al. The option keyboard: Combining skills in reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Dimitri P Bertsekas. *Dynamic Programming: Determinist. and Stochast. Models*. Prentice-Hall, 1987.

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

Szilvia Biro and Alan M Leslie. Infants' perception of goal-directed actions: development through cue-based bootstrapping. *Developmental science*, 10(3):379–398, 2007.

Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado Van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. *arXiv preprint arXiv:1812.07626*, 2018.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Junyi Chu, Joshua B Tenenbaum, and Laura E Schulz. In praise of folly: flexible goals and human cognition. *Trends in Cognitive Sciences*, 28(7):628–642, 2024.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

Guy Davidson and Todd Gureckis. Toward complex and structured goals in reinforcement learning. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

Nathaniel D. Daw and Daphna Shohamy. The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5):593–620, 2008.

P. Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Sergio A. Frasco, Clementine Domine, Geraud Nangue Tasse, and Devon Jarvis. Learning grid cells with world value functions. *The Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2025.

M. C. Fu. Alphago and monte carlo tree search: the simulation optimization perspective. In *Proceedings of the 2016 Winter Simulation Conference*, pp. 659–670. IEEE Press, 2016.

Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated Planning: theory and practice*. Elsevier, 2004.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Steven James, Benjamin Rosman, and George Konidaris. Autonomous learning of object-centric abstractions for high-level planning. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022.

Khurram Javed and Richard S Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. *Advances in neural information processing systems*, 22, 2009.

George Konidaris, Leslie P Kaelbling, and Tomas Lozano-Perez. Symbol acquisition for probabilistic high-level planning. AAAI Press/International Joint Conferences on Artificial Intelligence, 2015.

George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 61:215–289, 2018.

Saurabh Kumar, Hong Jun Jeon, Alex Lewandowski, and Benjamin Van Roy. The need for a big world simulator: A scientific challenge for continual learning. *arXiv preprint arXiv:2408.02930*, 2024.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

Grace Liu, Michael Tang, and Benjamin Eysenbach. A single goal is all you need: Skills and exploration emerge from contrastive rl without rewards, demonstrations, or subgoals. *arXiv preprint arXiv:2408.05804*, 2024.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022.

Ziyan Luo, Yijie Zhang, and Zhaoyue Wang. Does hierarchical reinforcement learning outperform standard reinforcement learning in goal-oriented environments? In *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*, 2023.

Chen Ma, Dylan R Ashley, Junfeng Wen, and Yoshua Bengio. Universal successor features for transfer reinforcement learning. *arXiv preprint arXiv:2001.04025*, 2020.

Ofir Marom and Benjamin Rosman. Transferable dynamics models for efficient object-oriented reinforcement learning. *Artificial Intelligence*, 329, 2024.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Gaia Molinaro and Anne GE Collins. A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 27(12):1150–1164, 2023.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A Boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9497–9507, 2020.

Geraud Nangue Tasse, Benjamin Rosman, and Steven James. World value functions: Knowledge representation for learning and planning. *arXiv preprint arXiv:2206.11940*, 2022.

Geraud Nangue Tasse, Devon Jarvis, Steven James, and Benjamin Rosman. Skill machines: Temporal logic skill composition in reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2024.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. Composition and zero-shot transfer with lattice structures in reinforcement learning. *Journal of Artificial Intelligence Research*, 82: 2325–2388, 2025.

Geraud Nangue Tasse, Tamlin Love, Mark Nemecek, Steven James, and Benjamin Rosman. An unreasonably simple approach to safe reinforcement learning. In *Reinforcement Learning Conference*, 2026. URL https://openreview.net/forum?id=b0dpupOgRw.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents need world models. *arXiv preprint arXiv:2506.01622*, 2025.

Matthias Rolf, Jochen J Steil, and Michael Gienger. Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, 2(3):216–229, 2010.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Matthew Schlegel, Andrew Jacobsen, Zaheer Abbas, Andrew Patterson, Adam White, and Martha White. General value function networks. *Journal of Artificial Intelligence Research*, 70:497–543, 2021.

Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990, 2022.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017.

R. Sutton, A. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

R. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

R. Sutton, J. Modayil, M. Delp, T. Degris, P. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 761–768. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Tadahiro Taniguchi, Emre Ugur, Matej Hoffmann, Lorenzo Jamone, Takayuki Nagai, Benjamin Rosman, Toshihiko Matsuka, Naoto Iwahashi, Erhan Oztop, Justus Piater, et al. Symbol emergence in cognitive developmental systems: a survey. *IEEE transactions on Cognitive and Developmental Systems*, 11(4):494–516, 2018.

V. Veeriah, J. Oh, and S. Singh. Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*, 2018.

C. Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

Supplementary Materials

The following content was not necessarily subject to peer review.

7 Agent-Centric GVF Example

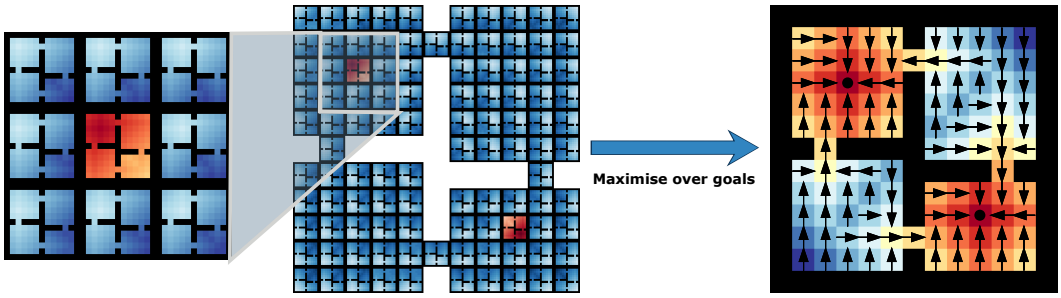

Figure 2: Learned ACGVF (**middle**), Close-up view of the ACGVF for "top-left" goal (**left**), and inferred values and policy for maximising task rewards (**left**). We can observe that the agent not only learns how to achieve all it's internal goals, but also learns which internal goals to pursue from each state to maximise environment rewards. The dark circles represent when the agent decides to terminate ($d = 1$).

Let's consider for example the four-rooms domain (Sutton et al., 1999) where an agent is required to reach various room positions. The agent can move in any of the four cardinal directions at each timestep (with reward -0.1), but colliding with a wall leaves it in the same state. The agent also recieves a reward of 10 when it takes action at the center of the "top-left" or "bottom-right" rooms, and then transitions uniformly at random to any state in the grid. We consider the case where the agent's goals are the entire state space ($\mathcal{G} = \mathcal{S}$).

We begin by illustrating how learned ACGVFs can attain mastery over all goals. Since the ACGVF satisfies the Bellman equations, $Q(s, g, a, d)$ can be learned using any suitable model free RL algorithm such as Q-learning (Watkins, 1989) and many goals update (Veeriah et al., 2018). See Algorithm 1 for the combination of both used in this experiment. The same results are also obtained with regular Q-learning (Algorithm 2).

Figure 2 shows the learned ACGVF, which is generated by plotting the value functions for every goal position and displaying them at their respective (x, y) positions. Note how the values with respect the "top-left" and "bottom-right" goals are high (red), reflecting the high rewards the agent receives for reaching the goals it intended to achieve. We can observe from the value gradient of the plots that the ACGVF does indeed learn how to reach all positions in the gridworld, and that maximising over goals does obtain the regular value function and policy.

Algorithm 1: Goal-Conditioned Q-learning with Many-Goals Update

Initialise: ACGVF $Q(s, g, a, d)$, exploration rate ϵ, learning rate α
Observe initial state $s \in \mathcal{S}$

$$g \leftarrow \begin{cases} argmax_g max_{a,d} Q(s, g, a, d) & \text{w.p. } 1 - \varepsilon \\ \text{random } g \in \mathcal{G} & \text{w.p. } \varepsilon \end{cases}$$

foreach *timestep* **do**

$$a \leftarrow \begin{cases} argmax_a max_d Q(s, g, a, d) & \text{w.p. } 1 - \varepsilon \\ \text{random } a \in \mathcal{A} & \text{w.p. } \varepsilon \end{cases}$$

Execute a, observe reward r and next state s'

$$d \leftarrow \begin{cases} argmax_d Q(s, g, a, d) & \text{w.p. } 1 - \varepsilon \\ \text{random } d \in \mathcal{D} & \text{w.p. } \varepsilon \end{cases}$$

$\gamma \leftarrow 1 - d$

for $g', d' \in \mathcal{G} \times \mathcal{D}$ **do**

$\gamma' \leftarrow 1 - d'$

$\gamma'_{\models} \leftarrow (1 - \gamma') P_{\models}(g' | s, a, s')$

$\bar{r} \leftarrow \left[\gamma' + \gamma'_{\models}\right] r + \left[(1 - \gamma') - \gamma'_{\models}\right] R_{MIN}$

$\delta \leftarrow \left[\bar{r} + \gamma' \max_{a', d''} Q(s', g', a', d'')\right] - Q(s, g', a, d')$

$Q(s, g', a, d') \leftarrow Q(s, g', a, d') + \alpha\delta$

$s \leftarrow s'$

if $True$ ***w.p.*** $1 - \gamma$ **then**

$$g \leftarrow \begin{cases} argmax_g max_{a,d} Q(s, g, a, d) & \text{w.p. } 1 - \varepsilon \\ \text{random } g \in \mathcal{G} & \text{w.p. } \varepsilon \end{cases}$$

Algorithm 2: Q-learning

Initialise: ACGVF $Q(s, g, a, d)$, exploration rate ϵ, learning rate α
Observe initial state $s \in \mathcal{S}$
foreach *timestep* **do**

$$g, a, d \leftarrow \begin{cases} argmax_{g,a,d} Q(s, g, a, d) & \text{w.p. } 1 - \varepsilon \\ \text{random } g, a, d \in \mathcal{G} \times \mathcal{A} \times \mathcal{D} & \text{w.p. } \varepsilon \end{cases}$$

Execute a, observe reward r and next state s'

$\gamma \leftarrow 1 - d$

$\gamma_{\models} \leftarrow (1 - \gamma) P_{\models}(g | s, a, s')$

$\bar{r} \leftarrow \left[\gamma + \gamma_{\models}\right] r + \left[(1 - \gamma) - \gamma_{\models}\right] R_{MIN}$

$\delta \leftarrow \left[\bar{r} + \gamma \max_{a', d'} Q(s', g, a', d')\right] - Q(s, g, a, d)$

$Q(s, g, a, d) \leftarrow Q(s, g, a, d) + \alpha\delta$

$s \leftarrow s'$