Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Survey Track

Adaptive Reward Design in Reinforcement Learning: A Taxonomy and Survey

Raphaela Baybas[1,2] , **Carlo D'Eramo**[2] , **Philipp Brune**[1]
[1]University of Applied Sciences Neu-Ulm
[2] Center for Artificial Intelligence and Data Science, University of Würzburg
{raphaela.baybas, philipp.brune}@hnu.de, carlo.deramo@uni-wuerzburg.de

Abstract

Adaptive Reward Design (ARD) is becoming a fundamental component for Reinforcement Learning (RL) agents, as they are deployed in increasingly complex settings where a single static reward across all phases of learning is rarely sufficient. Yet, ARD is rarely studied as a coherent topic: Relevant ideas are dispersed across reward shaping, curriculum learning, intrinsic motivation, non-stationary objectives, and preference- or feedback-based learning, which obscures conceptual connections and complicates method selection. This survey provides a unified view of ARD in RL by introducing a taxonomy, organized by the primary driver of the reward variation. The taxonomy distinguishes external-feedback-driven reward updates from reward adaptations driven by endogenous within-run signals and those conditioned on exogenous context signals. Using explicit assignment rules, we place work published between 2010 and 2025 within this taxonomy. By synthesizing typical RL settings and domains at the driver level, we simplify the method selection in ARD. Further, we describe the evolution and current trends of ARD and conclude by outlining promising future research directions.

1 Introduction

In the field of *Reinforcement Learning (RL)*, the reward function is an essential component. It determines the direction in which the agent learns and develops. This makes it notable that the reward function was long treated and taken for granted. Fortunately, this changed a few years ago, and the reward function is receiving the attention it deserves in science and research. Nowadays, *Adaptive Reward Design (ARD)* plays an increasingly important role in modern RL setups, and adaptive reward functions are likely to become even more central. As RL is increasingly applied to complex tasks, a static reward often fails to reflect what should matter at different stages of learning - for instance, initially incentivizing simple, discoverable behaviors and later emphasizing more sophisticated and higher-performing solutions. Adaptive reward functions can also support broader applicability by enabling robust generalization across varying contexts and deployment conditions. Moreover, updating the reward during training opens the possibility of a mechanism for mitigating reward misspecification by progressively correcting rewards that would otherwise steer the agent towards unintended behaviors. Finally, by shaping the reward in response to learning dynamics, ARD can improve the balance between exploration and exploitation.

In line with these possibilities, a large number of methodologically different approaches have already been published in the field of ARD, but rarely as a coherent topic: Relevant ideas are dispersed across reward shaping, curriculum learning, intrinsic motivation, non-stationary objectives, and preference- or feedback-based learning. This obscures conceptual connections and complicates method selection, which is why structuring and classification of the ARD research field is needed.

Padakandla [2021] directly covers scenarios of changing reward functions as part of non-stationary environments. However, he focuses on the algorithmic techniques (e.g. meta-learning, drift detection) without giving a taxonomy that would structure the field. Harib et al. [2022] provide a survey of the evolution of adaptive learning methods for nonlinear dynamic systems. The perspective is primarily historical-evolutionary and control-theoretical, but does not offer a taxonomic systematization of ARD. Ibrahim et al. [2024] provide a broader perspective on designing context-dependent rewards. They propose a taxonomy of reward design techniques. Their taxonomy emphasizes the methodologies and approaches that guide the design of rewards.
So far, no work presents a taxonomy that focuses directly on the ARD itself. This survey paper addresses this issue and establishes an initial systematic taxonomy for ARD, providing a uniform structure and formal classification.

Our survey provides a taxonomy that centers ARD itself. A significant advantage of our taxonomy is that even future, completely new approaches can be placed within our taxonomy. Additionally, we discuss the corresponding landscape of the taxonomy. Concretely, we formulate the following research questions:

RQ1 How can the existing work on ARD in RL be systematically structured into a taxonomy that centers the reward design itself, and what does the resulting ARD landscape look like?

RQ2 Along this taxonomy, what typical application scenarios, trends, and future research directions reveal for ARD?

We start with defining ARD and explaining its background in section 2. In section 3 we answer the research question RQ1. To conclude the survey we discuss RQ2 in section 4 and finally give a short conclusion in section 5.

2 Background of Adaptive Reward Design

2.1 Terminology and Fundamental Concepts

Reward design refers to the process of constructing a reward function such that the setting leads the agent to the desired behaviour. It comprises reward engineering and reward shaping [Ibrahim *et al.*, 2024].

The reward design process involves creating a reward function from scratch, often called *reward engineering* [Ibrahim *et al.*, 2024]. The task is to extract the reward function from the optimal behavior within a setting so that the agent relearns the most optimal behavior directly through this reward function [Ng *et al.*, 1999]. In some cases, the term reward engineering is used as a more practical synonym for reward design. Throughout this survey, reward engineering is understood in accordance with the definition provided by Ibrahim et al. [2024].

Additionally, reward design includes a technique known as *reward shaping*. This term describes the adjustment of the existing reward function by adding additional mathematical terms to the reward function, to steer the agent's learning process more strongly towards the learning goal [Ng *et al.*, 1999].
In their pioneering work, Ng et al. [1999] form the theoretical foundation for reward shaping. The central research goal of this work is to investigate the conditions under which a modification of the reward function leaves the optimal policy (π^*) of an *Markov Decision Process (MDP)* unchanged (called policy invariance). In their work, the authors identify potential-based reward shaping as the necessary and sufficient condition for ensuring policy invariance. The modified reward function R' is composed of $R' = R+F$, where R corresponds to the original reward function and F to a so-called shaping function. The shaping function must be expressible as the difference of a potential function Φ over the states:

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s). \tag{1}$$

Potential-based reward shaping systematically prevents the learning of suboptimal behaviors in which agents exploit reward loops without achieving the actual goal. It forms the basis for all further research in the field of reward shaping.

This survey paper takes a holistic understanding of the term reward design, consisting of reward engineering and reward shaping. The scope of this paper's topic is limited to ARD.

In *adaptive reward design*, the resulting reward function is not fixed, but rather dynamic and therefore changes and adapts within the agent's learning process. This definition explicitly includes settings with a deterministic, a priori defined reward function, provided that the reward nonetheless varies during the learning process, according to a predefined mechanism. In this case - having a predefined fixed reward function, the reward for the same state-action pair (s, a) within the agent's learning process must not be the same value, but vary.

Contextual reward design is a particular sub-form of ARD. In *adaptive* reward design, the reward is adapted to any driver. This adaptation to the driver causes the reward to be updated or changed whenever the driver signal changes. In the special case of *contextual* reward design, the driver is defined as a context variable within the given RL task. The contextual variable could be, e.g., a task or goal specification, an environment configuration, any multi-agent context, user preferences, safety settings, or other exogenous factors that are not captured by the agent's state representation.

In this survey, we use ARD as an umbrella term that includes contextual reward design.

2.2 Survey Methodology

To ensure a focused and methodologically coherent survey, this subsection explains our survey methodology.

The focus of this survey is on 'adaptive reward design' and its definition, provided in Subsection 2.1, establishes the scope of the survey.

To obtain a holistic overview of ARD, we examine work published between 2010 and 2025, while earlier publications are considered as fundamental benchmark work. We selected 2010 as our research start point, since deep (reinforcement) learning has been on the rise during that time, best seen in the famous paper 'Playing atari with deep reinforcement learning' from Silver et al. [2013]. To accurately capture the current development of ARD, an explicit focus was placed on recent work from 2025.

When concretely selecting contributions for this survey paper, we apply explicit inclusion and exclusion criteria. A paper is included only if it satisfies all of the following inclusion criteria:

I1 **Reinforcement learning setting:** The paper is situated in an RL/MDP framework.

I2 **Focus on reward design:** Reward design constitutes a substantive part of the contribution. Concretely, the paper may deliberately discuss reward design, or it may propose a method that modifies the reward function or its components in a way that falls under our definition of

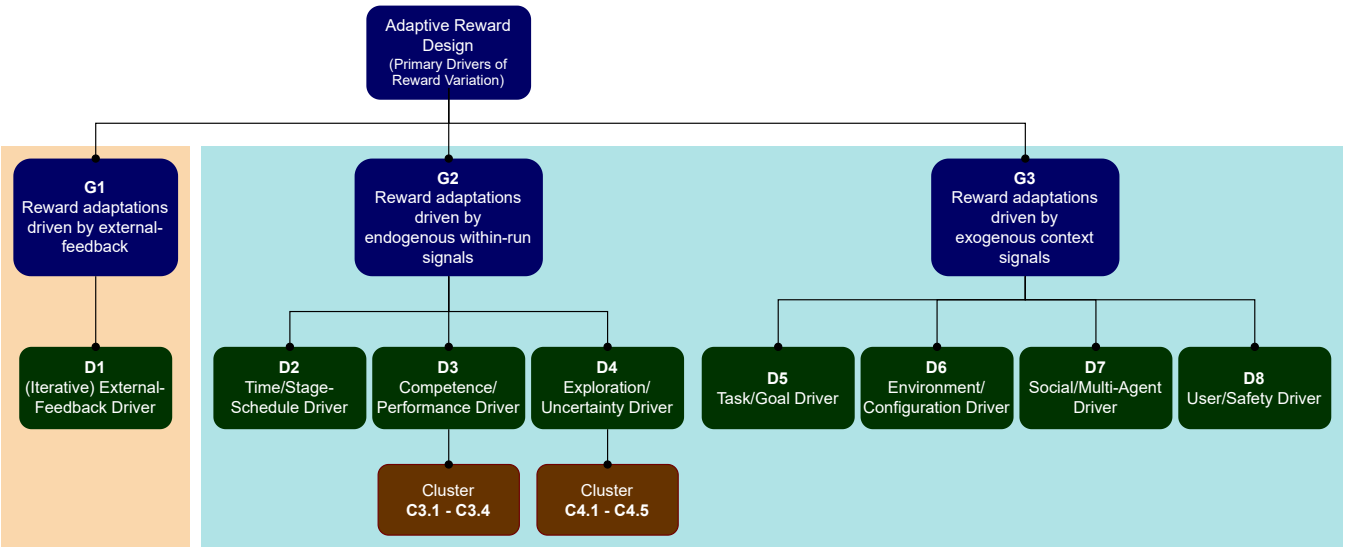

Figure 1: The taxonomy of adaptive reward design differs in the primary drivers that the reward update may be adapted to.

ARD - even if the authors frame the contribution as an RL algorithmic mechanism rather than as reward design.

I3 **Adaptive reward design:** The paper contains an (explicit or implicit) form of adaptation in reward design (i.e., the reward function, shaping terms, or reward-model parameters depend on a context signal or vary during training or execution).

If the following exclusion criterion applies, the paper is not considered further:

E1 **No generalizable contribution:** The paper presents only a single, ad-hoc reward formulation for a specific application, without a generalizable reward design method.

The taxonomy was developed using a synthetic approach. Initially, all relevant literature was collected following the methodology outlined above. Subsequently, a synthetic structural analysis was performed with the objective of establishing a taxonomy, centered on ARD. Various classification schemes were initially evaluated along several axes (e.g., separation by context type, adaptation mechanism, or design approach). Finally, reducing the classification to one central dimension - the primary driver of reward updates - provides the best analytical clarity.

3 Taxonomy and Landscape of Adaptive Reward Design

This section answers the research question RQ1. The motivation for establishing an ARD taxonomy arises from the tendency for ARD ideas to appear as secondary elements within broader RL contributions (e.g., curriculum learning, intrinsic motivation, non-stationary objectives, or preference-based learning), which obscures conceptual connections and complicates method selection. To organize the fragmented literature on ARD and to support consistent classification and method comparison, we introduce a taxonomy that centers on a single, operational question:

> *What is the primary driver that the reward update is adapted to?*

Answering this key question yields eight main drivers (D1 - D8), which can be categorized into three groups (G1 - G3): reward adaptations driven by external feedback, reward adaptations driven by endogenous within-run signals, and reward adaptations driven by exogenous context signals (see Figure 1). Due to fundamental differences, group G1 is separated from the other groups (for further explanations see Subsection 3.1). To support finer-grained synthesis within broad drivers, some of the Drivers are clustered into smaller subgroups (e.g., driver D3 contains clusters C3.1 - C3.4).

To structure the literature, we introduce clear assignment rules for the taxonomy, making the classification process reproducible. Each ARD paper is assigned to exactly one primary driver (D1 - D8) using the following decision questions:

DQ1 Does external feedback drive the reward updates?

DQ2 Is the primary driver of reward variation *endogenous* to the current run/learning process (i.e., generated or measured during the run/learning process itself, such as time/stage, competence/performance/progress, or uncertainty/novelty/visitation), or is it *exogenous* (i.e., specified by the task/environment/social/user setup and not triggered by within-run signals)?

Figure 2 summarizes our step-by-step assignment procedure, enabling a consistent and reproducible mapping of each paper into our taxonomy. Using DQ1, the decision flow first checks whether reward/objective updates are driven by external feedback (G1) or not. Otherwise, it uses DQ2 to distinguish endogenous within-run signals (G2) from exogenous context signals (G3) as the primary driver. Within G2 and G3, each paper is then assigned to a specific driver

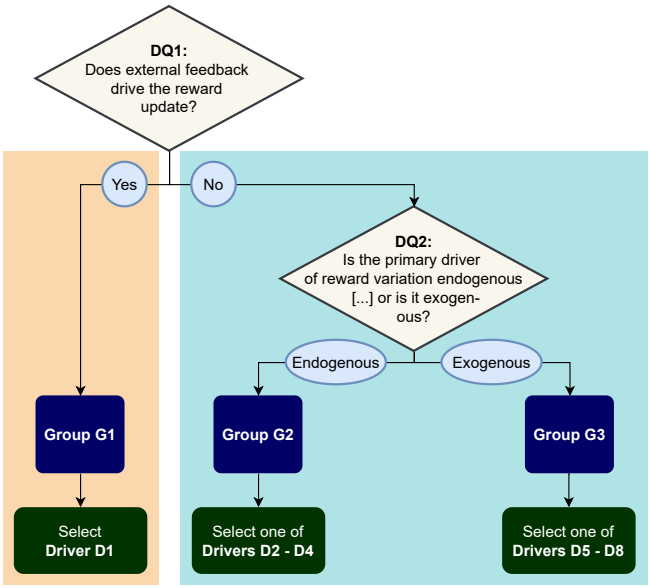

Figure 2: The assignment rules for the taxonomy ensure that each paper can be assigned to one primary driver.

within the group. The required definition of each individual driver can be found in the corresponding driver subsections within Sections 3.2 (G2, D2 - D4) and 3.3 (G3, D5 - D8). When multiple drivers appear plausible, using the following tie-break rules in the given order ensures a consistent primary driver assignment: (i) choose the driver that explicitly parameterizes or triggers the reward update (i.e., appears in the update equation or reward switch criterion); (ii) if multiple triggers are equally central, choose the one that the authors frame as the primary objective/motivation. Note that the first tie-break rule is responsible for the fact that, the name of a paper does not always reflect our primary driver assignment.

A significant advantage of our taxonomy is that even future, completely new approaches can be classified by asking our taxonomy key question.

The following subsections detail how the drivers within each group (G1–G3) can be distinguished and how each driver forms the ARD landscape.

3.1 G1 - Reward Updates Adapted to External-Feedback

The 'group' G1 contains only one driver as an entry: Driver D1. It contains ARD work, where the reward update is adapted to external feedback.

We treat feedback-driven reward adaptation as a top-level decision case. Meaning, as soon as any reward adaptation driven by external feedback is identified within a paper, it is assigned to G1 (and consequently to D1). Independent of any other potential main drivers. We decide to do so, because it relies on a fundamentally different information source than conventional reward. In G1 (and D1), reward updates are driven by exogenous evaluative signals that are not obtained from the agent–environment interaction itself, nor from an environment configuration. This distinction also carries practical implications: D1 methods typically require additional infrastructure, including human-in-the-loop processes or dedicated feedback models, along with data collection, annotation, and feedback pipelines, making them methodologically and operationally distinct from purely algorithmic reward adaptations. Finally, treating external feedback as its own driver group improves taxonomic clarity: if not separated, feedback-based adaptation would appear as a cross-cutting ingredient across many categories, diluting conceptual boundaries. Therefore, isolating D1 via a first gate (as given with DQ1), yields a more discriminative taxonomy and results in more methodologically homogeneous remaining groups, which supports clearer within-group comparisons.

D1 - (Iterative) External-Feedback Driver

As the use of DQ1 described above shows, a paper is assigned to D1 whenever the reward is adapted to external feedback (human- or AI-based). The key property is that the reward specification is updated using feedback that is not derived solely from the environment reward scalar or return. Typical instantiations include critique/corrections as iterative feedback in repeated rounds, as seen in [Xie *et al.*, 2024]. Preferences or demonstrations, given only once a time before the training starts, exemplarily seen in [Brys *et al.*, 2015], are not assigned to this group. In this case, they do not function as feedback. If, e.g., on the other hand, demonstrations are used as ongoing input for a reward model (compare [Ibarz *et al.*, 2018]), the demonstrations serve as feedback. Consequently, the corresponding papers are assigned to G1 and therefore to D1.

3.2 G2 - Reward Updates Adapted to Endogenous Within-Run Signals

Group G2 comprises reward updates that are adapted to a primary driver that is endogenous to the current run or learning process. I.e., this means a signal that is generated (or at least measured) within the run or learning process itself. Those inner-loop signals may be time or training stage, competence or performance indicators (e.g., success rates or returns), or exploration-related quantities (like uncertainty, novelty, or visitation counts). Crucially, the reason why the reward is updated - meaning the reward value for the same state-action pair (s, a) changes - is not an externally specified context, but a signal that arises from the learning process itself. As a result, G2 drivers reflect mechanisms in which reward adaptation is tightly coupled to the agent's internal learning trajectory, allowing the reward to evolve in tandem with progress, exploration, or training phase.

D2 - Time/Stage-Schedule Driver

With driver D2, the reward varies according to a predefined schedule tied to training progress, rather than being triggered by online signals such as performance or uncertainty. Concretely, the reward changes as

a function of time, steps, episodes, or curriculum stages, as proposed in [Miyashita and Sugawara, 2022; Wei, 2025]. A paper is assigned to D2 if the reward schedule would unfold identically regardless of whether the agent learns well or poorly.

D3 - Competence/Performance Driver

The driver D3 covers reward updates that are adapted to performance-related signals such as success rate, episodic return, constraint violations, error or loss measures, or other indicators of learning progress. A paper is assigned to D3 if the reward is updated according to the agent's improvement or deterioration.

This driver contains the largest number of publications. To support finer-grained synthesis within this broad driver, we further cluster the literature of D3. In this case, we define the clusters by (i) the performance/progress signal and (ii) the algorithmic mechanism by which it modulates the reward. Four clusters within D3 are the result.

C3.1: Outcome-based signals via gating or scaling form the biggest cluster within D3. C3.1 represents the branch of performance-based adaptations, where external success metrics (e.g. success rate, as in [Ma *et al.*, 2024; Florensa *et al.*, 2017; Wang *et al.*, 2025] or return, as seen in [Badia *et al.*, 2020; Cho *et al.*, 2025]) are utilized to shift the reward landscape via discrete thresholds (*gating*, exemplarly seen in [Florensa *et al.*, 2017; Klink *et al.*, 2024; Kwon *et al.*, 2025]) or statistical weighting (*scaling*, exemplarly seen in [Cho *et al.*, 2025; Ma *et al.*, 2024; Wang *et al.*, 2025]).

C3.2: Meta-gradient signals via bi-level optimization encompasses approaches that frame reward adaptation within D3 as a formal bi-level optimization problem. An outer loop explicitly optimizes reward parameters such that the inner loop—the policy learning process—achieves maximal performance with respect to an extrinsic objective. The driving signal is a meta-gradient, i.e., the derivative of the extrinsic performance measure with respect to the reward-function parameters. Typical papers that represent C3.2 are [Devidze *et al.*, 2024; Hu *et al.*, 2020; Lu *et al.*, 2025].

C3.3: Model-intrinsic signal via value estimates form the third cluster within D3. The Cluster C3.3 comprises approaches in which reward adaptation is driven not by external performance outcomes, but by the agent's internal representation of its own competence. Signals such as the estimated state-value function or model-internal confidence metrics drive the variation of the reward signal (see [Adamczyk *et al.*, 2025; Griesbach and D'Eramo, 2025; Klink *et al.*, 2020; Klink *et al.*, 2021; Müller *et al.*, 2025]).

C3.4: Policy-relation signal via alignment progress is the D3-cluster that comprises approaches in which reward adaptation is driven by relational competence. The driving signal is a policy-relation measure (e.g., a divergence between policies), and the adaptation mechanism modulates the reward according to alignment progress, i.e., how far the agent has progressed toward matching a behavioral standard such as expert behavior. Exemplary works for this are [Wu *et al.*, 2025; Zhang and Tan, 2023].

D4 - Exploration/Uncertainty Driver

The driver D4 covers reward updates that are adapted to exploration-related signals, such as epistemic uncertainty, novelty, visitation counts, or prediction uncertainty. A paper is assigned to D4 if the reward is modified to explicitly steer exploration — for instance, by providing intrinsic bonuses that decrease as uncertainty is reduced or as states become familiar. In the other way round, we explicitly exclude cases where uncertainty or novelty is used only for exploration in action selection without changing the reward.

A substantial fraction of recent ARD work falls under the primary driver D4 (Exploration/Uncertainty Driver). To carry our taxonomy construction and its practical application forward in a consistent and tractable way, we therefore cluster D4 methods into a small set of recurring mechanism families. Given the strict space constraints of this survey, we only provide a brief overview of these clusters, without conducting a D4 analysis at the cluster level. Instead, the clusters serve as a compact organizational scaffold for summarizing D4 at a high level. Concretely, we distinguish between the following five clusters:

- **C4.1 (count- & density-based novelty):** The reward update is derived from state(-action) visitation frequency via (pseudo-)counts or density models, as it is the case e.g. in [Bellemare *et al.*, 2016; Tang *et al.*, 2017; Castanyer *et al.*, 2023].
- **C4.2 (prediction-error surprise):** The reward update is proportional to the prediction error (surprise) of a learned dynamics model, as exemplarily seen in [Pathak *et al.*, 2017; Burda *et al.*, 2019].
- **C4.3 (epistemic uncertainty via disagreement):** The reward update are driven by an epistemic uncertainty quantified as disagreement across multiple plausible models. Unlike C4.2, which uses single-model prediction error ("How wrong am I?"), C4.3 uses epistemic uncertainty via model disagreement ("How much do hypotheses disagree?"). Examples of literature are [Houthooft *et al.*, 2016; Sekar *et al.*, 2020; Gao *et al.*, 2024].
- **C4.4 (episodic-memory & reachability Novelty):** The Novelty is defined via reachability to recently encountered states, using episodic memory or archive matching so that rewards decrease once states become easily reachable/familiar, as seen in [Savinov *et al.*, 2019; Ecoffet *et al.*, 2021].
- **C4.5 (Structure- & Context-Weighted Novelty):** The reward is computed from explicit structural state met-

rics and/or scaled by exogenous contextual factors to steer exploration in a more thoughtful and less short-sighted manner, as it is the case in [Pandian *et al.*, 2025; Hugessen *et al.*, 2024; Wan *et al.*, 2023].

3.3 G3 - Reward Updates Adapted to Exogenous Context

Group G3 includes reward updates whose primary driver is exogenous to the current run or learning process, i.e., the reward update is driven by the task, the environment, a social setting, or a user configuration, and not triggered by within-run signals. These methods can be viewed as outer-loop drivers: the reward specification for a given context instance is fixed within the run, but the resulting reward values differ across various context instantiations. Concretely, the reward value for the same state-action pair (s, a) changes because an externally chosen context (e.g., task/goal ID, environment parameters, agent roles, or user/safety settings) differs, rather than because the agent's online learning dynamics (time, performance, uncertainty) induce an update.
Note that exogenous context values may vary within a run/learning process (e.g., task sampling), but they remain exogenous, as long as these changes are not driven by an agent's within-run signals (such as stage, performance, or uncertainty).

D5 - Task/Goal Driver

The driver D5 covers reward updates that are adapted to task- or goal-related context, such as target states [Andrychowicz *et al.*, 2017], task/goal instructions [Mahmoudieh *et al.*, 2022], or subtask identifier [Azran *et al.*, 2024]. A paper is assigned to D5 if the reward is conditioned on which task or goal is currently specified — so that identical state-action pairs (s, a) can yield different reward values across different tasks or goals — without the variation being triggered by within-run signals such as time, performance, or uncertainty.

D6 - Environment/Configuration Driver

The driver D6 covers reward updates that are adapted to environment- or configuration-related context, such as domain parameters, physical configurations, layout variables, or other environment specifications (e.g., parameters used in domain randomization), exemplarily seen in [Belogolovsky *et al.*, 2021]. A paper is assigned to D6 if the reward is conditioned on the current environment configuration — so that identical state-action pairs (s, a) can yield different reward values under different environment settings — without the variation being triggered by within-run signals such as time, performance, or uncertainty.
Mapping the literature into our taxonomy suggests that the environment/configuration driver is most often not the primary driver of reward variation, but rather a secondary factor that co-occurs with other drivers, as is the case in [Mahmoudieh *et al.*, 2022; Azran *et al.*, 2024].

D7 - Social/Multi-Agent Driver

The driver D7 covers reward updates that are adapted to social- or multi-agent-related context, such as roles or identities of other agents, team compositions, opponent configurations, or cooperation/competition settings, as exemplarily seen in [Devlin *et al.*, 2014; Jaques *et al.*, 2019; Hughes *et al.*, 2018]. A paper is assigned to D7 if the reward is conditioned on such a social context — so that identical state-action pairs (s, a) can yield different reward values depending on the current social configuration.

D8 - User/Safety Driver

The driver D8 covers reward updates that are adapted to user-dependent or safety-related context, such as user preferences [Barreto *et al.*, 2025], safety settings [Hadfield-Menell *et al.*, 2017], or compliance requirements. A paper is assigned to D8 if the reward is conditioned on such a user/safety configuration — so that identical state-action pairs (s, a) can yield different reward values under different users or different constraint settings.

4 Cross-Cutting Analysis

This section answers the research question RQ2 in the following three subsections.

4.1 Typical Application Scenarios and Driver Selection

The following section outlines typical RL settings or domains in which several of the drivers are frequently applied. This provides an overview of when which driver could be used, thereby simplifying the method selection in ARD.

Driver **D1** (external feedback) is frequently used in the alignment of *Large Language Models (LLM)*, since typical LLM tasks cannot be adequately captured by a rigid mathematical definition of a reward function. Other application scenarios are quite rare.

If complex skills are to be learned by incrementally increasing the difficulty, Driver **D3** (competence/performance) is often used. The reward design responds directly to the agent's learning progress, ensuring that the agent receives an appropriate reward according to the agent's current level of competence.
Approaches assigned to driver **D4** (exploration/uncertainty) cope with RL settings of complex state spaces, such as physical control tasks, similarly to D3. The problem here, analogous to the typical use cases of D3, is the sparse rewarding. Approaches from the driver D4 address this problem by rewarding unknown states or state-action pairs more highly. This maximizes the information gain about the environment and enables the drive 'learning from itself'. Until the actual task-specific goal is achieved. This is why the term "intrinsic reward" is often used here, which enables a focus on exploring new things.
A D4 approach is preferable (i) if the goal is to learn a highly accurate model of the environment dynamics itself, or (ii)

if the environment contains misleading local optima. In contrast, choosing a D3 approach is often preferable when having a controllable task complexity (i.e., via choosable initial states or variable distances to the goal). In this case, target-oriented adjustment of the reward may result in higher sample efficiency. In several approaches, both drivers are used in common.

Furthermore, the driver **D7** (social/multi-agent) is, naturally, used in multi-agent systems. However, D7 is rarely used as the primary driver and is more commonly employed as a secondary driver alongside other reward adaptation mechanisms.

4.2 Evolution of Adaptive Reward Design

Within ARD, several driver-level changes can be identified:

A development can be observed within Group G1 and Driver **D1** (external feedback): There is a trend away from selective human feedback to approaches that automate the feedback process or scale human feedback while using AI models. To keep the effort and costs as low as possible, the efficient use of human feedback plays an ever-present role within these driver approaches.

A new trend can also be observed within Group G2. The importance of purely time-based or predefined stage schedules (**D2**) is declining. Approaches related to Driver **D4** (exploration/uncertainty) have been studied for some time. The focus of new research is currently particularly on **D3** (competence/performance). Nevertheless, Driver D4 remains important, as it continues to be employed in combination with D3. Within driver D3, in particular cluster C3.1 (outcome-based signal via gating or scaling) already includes earlier work. In contrast, the research within the direction of clusters C3.2 (meta-gradient signals via bi-level optimization) and C3.3 (model-intrinsic signal via value estimates) is relatively new. Cluster C3.4 (policy-relation signal via alignment progress) contains relatively new and limited research.

Within the group of drivers that are exogenous to the current run or learning process (G3), the drivers **D5** (task/goal) and **D8** (user/safety) are the ones of newer interest. The driver **D7** (social/multi-agent) is an ever-present area of research, but still tends to serve more as a secondary driver within the field of reward adaptation. Finally, only a few papers are associated with Driver **D6** (environment/configuration). Given the apparent straightforwardness of this driver, this does not necessarily indicate a neglected research area.

In general, more research has been conducted on G2 than on G3. In the overall breakdown, it is particularly striking that drivers D1, D3, and D4 play a major role in ARD research and represent the focus of current methodologies. The scientific evolution of ARD can be described as a process of increasing abstraction: from manual adjustment of reward to autonomous discovery of context-specific reward functions. It can be assumed that the importance of ARD will continue to grow in the future as RL scenarios become increasingly complex.

4.3 Future Research Directions

Based on the two preceding subsections, this section outlines future research directions within ARD.

A first direction is to broaden the applicability of **D1** (external feedback) to further domains beyond LLM. There are significantly more application scenarios in which tasks cannot be adequately captured by a rigid mathematical definition of a reward function (i.e., other complex settings with continuous state spaces, like in control scenarios). A key bottleneck is collecting human reward data, due to its high cost and strong dependence on the specific domain. Future work will therefore continue to focus on improving the efficient use of human feedback to keep effort and costs as low as possible. Furthermore, additional research is needed to better address the issue of subjectivity and inconsistency in feedback.

On the taxonomy side, **D6** (environment/configuration) and **D7** (social/multi-agent) remain underexplored as primary drivers of reward variation. More work is needed to assess the value of this research field.

The driver **D3** (competence/performance) represents a trending research direction, especially the clusters C3.2 (meta-gradient signals via bi-level optimization) and C3.3 (model-intrinsic signal via value estimates). It is certainly beneficial to focus on further improvements within this area. The cluster 3.4 (policy-relation signal via alignment progress) currently represents a research vacuum within this taxonomy and therefore offers a considerable scope for innovation.

5 Conclusion

This survey introduced the core terminology of ARD. We presented a taxonomy that systematically structures the existing work on ARD. The new taxonomy is guided by the question “What is the primary driver that the reward update is adapted to?”, which places the ARD at its center. According to this taxonomy, the landscape of ARD is systematically presented. By extracting typical RL settings and domains at the driver level, this survey facilitates method selection in ARD. Finally, we highlighted the evolution of ARD and concluded by outlining promising future research directions.

References

[Adamczyk *et al.*, 2025] Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V. Kulkarni. Bootstrapped Reward Shaping. *Proceedings of AAAI*, 39(15):15302–15310, 2025.

[Andrychowicz *et al.*, 2017] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight Experience Replay. In *NeurIPS*, volume 30, 2017.

[Azran *et al.*, 2024] Guy Azran, Mohamad H. Danesh, Stefano V. Albrecht, and Sarah Keren. Contextual Preplanning on Reward Machine Abstractions for Enhanced Transfer in Deep Reinforcement Learning. *Proceedings of AAAI*, 38(10):10953–10961, 2024.

[Badia *et al.*, 2020] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the Atari Human Benchmark. In *Proceedings of ICML*, pages 507–517, 2020. ISSN: 2640-3498.

[Barreto *et al.*, 2025] Andre Barreto, Vincent Dumoulin, Yiran Mao, Mark Rowland, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing Individual Human Preferences with Reward Features. 2025.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. In *NeurIPS*, volume 29, 2016.

[Belogolovsky *et al.*, 2021] Stav Belogolovsky, Philip Korsunsky, Shie Mannor, Chen Tessler, and Tom Zahavy. Inverse reinforcement learning in contextual MDPs. *Machine Learning*, 110(9):2295–2334, 2021.

[Brys *et al.*, 2015] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, and Matthew E Taylor. Reinforcement Learning from Demonstration through Shaping. In *Proceedings of IJCAI*, pages 3352–3358, 2015.

[Burda *et al.*, 2019] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. In *ICLR*, 2019.

[Castanyer *et al.*, 2023] Roger Creus Castanyer, Joshua Romoff, and Glen Berseth. Improving Intrinsic Exploration by Creating Stationary Objectives. 2023.

[Cho *et al.*, 2025] Myungsik Cho, Jongeui Park, Jeonghye Kim, and Youngchul Sung. ARS: Adaptive Reward Scaling for Multi-Task Reinforcement Learning. 2025.

[Devidze *et al.*, 2024] Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Informativeness of Reward Functions in Reinforcement Learning. In *Proceedings of AAMAS*, pages 444–452, 2024.

[Devlin *et al.*, 2014] Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *Proceedings of AAMAS*, pages 165–172, 2014.

[Ecoffet *et al.*, 2021] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[Florensa *et al.*, 2017] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse Curriculum Generation for Reinforcement Learning. In *Proceedings of CoRL*, pages 482–495, 2017.

[Gao *et al.*, 2024] Zijian Gao, Kele Xu, Yuanzhao Zhai, Bo Ding, Dawei Feng, Xinjun Mao, and Huaimin Wang. Self-Supervised Exploration via Temporal Inconsistency in Reinforcement Learning. *IEEE TAI*, 5(11):5530–5539, 2024.

[Griesbach and D'Eramo, 2025] Sebastian Griesbach and Carlo D'Eramo. Learning to Explore in Diverse Reward Settings via Temporal-Difference-Error Maximization. *RLJ*, pages 1140–1157, 2025.

[Hadfield-Menell *et al.*, 2017] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse Reward Design. In *NeurIPS*, 30, 2017.

[Harib *et al.*, 2022] Mouhcine Harib, Hicham Chaoui, and Suruz Miah. Evolution of adaptive learning for nonlinear dynamic systems: a systematic survey. *Intelligence & Robotics*, 2(1):37–71, March 2022.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational Information Maximizing Exploration. In *NeurIPS*, volume 29, 2016.

[Hu *et al.*, 2020] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping. In *NeurIPS*, volume 33, pages 15931–15941, 2020.

[Hugessen *et al.*, 2024] Adriana Hugessen, Roger Creus Castanyer, Faisal Mohamed, and Glen Berseth. Surprise-Adaptive Intrinsic Motivation for Unsupervised Reinforcement Learning. 2024.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NeurIPS*, volume 31, 2018.

[Ibarz *et al.*, 2018] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in Atari. In *NeurIPS*, volume 31, 2018.

[Ibrahim *et al.*, 2024] Sinan Ibrahim, Mostafa Mostafa, Ali Jnadi, Hadi Salloum, and Pavel Osinenko. Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications. *IEEE Access*, 12:175473–175500, 2024.

[Jaques *et al.*, 2019] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of ICML*, pages 3040–3049, 2019. ISSN: 2640-3498.

[Klink *et al.*, 2020] Pascal Klink, Carlo D' Eramo, Jan R Peters, and Joni Pajarinen. Self-Paced Deep Reinforcement Learning. In *NeurIPS*, volume 33, pages 9216–9227, 2020.

[Klink *et al.*, 2021] Pascal Klink, Hany Abdulsamad, Boris Belousov, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. A Probabilistic Interpretation of Self-Paced Learning with Applications to Reinforcement Learning. *Journal of Machine Learning Research*, 22(182):1–52, 2021.

[Klink *et al.*, 2024] Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. On the Benefit of Optimal Transport for Curriculum Reinforcement Learning. *IEEE TPAMI*, 46(11):7191–7204, 2024.

[Kwon *et al.*, 2025] Minjae Kwon, Ingy ElSayed-Aly, and Lu Feng. Adaptive Reward Design for Reinforcement Learning. 2025.

[Lu *et al.*, 2025] Renzhi Lu, Zonghe Shao, Yuemin Ding, Ruijuan Chen, Dongrui Wu, Housheng Su, Tao Yang, Fumin Zhang, Jun Wang, Yang Shi, Zhong-Ping Jiang, Han Ding, and Hai-Tao Zhang. Discovery of the reward function for embodied reinforcement learning agents. *Nature Communications*, 16(1):11064, 2025. Publisher: Nature Publishing Group.

[Ma *et al.*, 2024] Haozhe Ma, Zhengding Luo, Thanh Vinh Vo, Kuankuan Sima, and Tze-Yun Leong. Highly Efficient Self-Adaptive Reward Shaping for Reinforcement Learning. In *ICLR*, 2024.

[Mahmoudieh *et al.*, 2022] Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-Shot Reward Specification via Grounded Natural Language. In *Proceedings of ICML*, pages 14743–14752, 2022. ISSN: 2640-3498.

[Miyashita and Sugawara, 2022] Yuki Miyashita and Toshiharu Sugawara. Two-stage reward allocation with decay for multi-agent coordinated behavior for sequential cooperative task by using deep reinforcement learning. *Autonomous Intelligent Systems*, 2(1):10, 2022.

[Müller *et al.*, 2025] Henrik Müller, Lukas Berg, and Daniel Kudenko. Using incomplete and incorrect plans to shape reinforcement learning in long-sequence sparse-reward tasks. *Neural Computing and Applications*, 37(23):18851–18866, 2025.

[Ng *et al.*, 1999] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *ICML*, pages 278–287, 1999.

[Padakandla, 2021] Sindhu Padakandla. A Survey of Reinforcement Learning Algorithms for Dynamically Varying Environments. *ACM Computing Surveys (CSUR)*, July 2021. Publisher: ACMPUB27New York, NY, USA.

[Pandian *et al.*, 2025] J. Arun Pandian, Ramkumar Thirunavukarasu, and Rajganesh Nagarajan. Enhanced exploration in reinforcement learning using graph neural network based intrinsic reward mechanism. *Scientific Reports*, 15(1):39986, 2025. Publisher: Nature Publishing Group.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of ICML*, pages 2778–2787, 2017. ISSN: 2640-3498.

[Savinov *et al.*, 2019] Nikolay Savinov, Anton Raichuk, Raphael Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic Curiosity through Reachability. In *ICLR*, 2019.

[Sekar *et al.*, 2020] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *Proceedings of ICML*, pages 8583–8592, 2020.

[Silver *et al.*, 2013] David Silver, Alex Graves Ioannis Antonoglou Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. arXiv:1312.5602v1 [cs].

[Tang *et al.*, 2017] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning. In *NeurIPS*, volume 30, 2017.

[Wan *et al.*, 2023] Shanchuan Wan, Yujin Tang, Yingtao Tian, and Tomoyuki Kaneko. DEIR: Efficient and Robust Exploration through Discriminative-Model-Based Episodic Intrinsic Rewards. In *Proceedings of IJCAI*, pages 4289–4298, 2023.

[Wang *et al.*, 2025] Linji Wang, Tong Xu, Yuanjie Lu, and Xuesu Xiao. Reward Training Wheels: Adaptive Auxiliary Rewards for Robotics Reinforcement Learning. In *2025 IEEE/RSJ IROS*, pages 15262–15267, 2025.

[Wei, 2025] Xinghai Wei. ReSCOM: Reward-Shaped Curriculum for Efficient Multi-Agent Communication Learning. 2025.

[Wu *et al.*, 2025] Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. AlphaDPO: Adaptive Reward Margin for Direct Preference Optimization. 2025.

[Xie *et al.*, 2024] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. In *ICLR*, 2024.

[Zhang and Tan, 2023] Zhe Zhang and Xiaoyang Tan. Adaptive Reward Shifting Based on Behavior Proximity for Offline Reinforcement Learning. In *Proceedings of IJCAI*, pages 4620–4628, 2023.