Towards a Unifying Framework for Formal Theories of Novelty

T. E. Boult[1], P. A. Grabowicz[5], D. S. Prijatelj[2], R. Stern[6], L. Holder[4], J. Alspector[3], M. Jafarzadeh[1], T. Ahmad[1], A. R. Dhamija[1], C. Li[1], S. Cruz[1], A. Shrivastava[7], C. Vondrick[8], W. J. Scheirer[2]

[1] U. Col. Col. Springs, [2] U. Notre Dame, [3] IDA/ITSD, [4] Wash. State U., [5] U. Mass., [6] PARC, BGU, [7] U. Maryland, [8] Columbia U.
{tboult | mjafarzadeh | tahmad | adhamija}@vast.uccs.edu, grabowicz@cs.umass.edu, derek.prijatelj@nd.edu, rstern@parc.com, jalspect@ida.org, holder@wsu.edu, abhinav@cs.umd.edu, vondrick@cs.columbia.edu, walter.scheirer@nd.edu

Abstract

Managing inputs that are novel, unknown, or out-of-distribution is critical as an agent moves from the lab to the open world. Novelty-related problems include being tolerant to novel perturbations of the normal input, detecting when the input includes novel items, and adapting to novel inputs. While significant research has been undertaken in these areas, a noticeable gap exists in the lack of a formalized definition of novelty that transcends problem domains. As a team of researchers spanning multiple research groups and different domains, we have seen, first hand, the difficulties that arise from ill-specified novelty problems, as well as inconsistent definitions and terminology. Therefore, we present the first unified framework for formal theories of novelty and use the framework to formally define a family of novelty types. Our framework can be applied across a wide range of domains, from symbolic AI to reinforcement learning, and beyond to open world image recognition. Thus, it can be used to help kick-start new research efforts and accelerate ongoing work on these important novelty-related problems.

Introduction

"What is novel?" is an important AI research question that informs the design of agents tolerant to novel inputs. Is a noticeable change in the world that does not impact an agent's task performance a novelty? How about a change that impacts performance but is not directly perceptible? If the world has not changed but the agent senses a random error that produces an input that leads to an unexpected state, is that novel?

With decades of work and thousands of papers covering novelty detection and related research in anomaly detection, out-of-distribution detection, open set recognition, and open world recognition, one would think that a consistent unified definition of novelty would have been developed. Unfortunately, that is not the case. Instead, we find a plethora of variations on this theme, as well as *ad hoc* use and inconsistent reuse of terminology, all of which injects confusion as researchers discuss these topics.

This paper introduces a unifying formal framework of novelty. The framework seeks to formalize what it means for an input to be a novelty in the context of agents in artificial intelligence or in other learning-based systems. Using the proposed framework, we formally define multiple types of novelty an agent can encounter. The goal of these definitions is to be broad enough to encompass and unify the full range of novelty models that have been proposed in the literature (Pimentel et al. 2014; Markou and Singh 2003a,b; Scheirer et al. 2013; Bendale and Boult 2015; Langley 2020). An important generalization beyond prior work is that we consider novelty in the world, observed space, and agent space (see Fig. 1), with dissimilarity and regret operators critical to our definitions. The overarching goal is a framework such that researchers have clear definitions for the development of agents that must handle novelty, including support for agents / algorithms that incrementally learn from novel inputs. A longer version of this theory with example applications to three different domains can be found at (Boult et al. 2020).

Our framework supports *implicit theories of novelty*, meaning the definitions use functions to implicitly specify if something is novel. The framework does not require a way to generate novelties, but rather it provides functions that can be used to evaluate if a given input is novel. This is similar to how any 2D shape can be implicitly defined by a function $f(x, y) = 0$, whether or not there is a procedure for generating the shape. We contend any constructive or generative theory of novelty (Langley 2020) must be incomplete because the construction or generation of defined worlds, states, and any enumerable set of transformations between them form, by definition, a closed world. We note, however, that a constructive model can be consistent with our definition, but we do not require a constructive model.

Formalizing Novelty

We present frameworks for formalizing novelty for static or learning-based agents, operating in a setting where handling unknown items is required. Fig. 1 shows the main elements of a novelty problem for task $\mathcal{T}$. The formulation can support a wide range of novelty problems including being robust to novelties, detecting novelties, learning from novelties or generating novelties. The paper's formalization is about theories, rather than "a theory," because when the definition's set of items and associated functions are provided, a different theory of novelty is defined. There are infinitely many such theories of novelty for any given task.

For simplicity of presentation, our world and observation space are d' and d dimensional spaces of real numbers. Let

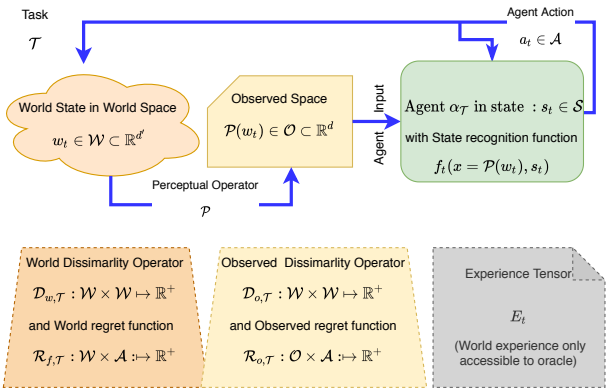

Figure 1: Main elements of the implicit theories of novelty. The agent can only access world information indirectly through a perceptual operator $\mathcal{P}$. It can then update its internal state and act on the world state. Items with dashed outlines are outside of the task or agent but are critical to defining novelty. In the framework a theory of novelty is obtained by specifying: world $\mathcal{W}$ and the world dimensionality d', observation space $\mathcal{O}$ accessible to the agent and its dimensionality d, agent state space $\mathcal{S}$, a perceptual operator $\mathcal{P}$ that processes world regions, task-dependent world dissimilarity functions $\mathcal{D}_{w,\mathcal{T}}$ with associated threshold δ_w, task-dependent observation-space dissimilarity functions $\mathcal{D}_{o,\mathcal{T}}$ with threshold δ_o, agent α in state $s_t \in \mathcal{S}$ at time t, using state recognition function $f_t(x, s)$ to determine the action $a_t \in \mathcal{A}$ to be taken, world regret function $\mathcal{R}_{w,\mathcal{T}}$, observation-space regret function $\mathcal{R}_{o,\mathcal{T}}$, and agent-space regret function $\mathcal{R}_{a,\mathcal{T}}$. Every set of these operators / functions / values defines a different theory of novelty for its associated task.

a world state be $w_t \in \mathcal{W}$, and let $\mathcal{W} \subset \mathbb{R}^{d'}, d' \geq d$, be the representation of the world state at time t obtained from a subset of $\mathcal{W}$, the allowed world states. Note that $\mathcal{W}$ need not be the entire world. It can be only that region of space and time which is sampled during operation or that is relevant for the task. It could even be a finite set of items or relationships to be considered by the system.

Let our observation at time t be $x_t \in \mathcal{O}$ in observation-space $\mathcal{O} \subset \mathbb{R}^d$. Let $\mathcal{P} : \mathbb{R}^{d'} \mapsto \mathbb{R}^d$ be the perceptual operator at time t, which maps world spaces to observation spaces, *i.e.*, $x_t = \mathcal{P}(w_t)$. The agent never has direct access to the world and can only access it via the perceptual operator. This operator is generally a combination of real-time sensing plus external processing on that sensed data. It can also include pre-processing on stored sensory data. But everything in this operator is to be considered external to the agent. Potential changes to system hardware can influence the outcomes of the perceptual operator and are represented as parameters of $\mathcal{P}$ stored within w_t. If the perceptual operator processes only a subregion of any world state, we let $\mathcal{W}_t$ be the subregion of the world that has been processed up to time t. Accordingly, any world states that differ only outside the processed subregions are indistinguishable.

Let agent $\alpha_\mathcal{T}$ solving task $\mathcal{T}$ at time t have an internal state representation $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the space of possible states. An agent reacts to the environment, but a common architecture in the design of autonomous agents is for the agent to have an internal state s_t that influences how the agent will act at time t. To capture this common agent architecture, we define a *state recognition function* that maps an observation-space input, x_t, plus the current agent state, s_t, into its new internal state and action. Formally, let $f_t(x_t, s_t) : \mathbb{R}^d \times \mathcal{S} \mapsto \mathcal{S} \times \mathcal{A}$ be a state recognition function at time t mapping an observation-space input x_t and its internal state s_{t-1} to its new state s_t and an action $a_t \in \mathcal{A}$ to be taken, where $\mathcal{A}$ is the space of actions.

Based on the recognition function output, the agent takes an action. Then, the state transition function, $T(w_t, a_t) : \mathcal{W} \times \mathcal{A} \mapsto \mathcal{W}$, maps the current world state and agent's selected action at time t to a new world state. This mapping can be stochastic, e.g., a Markov decision process. Note that for many real problems the world may change outside of any assumptions, including oracle's assumptions, in which case we cannot assume $w_{t+1} = T(w_t, a_t)$. For problems which traditionally do not have state transition function or specify an action, such as machine learning or open set classification, agent's "action" is the predicted outcome for a given sample, e.g., a class reported by the classifier.

Let $\mathcal{N} \subseteq \mathcal{S}$ be the possible empty set of internal states which are associated with the agent determining the world is novel. The world state w_t is identified by the agent as novel if $s_t \in \mathcal{N}$. When dealing with novelty, the agent may obtain unexpected observation states, which could map to potentially unexpected internal states (*e.g.*, a "crash" state).

To represent history, let $E_t = \{(w_1) \dots (w_t)\}$ be our experience tensor of states. It is important to note that the experience tensor is not about what the agent "remembers". It is external to the agent and is about what world states the agent has experienced up to and including time t. The experience tensor is integral to defining novelty, which depends on dissimilarity between the potentially novel world and the experience of some non-novel worlds.

Dissimilarity and Regret Measures

For a given task $\mathcal{T}$, data generally do not need to match exactly to be considered "the same" with respect to the task's objectives. Note that while we call it "a task," it could include a set of objectives (*e.g.*, a multi-task problem). In any task, multiple dimensions can impact the task performance and determine when a state is effectively the same or how different a state is from prior experience. We formulate this in terms of a measure of dissimilarity, which depends on the task. It is important to note that one should be careful in *a priori* definitions of what matters to a task, as novel world states may have an unpredictable impact on it.

Essential to our framework is a set of task-dependent dissimilarity functions $\mathcal{D}_{w,\mathcal{T}} : \mathcal{W} \times \mathcal{W} \mapsto \mathbb{R}^+$ and $\mathcal{D}_{o,\mathcal{T}} : \mathcal{W} \times \mathcal{W} \mapsto \mathbb{R}^+$, which measure dissimilarity between states in the world space and the observation space respectively. The perceptual dissimilarity may access the world but must map to observed space with the perceptual operator before mapping to $\mathbb{R}^+$. Both dissimilarities can use experience tensor E_t as an optional parameter. Dissimilarity measures produce

non-negative values that are a generalization of distance metrics. It is possible to have zero dissimilarity for states that are identical in terms of the task and hidden variables, even if they are distinct states in the world or observation space. Novelty definitions will generally include a similarity threshold of δ_w and δ_o, beyond which a state is treated as novel, which depends on the task and user requirements. Dissimilarity may be an actual distance metric, a statistical measure, an information-theoretic measure, or such measures combined with ontological or hierarchical information. If the world modeling is probabilistic, then the dissimilarity function may, but need not, be based on probabilistic computations. We use dissimilarity rather than distance since it is well known that human perception / recognition is non-metric (Tversky 1977; Scheirer et al. 2014).

Not all novel states are of interest or present a risk to the agent. While we cannot know the risk of an unknown state until it becomes known, we can, after the fact, assign a regret score associated with new world / observed / agent state after the action $a_t(w_t, s_t)$. We let $\mathcal{R}_{f,\mathcal{T}} : (\mathcal{O} \times \mathcal{A}) \mapsto \mathbb{R}$, $\mathcal{R}_{o,\mathcal{T}} : (\mathcal{O} \times \mathcal{A}) \mapsto \mathbb{R}$ and $\mathcal{R}_{w,\mathcal{T}} : (\mathcal{W} \times \mathcal{A}) \mapsto \mathbb{R}$ be the regret operators for task $\mathcal{T}$ in agent, observation, and world space respectively. Agent and observation regrets are functions of an observed state, whereas world regret of a world state. A suboptimal agent may have an agent regret higher than an observation regret, which should be defined with respect to an oracle or the best agent on observed data.

We separate these three because it allows one to reason about regret in terms of specific models. Only an oracle that has access to ground-truth data has the ability to actually compute regret in world space. However, an agent can approximate regret in observation space, especially given the ground-truth answers. It is important to note that agent-computed regret can only be an approximation, even given the ground-truth action / outcome, because the optimal decision with limited data can still lead to bad outcomes. Hence, an agent might estimate regret even if there should be none.

Example: CartPole Domain

As an example of using the constructs defined above, consider the CartPole in the OpenAI Gym.[1] In this domain, a cart has a pole connected to it, and the task $\mathcal{T}$ is to push the cart left or right so as to prevent the attached pole from falling. The world state w in the CartPole domain comprises the following real values (with default / initial values in parentheses): gravity G (9.8), mass of cart M_c (1.0), mass of pole per unit length M_p (0.1), length of pole L (1.0), force of push F_p (10.0), horizontal force acting on the cart F_h (0), min / max cart position $z^{\min}(-2.4), z^{\max}(+2.4)$, min / max pole angle $\phi^{\min}$ $(-12°)$, $\phi^{\max}$ $(+12°)$, time between state updates τ $(0.02$ seconds), start time t (0). The initial cart position z_0, cart velocity $\dot{x}_0$, pole angle ϕ_0, and pole angular velocity $\dot{\phi}_0$ are all i.i.d. random samples from $[-0.05...0.05]$. The perceptual operator $\mathcal{P}$ in this domain is a projection of the world state that returns only the cart position z, cart velocity $\dot{z}$, pole angle ϕ, and pole angular velocity $\dot{\phi}$ as the 4D observed state vector $x = (z, \dot{z}, \phi, \dot{\phi})$.

Based on these world state features, the task is more precisely defined as: given x, select an action from the space of $\mathcal{A} = \{\text{Left}, \text{Right}\}$ to maintain the cart position within the min / max cart position and maintain the pole angle within the min / max pole angle. Note that the last four features (cart position, cart velocity, pole angle, pole angular velocity) are determined via a deterministic physics model based on the full world state combined with agent actions.

Dissimilarity and Regret in CartPole

State transitions in the CartPole domain are determined by the equations of motion and can be simulated in discrete time with numerical integration. In this example we assume transitions between observed states are Markovian, which simplifies presentation; other theories for this domain could consider dissimilarity and regret in more general settings.

The dissimilarity measure for CartPole might take a simple form (*e.g.*, the Euclidean distance in the world or observed space). However, Euclidean distance between world states is affected by factors other than novelties, including the choice of units. It is also insensitive to the variation in the impact of different variables on the state evolution or task outcome. Proper conditioning would reduce dependency on units and account for states that correspond to different samples from the same world (*e.g.*, the same CartPole world with a different initial position of the pole is not considered novel).

To avoid these issues, we compare two worlds, w and $\check{w}$, the states that proceed from a common observed state and action. We consider an action of an optimal agent, a^*, in the first world, w, and choose as the common observed states the states that the agent encounters, $\check{x}_t$, in the second world, $\check{w}$. Then, we average over all these states, including the initial observed state:

$$\mathcal{D}_{o,\mathcal{T}}(w, \check{w}) = \mathbb{E}_{\check{x}_0,t}(\mathcal{P}(T(M(w, \check{x}_t), a_t^*)) - \mathcal{P}(T(M(\check{w}, \check{x}_t), a_t^*))^2,$$

where $M(w, x) : \mathcal{W} \mapsto \mathcal{W}$ is a function that returns a modified w whose observed components are replaced with the values from x, such that $\mathcal{P}(M(w, x)) = x$, while all other components remain unchanged. Overall, the dissimilarity measures the average distance between observed states in two different worlds that proceed from a common observed state and action, $\check{x}_t$ and a_t^*. The agent is optimal in the first world, while the trajectory is from the second world, so this dissimilarity measure can be seen as an expected state prediction error of the optimal agent trained in the first world and tested in the second world. The world dissimilarity is defined analogously, except it does not use the perceptual operator. Note that these dissimilarity measures depend on multiple experience tensors of the respective optimal agents, i.e., the average is over their trajectories.

This is an asymmetric dissimilarity measure as the selected agent is optimal for the first world and need not be optimal for the second, and then marginalizes over the initial conditions and time in the second world. Due to the conditioning, any pair of states from the same world will have zero dissimilarity.

[1] https://github.com/openai/gym/blob/master/gym/envs/classic_control/cartpole.py

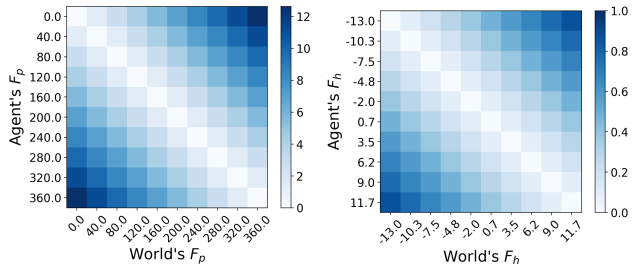

Figure 2: Observation dissimilarity, $\mathcal{D}_{o,\mathcal{T}}(w, \breve{w})$, between a world expected by an optimal non-adaptive agent and the observed world. The agents are optimal in a world having incorrect value of the magnitude of pushing force, F_p (left panel), or a horizontal force acting on the cart, F_h (right panel). The expectation is computed over 20 samples of the initial world state w_0.

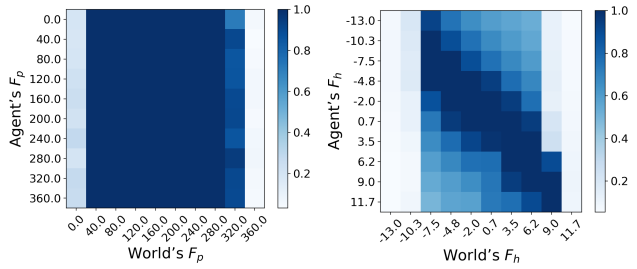

Figure 3: Average reward, 1-$\mathbb{E}_{x_0,t}\ell_{\mathcal{T}}(x_t, a_t^*)$, of non-adaptive agents that are trained to act optimally in a certain world but are tested in another world.

Furthermore, since these depend on the choice of "optimal action" a_t^* from the first world, it implicitly normalizes for how different dimensions (variables) impact the evolution of the world state. One can consider actions of an optimal reference agent under given conditions (*e.g.*, a non-adaptive agent that performs optimally w.r.t. the main task $\mathcal{T}$ in a non-novel CartPole world. If it is infeasible to obtain an optimal agent in practice, then one may use an arbitrary reference agent and its expectation, with the caveat the dissimilarity measure will depend on the oracle's reference agent.

Regret for the agent's action at state x_t w.r.t. $\mathcal{T}$ is

$$\mathcal{R}_{o,\mathcal{T}}(x_t, a_t) = \ell_{\mathcal{T}}(x_t, a_t) - \ell_{\mathcal{T}}(x_t, a_t^*),$$

where $\ell_{\mathcal{T}}(x_t, a_t)$ is the loss incurred in the state x_t by an agent that performs action a_t, while $\ell_{\mathcal{T}}(x_t, a_t^*)$ is the loss of an agent that performs the optimal action a_t^*, given the same observation. In CartPole, the loss at the given time step is 1 if the pole angle or cart's position in the next time steps is beyond the threshold, $|\phi| > \phi^{\max}$ or $|z| > z^{\max}$, otherwise the loss is 0. In this CartPole domain, the world regret is the same as observation regret, $\mathcal{R}_{w,\mathcal{T}}(w_t, a_t) = \mathcal{R}_{o,\mathcal{T}}(\mathcal{P}(w_t), a_t)$, because there are no hidden dynamic elements interacting with pole or cart, such as an invisible pendulum that hangs above the pole and sometimes hits it. Once such elements are introduced, the two regrets may differ.

Agent State and Optimal Non-Adaptive Agents

In simple environments, like CartPole, it is possible to obtain optimal or near-optimal non-adaptive agents. A particularly simple version of CartPole is the one where the agent's action space is binary, *i.e.*, the agent can choose to push the cart left or right, and its reward is the time that the pole is up. In such a CartPole environment, an agent can find optimal actions by performing what-if simulations of the world and searching for actions that result in the best performance, *i.e.*, the agent can simulate what would happen if it pushed the cart left or right and then choose the next action that results in better performance. In this paper, for simplicity, we present the results for a near-optimal single-look-ahead agent. The action is chosen based on the distance, $||\beta^\intercal(x_t - x^{\mathrm{s}})||$, between the expected state resulting from the action, x_t, and the desired state, $x^{\mathrm{s}} = (0, 0, 0, 0)$, where the weight vector $\beta = (0, 0, 1, 0.005)$ weighs discrepancy in ϕ the most and ignores the discrepancies in z and $\dot{z}$. The state of this agent is described by the parameters used to simulate system dynamics and the weight vector, $s = (G, M_c, M_p, L, F_p, F_h, \tau, \beta)$. In non-adaptive agents, these parameters are fixed to the values that correspond to the non-novel world.

Measurements and Observations

As expected, the observed dissimilarity captures the novelties in the magnitude of pushing force and in a horizontal force acting on the cart (Figure 2). The larger the distance between the value of a given parameter in the world and its value assumed by the agent, the larger the dissimilarity in state prediction.

Surprisingly, in CartPole, reasonable changes to many of the aforementioned latent parameters, like the gravity or the magnitude of pushing force, do not impact an optimal non-adaptive agent's performance (left Figure 3). The columns on the verge of the heat map correspond to the achievable limits in performance, *i.e.*, the leftmost column corresponds to such a small force magnitude that the push is insufficient to counter the gravity, whereas the rightmost column corresponds to such a large force magnitude that the push instantly rotates the pole beyond the allowed region. We conclude that the optimal agent performs just as well in the novel world where gravity or pole length have new values, despite making simulations that assume incorrect values of gravity or pole length, *i.e.*, the values from the non-novel world. Note that this agent is non-adaptive, so it does not change its internal model to adapt to the novel world, *i.e.*, internally it uses the model of the non-novel world to take actions. Again, the columns at the verge of the heat map mark inherent performance limits, *i.e.*, if the magnitude of the horizontal force is larger than the pushing force, $|F_h| > F_p = 10$, then the pushes are insufficient to counteract the horizontal force and the pole is destined to fall, despite taking optimal actions.

Less surprisingly, some of the latent parameters impact the agent's performance in an intuitive way (*e.g.*, a horizontal force applied to the cart) (right Figure 3). This happens because the horizontal force directly impacts the action that the agent should choose: if the horizontal force pushes the cart right, then the agent should probably push it left, and vice

versa. The larger the difference between the horizontal force in the environment and assumed by the agent, the higher the drop in the agent's performance.

To distinguish between novelties that affect or do not affect the task performance, we use the regret of an optimal agent trained in the non-novel world and tested in the novel world, *i.e.*, $\mathcal{R}_{o,\mathcal{T}}(x_t, a_t^*)$. Novelties with $\mathcal{R}_{o,\mathcal{T}}(x_t, a_t^*) = 0$ do not affect task performance and can be ignored by agents that are near-optimal without a performance drop. By contrast, $\mathcal{R}_{o,\mathcal{T}}(x_t, a_t^*) > 0$ tells us that the novelty impacts the performance of an optimal non-adaptive agent and that the agent can update its state to improve its performance, *i.e.*, learn the novelty. Naturally, one can develop an adaptive version of this agent by learning an estimate of world state parameters from observations and using them to perform more accurate simulations and taking better actions.

Types of Novelty

For the definitions, we introduce the primary types of novelty, with the subtypes defined by combining primary types and regret (see Table 1 and (Boult et al. 2020)). Using our definitions of world, observation space, internal states, and the corresponding dissimilarity functions, we can formally define the following primary types of novelty.

World novelty. A world state $\breve{w} \in \mathcal{W}$ is considered a world novelty for agent α at time t if $\min_{w \in E_{t-1}} \mathcal{D}_{w,\mathcal{T}}(w, \breve{w}; E_{t-1}) > \delta_w$. That is, any world state $\breve{w}$ sufficiently dissimilar from every world state in the experience tensor is a world-level novelty. Only an oracle with access to E_{t-1} and $\mathcal{D}_{w,\mathcal{T}}$ can determine that a world state is truly novel. If the world representation is viewed as including distributional information (*e.g.*, probabilities of various items occurring) then a change in distributional parameters can be a world-level novelty even if no new "objects" occur in the world. Thus world-level novelty can produce problems of domain adaption, not just domain transfer.

Observation novelty. A world state $\breve{w} \in \mathcal{W}$ is considered an observation novelty for an agent α at time t iff $\min_{w \in E_{t-1}} \mathcal{D}_{o,\mathcal{T}}(w, \breve{w}; E_{t-1}) > \delta_o$. That is, an observation novelty is the observation-space state obtained for any world state $\breve{w}$ that, when projected through a perceptual operator, is sufficiently dissimilar from every observation-space state in the agent's experience tensor. Note that in this definition, the observed world state, $\breve{w}$, is subject to the current perceptual operator $\mathcal{P}$ at time t and is compared to the observation-space states in the experience tensor, which may have used the perceptual operator with potentially different parameter values (stored in world states). It is not surprising that the same world state may be novel at one point in time but not novel at another. However, it may be surprising that if the perceptual operator changes over time, then something can be perceptually novel at time t even if it was not perceptually novel at time $t-1$. For example, consider a transmission glitch creating errors in a static scene that has been viewed previously. It is important to note that observation novelty is defined considering all experience, which permits observation novelty that includes distributional shifts or reasoning about consecutive states to detect novelty in dynamics. If the agent had access to the true dissimilarity $\mathcal{D}_{o,\mathcal{T}}$, it could use that to define its state recognition function f. However, in practice, an agent will not have access to $\mathcal{D}_{o,\mathcal{T}}$, since it is trying to learn such a function from the data or was programmed with static rules to approximate it. Furthermore, agents rarely store all inputs.

Agent novelty. An observation-space state $x = \mathcal{P}(\breve{w} \in \mathcal{W})$ is considered an agent novelty for an agent α at time t *iff* $f_t(x) = \mathcal{N}$. That is, x is an agent novelty *iff* the agent at time t cannot map x to any of its internal states or maps to a special state for when it detects novel inputs. We note that this definition does not consider something novel if the state recognition functions f_t associate x with an incorrect state.

These novelty types are not mutually exclusive, and their combinations define the following notable novelty sub-types:

- **Unanimous novelty** is any world novelty w for which the perceptual operator produces an observation-space state that is both an observation novelty and an agent novelty. Unanimous novelty is correctly detected by the agent.
- **Imperceptible novelty** is any world novelty w for which the perceptual operator produces an observation space state x that is not an observation novelty. Accordingly, the agent cannot directly react to such novelties.
- **Faux novelty** is a world state w that is not a world novelty but its corresponding observation state x is an observation novelty or an agent novelty.
- **Ignored novelty** is any world state w such that its corresponding observation state x is not an agent novelty while either w is a world novelty or x is an observation novelty. Ignored novelty does not have to result in poor performance (*e.g.*, a non-adaptive agent may ignore all novelties while still performing well in the presence of them).

Combining these novelty types and sub-types with the regret functions ($\mathcal{R}_{f,\mathcal{T}}$, $\mathcal{R}_{o,\mathcal{T}}$, and $\mathcal{R}_{w,\mathcal{T}}$) allows us to formally define additional useful novelty sub-types including:

- **Managed novelty** is a world novelty w such that its implication on regret (performance) is minimal, *i.e.*, $\mathcal{R}_{f,\mathcal{T}}(w) < \epsilon$.
- **Nuisance novelty** is a novelty for which the world regret and the observation regret significantly disagree.

These are important for evaluations defining novelty ground-truth and associated world-regret, these sub-types need to be avoided or at least accounted for in evaluation metrics.

Novelty Types in CartPole

In the previous section, we defined and measured the dissimilarity and regret in the observation and world spaces in the CartPole domain. The novelties discussed there are world and observation novelties, since both world and perceptual dissimilarities are larger than zero, $\mathcal{D}_{o,\mathcal{T}}(w, \breve{w}) > 0$ and $\mathcal{D}_{w,\mathcal{T}}(w, \breve{w}) > 0$. The novelty in the pushing force magnitude, F_p, does not impact the two regrets of optimal non-novel agents, so this is either an ignored or unanimous managed novelty, depending whether the agent detects it.

If the CartPole environment had an additional unobserved cart and pole that did not influence the main cart

World Novelty	Observation Novelty	Agent Novelty	World Regret $\mathcal{R}_{w,\mathcal{T}} > \epsilon_w$		No World Regret $\mathcal{R}_{w,\mathcal{T}} \leq \epsilon_w$	
			Perceptual Regret $\mathcal{R}_{o,\mathcal{T}} > \epsilon_o$	No Perceptual Regret $\mathcal{R}_{o,\mathcal{T}} \leq \epsilon_o$	Perceptual Regret $\mathcal{R}_{o,\mathcal{T}} > \epsilon_o$	No Perceptual Regret $\mathcal{R}_{o,\mathcal{T}} \leq \epsilon_o$
Yes	Yes	Yes	Unanimous w/ Regret	Unanimous Nuisance	Unanimous Nuisance	Unanimous Managed
		No	Ignored	Ignored Nuisance	Ignored Nuisance	Ignored Managed
	No	Yes	Imperceptible	Imperceptible Nuis.	Imperceptible Nuis.	Managed Imperceptible
		No	Imperceptible Ignored	Imper. Ignored Nuis.	Imper. Ignored Nuis.	Managed Imperceptible
No	Yes	Yes	Faux	Faux Nuis.	Faux Nuis.	Managed Faux
		No	Faux Ignored	Faux Ignored Nuis.	Faux Ignored Nuis.	Managed Faux
	No	Yes	Faux	Faux Nuis.	Faux Nuis.	Managed Faux
		No	No novelty	No novelty Nuis.	No novelty Nuis.	No Novelty

Table 1: Subtypes of novelty defined by interaction of primary novelty types and regret. Some combinations of states get multiple labels (*e.g.*, Unanimous Nuisance is both Unanimous (all types of novelty present) and Nuisance (inconsistent regret values)).

and pole, then a change in the parameters of that unobserved pole would be an imperceptible world novelty, since $\mathcal{D}_{o,\mathcal{T}}(w, \breve{w}) = 0$ and $\mathcal{D}_{w,\mathcal{T}}(w, \breve{w}) > 0$. Since regrets do not depend on the additional cart, this is a managed imperceptible novelty. If world regret would also depend on the detection of such novelty, then it would be an imperceptible nuisance novelty, because observation regret does not depend on it.

Conclusion

We see three primary contributions of this formalization of novelty that will spur further research. First, formalization forces one to specify (or intentionally disregard) the required items in the theory. This can lead to insights about the problem and fill in knowledge gaps. For example, when applying the theory to the CartPole problem, numerous unanticipated issues were highlighted, new predictions made, and new experiments validated the new insights.

Second, formalization provides a common language to define and compare models of novelty across problems. The precision of terms reduces confusion, while the flexibility allows it to be applied to a wide range of problems.

Third, the formalization allows one to make predictions about where or why experiments incorporating some form of novelty might run into difficulties. For example, when the world-level and perceptual-level dissimilarity assessments disagree, we predict novelty problems will be more difficult. One example of difficulty is world-disparity using variables not represented in perceptual space. Another is when there are many possible world labels, but the input is only assigned one label that is used for assessing world-level dissimilarity. In this case, the theory predicts a greater difficulty with such novelty, especially if the assigned label is associated with a physically smaller aspect of the observation.

Biological intelligence has a remarkable capacity to generalize novel inputs with ease, yet artificial agents continue to struggle with this behavior. It is our hope that the adoption and use of the framework proposed here leads to the development of more effective solutions for novelty management and to make agents more robust to novel changes in their world.

By formalizing CartPole using our novelty framework, we gained insights into what are meaningful "novelty" problems for this task. We showed how to develop better measures to predict when novelty would be easy or hard to manage or to detect. In line with this, our team of researchers has been refining this theory and applying it to multiple problem domains. More details can be found in the longer arXiv version (Boult et al. 2020).

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under multiple contracts/agreements including HR001120C0055, W911NF-20-2-0005,W911NF-20-2-0004,HQ0034-19-D-0001, W911NF2020009. The views contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government.

References

Bendale, A.; and Boult, T. E. 2015. Towards Open World Recognition. In *IEEE CVPR*, 1893–1902.

Boult, T.; Grabowicz, P.; Prijatelj, D.; Stern, R.; Holder, L.; Alspector, J.; Jafarzadeh, M.; Ahmad, T.; Dhamija, A.; C.Li; S.Cruz; Shrivastava, A.; Vondrick, C.; and Scheirer, W. 2020. A Unifying Framework for Formal Theories of Novelty:Framework, Examples and Discussion. *arXiv preprint 2020* .

Langley, P. 2020. Open-World Learning for Radically Autonomous Agents. In *AAAI*, 13539–13543. JSTOR.

Markou, M.; and Singh, S. 2003a. Novelty detection: a review part 1: statistical approaches. *Signal processing* 83(12): 2481–2497.

Markou, M.; and Singh, S. 2003b. Novelty detection: a review part 2: neural network based approaches. *Signal processing* 83(12): 2499–2521.

Pimentel, M. A.; Clifton, D. A.; Clifton, L.; and Tarassenko, L. 2014. A review of novelty detection. *Signal Processing* 99: 215–249.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2013. Toward open set recognition. *IEEE TPAMI* 35(7): 1757–1772.

Scheirer, W. J.; Wilber, M. J.; Eckmann, M.; and Boult, T. E. 2014. Good recognition is non-metric. *Pattern Recognition* 47(8): 2721–2731.

Tversky, A. 1977. Features of similarity. *Psychological review* 84(4): 327.