Measuring the Performance of Open-World AI Systems

Vimukthini Pinto[1], Jochen Renz [1], Cheng Xue [1], Peng Zhang [1], Katarina Doctor [2], David W. Aha [2]

[1] School of Computing, The Australian National University, Canberra, Australia
[2] Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, D.C.
{vimukthini.inguruwattage, jochen.renz, cheng.xue, p.zhang}@anu.edu.au, {katarina.doctor, david.aha}@nrl.navy.mil

Abstract

Detecting and responding to novel and unforeseen situations is a key capability of human intelligence and remains a major challenge in modern artificial intelligence (AI). Open-world learning (OWL), an active and new research area, focuses on this challenge. However, there is a lack of systematic measures for evaluating OWL approaches. We address the question of how to evaluate the performance of AI methods for OWL by considering two tasks: Novelty detection and novelty reaction. We (1) argue that existing measures (e.g., accuracy, precision, and recall) are inappropriate for these tasks, (2) propose new performance measures for novelty detection and novelty reaction, and (3) evaluate them for a sample domain where novelty (i.e., an abrupt change in problem distribution) is introduced.

1 Introduction

Functioning in open-world environments (i.e., with novel and unforeseen situations) is a hallmark of human cognition. The field of AI has recently focused on creating intelligent systems that can detect and respond to sudden, long-term changes in their environment (Langley 2020; Boult et al. 2021). Such changes are very common in everyday situations. For example, every new item a person buys for a house is a sudden and a long term change to a household robot. DARPA has initiated a research program that focuses on open-world novelty (Senator 2019), which indicates the importance of this topic. Moving along in the same direction, domains such as Angry Birds (AIBirds 2021), Polycraft (Horner 2020), Monopoly (Baker 2020), and CartPole (Boult et al. 2021) have been used to create environments in which an agent can encounter novel situations. Simultaneously, AI systems are being developed to detect when a shift to a distribution occurs and respond appropriately (Klenk et al. 2020; Jafarzadeh et al. 2020; Schmitt 2020; Peng, Balloch, and Riedl 2021).

Empirical evaluations are critical for assessing and comparing the performance of learning algorithms. For example, in machine learning, many performance measures have been proposed to evaluate a model based on the learning task. Threshold, rank, and probability measures are three main groups of measures used for classification tasks (Caruana and Niculescu-Mizil 2006). Similarly, many evaluation measures are used for regression tasks (e.g., Mean Absolute Error, Root Mean Squared Error, R-Squared, and Adjusted R-Squared). Although many measures are already available, they exhibit limitations under certain circumstances (Yi et al. 2013). Therefore, an evaluator should consider factors such as the nature of the data, evaluation protocol, and the learning task when selecting a performance measure.

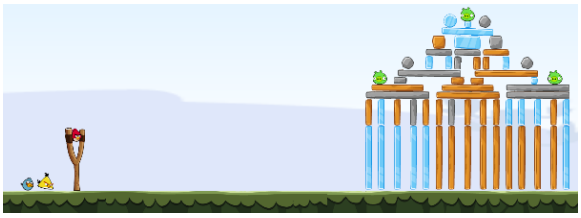
Appropriate measures for evaluating OWL systems have not been studied previously but are of critical importance (Langley 2020). Evaluation measures should characterize the performance of OWL agents for two tasks. The first, novelty detection, cannot be evaluated as a typical classification task because the measure must quantify both the agent's ability to correctly detect the novel problem distribution and the timeliness of that detection. The second, novelty adaptation, concerns the ability of the agent to react to the situations of the perceived novel distribution. Simply recording the agent's task performance is misleading as adaptation depends on the change in problem distribution as well as the agent's performance task. Therefore, we propose domain-independent evaluation measures to quantify agent performance for OWL environments.

2 Background

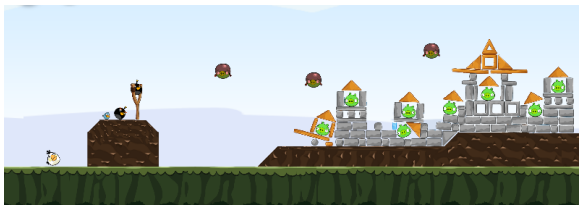
This section briefly describes the meaning of novelty, proposes a protocol to measure agent performance, and discusses related research to OWL.

2.1 Novelty Definition

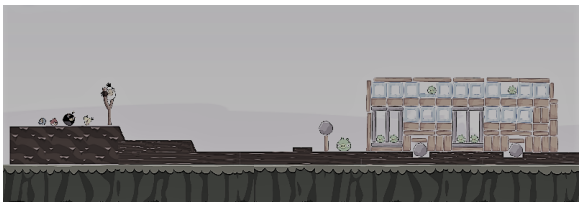
Novel problem distributions include situations that violate implicit or explicit assumptions about the agents, the environment, or their interactions (SAIL-ON 2019). Following this, Langley (2020) highlighted the need for a *theory of novelty* on OWL and described example transformations to the problem distribution that can each be considered as novel. Moreover, Boult et al. (2021) introduced a unifying framework that formalizes what it means for a problem to be considered as drawn from a novel distribution. According to these formalizations, we introduced novel problem distributions in an example domain, namely the video game Angry Birds. Figure 1 shows example novel probem distributions in Angry Birds.



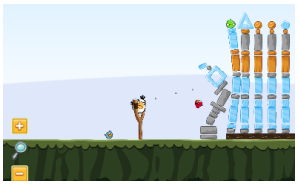
(a) Normal Angry Birds game environment with different types of blocks (wood, ice, stone), pigs, and birds



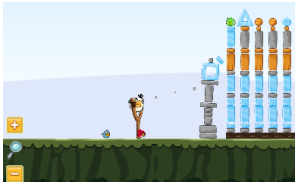
(b) An example novel transformation with a floating pig which is not available in the normal game environment



(c) An example novel transformation where the representation of the game environment is changed



(d) Original movements of the game



(e) Movements after a novel physical property in stone with the same shot made in 1d

Figure 1: Example novel problem distributions in Angry Birds. According to Langley's taxonomy (Langley 2020) 1b is an example *structural transformation*: a new category of an object is introduced. 1c is an example *spatio-temporal transformation*: it has altered the vision of an agent, and 1e is an example *structural transformation*: it has altered an existing attribute.

2.2 Evaluation Protocol

It is beneficial to have an agreed-upon protocol to empirically assess an agent's performance that can be applied to different domains. Based on the idea of having agents responding to sudden, long term changes in the environment (Langley 2020; Jafarzadeh et al. 2020), we want to compare how agents react when a novel problem distribution is introduced, whether and how long it takes them to detect this change, and how they can adjust their performance to successfully react to such distribution changes. We assume that we have state-of-the-art agents, which we call *baseline agents*, for operating in a given domain (i.e., on the non-novel problem distribution). After a problem distribution changes, we assume that novel distribution persists in the environment. While baseline agents may not be able to detect and properly react to novel distributions, we expect *novelty agents* to detect, and once detected, to adjust to novel distributions. Taking these considerations into account, we propose a general protocol for evaluating agent performance with novel problem distributions in a given domain. We call D_{pre} the pre-novelty distribution and D^n_{post} the post-novelty distribution with a specific (novel) distribution change n. Following points explains the protocol:

1. Agents are first exposed to a sequence of pre-novelty (i.e., non-novel) instances drawn from D_{pre}. The number of pre-novelty instances is not known to the agent. Agents can attempt to solve each problem instance once in the given order. [1]
2. At some point, the problem distribution switches from D_{pre} to D^n_{post}. We refer to this switch as the *distribution change*. All subsequent problem instances are drawn from D^n_{post}. The number of post-novelty instances is unknown to the agent. Agents can attempt to solve problem instances only once in the given order.
3. For every instance (pre- or post-novelty) i an agent attempts to solve, we record its task performance (e.g., score) TSP_i and p_i, the probability that the agent believes a distribution change has occurred. Task performance reflects how well the agent solves a problem instance; it is a task- and domain-dependent measure.

We refer to the above sequence of pre- and post-novelty instances as a single *trial*. T^n_j is the j^{th} *trial* for a given problem distribution change n. We refer a set of trials with the same post-novelty distribution as a *trial-set*. An *experiment* is a set of trial-sets. When an agent completes a trial, it is reset to its initial state before it begins the next trial (i.e., agents are permitted to learn throughout a trial, but learned models are not transferred between trials[2]). The agent also reports a *detection threshold* where each p_i exceeding the threshold indicates a predicted distribution change (i.e., the agent predicts that a distribution change has occurred).

During the training stage, agents are given problems drawn from the pre-novelty distribution, but not from D^n_{post}. For this paper, we further assume that, in addition to being distinct, D_{pre} and D^n_{post} are disjoint (i.e., that the agents were not trained on any problem instances from D_{post}). Figure 2 illustrates this protocol.

[1] This design decision ensures that agents do not have the choice of selecting the order and the agent is always presented with pre-novelty instances before the post-novelty instances.

[2] This is to ensure that the same agent is tested on all trials within a trial-set.

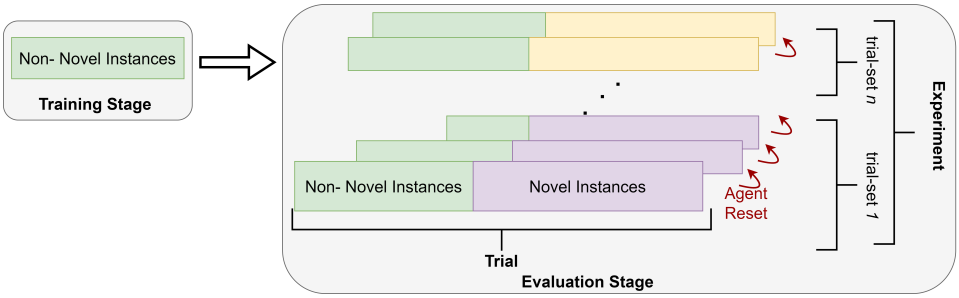


Figure 2: Training stage followed by the evaluation stage. An experiment contains *trial-sets* that are drawn from different novel distributions. A trial contains variable lengths of problem instances drawn first from a pre-novelty distribution and then from a post-novelty distribution. We assume the agent is an online learner, and we record its performance throughout a trial. The agent's model is reset at the end of each trial to eliminate model transfer.

2.3 Related Research

OWL is related to many other research paradigms (Langley 2020). For instance, the novelty detection task in OWL is similar to tasks such as anomaly detection, outlier detection, and out-of-distribution detection (Pimentel et al. 2014; Hodge and Austin 2004; Markou and Singh 2003). However, the standard performance measures such as accuracy, precision, and recall used for these tasks become less useful for OWL mainly due to the ordered sequence of instances present in the OWL protocol. Change detection problems (Pears, Sripirakas, and Koh 2014; Sebastião and Gama 2009) that contain a data stream are related to OWL, and some of the measures we propose are inspired by this research field. However, unlike OWL, these tasks do not involve a trial setup (detection metrics are discussed in detail in Section 3.1).

Learning in streams (Lu et al. 2018) and transfer of learned expertise (Senator 2011) are related to the novelty reaction task in OWL. However, learning in streams typically addresses classification tasks and transfer of learned expertise is used when a change of distribution is informed (Langley 2020). As OWL is not limited to classification tasks and as the change in distribution is not informed, measures in these areas become less useful in the OWL protocol. Moreover, other paradigms such as few-shot learning (Wang et al. 2020), zero-shot learning (Wang et al. 2019), and incremental learning (Khreich et al. 2012; Yan, Xie, and He 2021a) do not consist of a change detection task.

3 Evaluation Measures

We discuss an agent's empirical evaluation using two types of measures: 1) *Detection measures*, which quantify the agent's ability to detect a problem distribution change, and 2) *Reaction measures*, which quantify the agent's ability to adjust to such a change.

3.1 Detection Measures

Agents that operate in an OWL environment should not predict a distribution change before the point it changes. After the change occurs, the agent should quickly identify that a problem distribution has changed. Figure 3 illustrates six possible variations of an agent's detection performance within a single trial. Assuming the detection threshold is 0.5, the perfect agent estimates low probabilities in the pre-novelty instances and peaks as soon as the distribution change occurs. Agents 3 and 4 represent two extreme scenarios where Agent 3 believes the pre-novelty distribution to be novel while Agent 4 believes the post-novelty distribution to be non-novel throughout the trial. Agent 5 detects the post-novelty distribution before it is introduced. None of these agents correctly detect this distribution change. Agents 1 and 2 are desirable, where Agent 1 detects the novel distribution faster. Thus, our performance measures should ideally capture the correct identification of the distribution change and the timeliness of that detection.

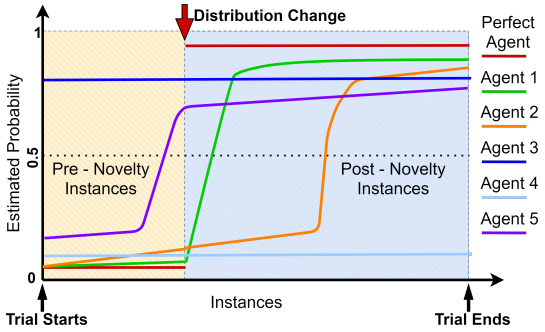


Figure 3: Estimated probability of the distribution change as reported by six agents over a single trial.

Existing Measures Novelty detection - the identification of dataset instances that do not match well to a known distribution - is a widely studied problem for learning systems (Markou and Singh 2003; Marsland 2001). Several measures have been used to assess the quality of detection such as accuracy, balanced accuracy (BA), precision, recall, F-measure, and area under the ROC (Straube and Krell 2014; Hernández-Orallo, Flach, and Ferri 2012) Table 1 summarizes how each of the existing measures can be formulated to the OWL protocol. Unfortunately, these measures are less useful for the novelty detection task in OWL environments due to the nature of the protocol. The OWL protocol con-

Measure	Formulation of the measure for OWL protocol
Accuracy	*Accuracy* = $\frac{1}{\|T\|}\sum_{t=1}^{\|T\|} Accuracy_t$ For each trial, *t*: $Accuracy_t = \frac{TP_t+TN_t}{TP_t+TN_t+FP_t+FN_t}$
Balanced Accuracy (BA)	*BA* = $\frac{1}{\|T\|}\sum_{t=1}^{\|T\|} BA_t$ For each trial, *t*: $BA_t = 0.5 \times (TPR_t + TNR_t)$ where: $TPR_t = \frac{TP_t}{TP_t+FN_t}$ $TNR_t = \frac{TN_t}{TN_t+FP_t}$
Precision	*Precision* = $\frac{1}{\|T\|}\sum_{t=1}^{\|T\|} Precision_t$ For each trial, *t*: $Precision_t = \frac{TP_t}{TP_t+FP_t}$
Recall	*Recall* = $\frac{1}{\|T\|}\sum_{t=1}^{\|T\|} Recall_t$ For each trial, *t*: $Recall_t = \frac{TP_t}{TP_t+FN_t}$
F1	*F1* = $\frac{1}{\|T\|}\sum_{t=1}^{\|T\|} F1_t$ For each trial, *t*: $F1_t = \frac{2\times precision_t \times recall_t}{precision_t+recall_t}$

Table 1: Formulation of the existing measures that can be used in the novelty detection task for OWL. TP_t,TN_t,FP_t,FN_t are true-positive, true-negative, false-positive, and false-negative in trial *t* where positives are considered as post-novelty instances (instances from the post-novelty distribution) and negatives are pre-novelty instances (instances from the pre-novelty distribution). T is the trial-set.

tains ordered data (i.e., a sequence of post-novelty problem instances that follow a sequence of pre-novelty instances), whereas there is no such order in (batch) classification tasks. Moreover, measures derived from a confusion matrix are affected by the number and proportion of pre- and post-novelty instances in trials, which can yield misleading performance comparisons. If an agent reports detection starting from pre-novelty instances in a trial with few pre-novelty instances, measures that are sensitive to class imbalance may indicate a high performance without penalizing its false detection. Measures that are insensitive to class imbalance, such as BA, also fail in certain cases. For example, a good measure should indicate 0 for Agents 3 and 4 in Figure 3 as none of them correctly detect the changed distribution. However, BA misleadingly indicates 0.5.

Measures such as accuracy and BA depend on the number of false positive (FP) predictions. However, according to the protocol, an agent that detects early does not know how many more pre-novelty instances appear before the distribution change. Therefore, an ideal detection measure should not distinguish where an agent's first detection occurs if it is a FP prediction, all FP detections are equally problematic for OWL tasks. Otherwise, all measures derived from a confusion matrix suffer from not quantifying how quickly an agent detects the distribution change. This makes the OWL task seem similar to change point detection (Aminikhanghahi and Cook 2017). The measure should ideally capture how many post-novelty problems instances an agent requires to detect the distribution change. Cumulative sum control (CUSUM) chart (Page 1954), a statistical quality control chart, and the activity monitoring operating characteristic (AMOC) curve (Fawcett and Provost 1999), which is often used to measure the performance of event surveillance systems, are preferable alternatives. However, we cannot directly use the CUSUM chart as it considers mean shifts whereas our evaluation considers the detection threshold that the agent provides. AMOC generally evaluates the trade-off between timeliness of detection and the false alarm rate (Jiang, Cooper, and Neill 2009). However, in our case of novelty detection performance, this technique becomes less useful because AMOC considers the false alarm rate whereas we are only interested in the presence of FP but not the FP rate. Change detection problems (Pears, Sripirakas, and Koh 2014; Sebastião and Gama 2009) that generally consists of gradual multiple changes throughout a data stream also suggests to evaluate models using false alarm rate, detection accuracy, and detection delay. As mentioned earlier, we are concerned with the presence of a FP and not the rate. Moreover, when the agent indicates a detection in a data steam with multiple change points, it is not clear if the agent's detection is correct unless the detection is made on the change point. On the other hand, change detection problems do not comprise of trials as in the OWL protocol. Inspired by all of these measures, we propose the following measures to suit our protocol.

Ideal OWL detection measures should be independent of the number of pre- and post-novelty instances, trial order, and the number of FP instances. Measures should capture the correctness and timeliness of detection.

Proposed Measures We propose two measures that avoid the limitations of standard measures. Ours measure an agent's novelty detection ability in terms of its correctness and timeliness, respectively.

Percentage of correctly detected trials (CDT): A correctly detected trial is one where an agent predicts that the distribution has changed among only the post-novelty instances (i.e., there is at least one true positive (TP) but no false positives (FP)).

$$CDT = \frac{1}{|T|}\sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t = 0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where FP_t and TP_t are the number of false and true positive detections made in the t^{th} trial, and *T* is a trial-set.

Average number of instances to detect novelty (IDN): This quantifies the timeliness of detection using the number of problem instances required to correctly detect the novel

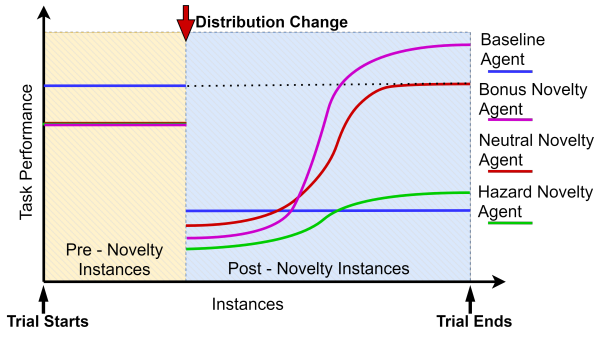


Figure 4: Distinction between *bonus, hazard*, and *neutral* novelty categories.

distribution.

$$IDN = \frac{1}{N_{cdt}} \sum_{t=1}^{|T|} \begin{cases} n_t, & \text{if } FP_t{=}0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where,

$$N_{cdt} = \sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t{=}0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

and n_t is the number of *FN* instances until the first *TP* instance in trial *t*.

These measures can be collectively used to assess an agent's ability to correctly detect the novel distribution and to quantify how quickly an agent detects it. These measures assume an agent provides a consistent detection (i.e., its estimated probability exceeds the detection threshold only once and then remains above threshold for the rest of the trial). This assumption is made according to the OWL task, as we assume OWL agents are told that a distribution change occurs only once within a trial. However, if this assumption is violated, we can use the percentage of correctly and consistently detected trials and the number of instances to detect the distribution change consistently. Consistent detection means all reported probabilities exceed the detection threshold after the distribution change until a trial ends.

These two measures are independent of trial length, the number of trials, trial order, the number of pre- and post-novelty instances and the number of FP instances. They independently quantify how effective and efficient an agent is in detecting a distribution change. They can be used to compare the detection performance of multiple agents and across different distribution changes. However, they do not distinguish between trials with FP detections and trials with no positive detections. If this distinction is important, we can consider the percentage of wrongly detected trials (WDT) (i.e., trials with FP detections) as an additional measure.

$$WDT = \frac{1}{|T|} \sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

3.2 Reaction Measures

Reaction measures are used to quantify an agent's ability to adjust to the novel instance distribution. For explanation purposes, we define three novelty categories by considering the maximum possible performance of an agent on post-novelty instances (see Figure 4). We refer to a post-novelty distribution to be a *bonus*, *hazard*, or *neutral* novelty distribution if the best agents can achieve higher, lower, or roughly equal performance, respectively, than a baseline agent can achieve on pre-novelty instances. An ideal reaction measure should quantify how well an agent adjusts to a distribution irrespective of the novelty category distribution.

Existing Measures Jafarzadeh et al. (2020) introduced a measure for open-world classification tasks. However, this cannot be used in domains such as Angry Birds, where the performance task is not classification. Moreover, Jafarzadeh et al. highlight the use of normalized mutual information, which is not applicable to tasks where performance is measured in an unbounded numeric scale.

Incremental learning (Yan, Xie, and He 2021b; Ade and Deshmukh 2013), where the learning process takes place when new examples emerge, employs three criteria that should be considered when evaluating models in different domains: stability, improvement, and recoverability (Syed, Liu, and Sung 1999). There are no explicit formulations available as evaluation measures can be defined to suit a domain satisfying the three criteria. Moreover, there is no trial setup in incremental learning problems. Other fields of research such as learning under concept drift (Lu et al. 2018), few-shot learning (Wang et al. 2020), zero-shot learning (Wang et al. 2019), and transfer learning (Pan and Yang 2010) can be viewed as reaction tasks. However, these fields use task-specific performance evaluation measures (e.g., classification accuracy in few shot-shot learning classification tasks, game score in an Atari game with few-shot learning).

We expect OWL reaction measures to be independent of the domain, novelty distribution category, trial order, and number of pre- and post-novelty instances in a trial. The measures should enable us to compare and quantify each agent's performance.

Proposed Measures We first show a baseline measure that suits the OWL protocol and then propose two measures that can be used collectively to assess novelty reaction behaviours of agents.

Novelty reaction performance (*NRP*): To our best knowledge, there are no standard methods we can directly adapt to suit the OWL protocol. One obvious novelty reaction performance measure is to capture if the novelty agent (NT) performs on post-novelty instances at least as well as the baseline agent (BL) performance on pre-novelty instances. This measure is considered as the baseline measure. The measure is as given below:

$$NRP = \frac{1}{|T|} \sum_{t=1}^{|T|} NRP_t \quad (5)$$

For each trial, *t*:

$$NRP_t = \frac{P_{post,NT,t}}{P_{pre,BL,t} + P_{post,NT,t}} \quad (6)$$

where,

$$P_{post,NT,t} = \frac{1}{n_{post,t}} \sum_{i=n_{pre,t}}^{n_t} TSP_{i,NT,t} \quad (7)$$

$$P_{pre,BL,t} = \frac{1}{n_{pre,t}} \sum_{i=0}^{n_{pre,t}} TSP_{i,BL,t} \quad (8)$$

$TSP_{i,j,t}$: Agent *j*'s task performance for instance *i* in t^{th} trial

j : {NT, BL}

n_t : Total number of instances trial *t*

$n_{pre,t}$: Number of pre-novelty instances in trial *t*

$n_{post,t}$: Number of post-novelty instances in trial *t*

NRP>0.5 indicates that the novelty agent outperforms BL on pre-novelty instances. Therefore, this measure only enables us to determine whether an agent yields the expected performance for each novelty category. However, it does not allow us to identify whether an agent performs at least as well as BL on post-novelty instances. Moreover, we cannot distinguish agent performance based solely on this measure. For example, even though an agent applied to hazard novelty distributions may learn, it always yields a value of *NRP*<0.5. Similarly, an *NRP*>0.5 may be found for bonus novelty distributions even for agents that do not adapt to the novel distribution.

To overcome the drawbacks of the *NRP* measure, we propose the following two measures that are independent of the novelty distribution category.

Asymptotic novelty reaction performance (*ANRP*): *ANRP* attempts to quantify the performance of an agent independent of the novelty distribution category. This measures the performance of the novelty agent versus the BL on the *same* set of post-novelty instances.

$$ANRP = \frac{1}{|T|} \sum_{t=1}^{|T|} ANRP_t \quad (9)$$

For each trial, *t*:

$$ANRP_t = \frac{P_{post\_asymptotic,NT,t}}{P_{post\_asymptotic,BL,t} + P_{post\_asymptotic,NT,t}} \quad (10)$$

where,

$$P_{post\_asymptotic,j,t} = \frac{1}{m_{1,t}} \sum_{i=n_t - m_{1,t}}^{n_t} TSP_{i,j,t} \quad (11)$$

$m_{1,t}$: Length of the final subsequence of the post-novelty instances in trial *t*

and $TSP_{i,j,t}$, *j* and n_t are defined as in *NRP*.

The value of $m_{1,t}$ can be adjusted based on the domain and the experimental setting using a suitable percentage of $n_{post,t}$ (e.g., 10% of $n_{post,t}$). Thus, $P_{post\_asymptotic,j,t}$ captures the j^{th} agent's average performance at the end of trial *t* based on the predefined asymptotic length (length of the final subsequence of the post-novelty instances). If $P_{post\_asymptotic,NT,t}$ = 0, we consider $ANRP_t$ to be zero as there is no performance. *ANRP* > 0.5 indicates that the novelty agent outperforms BL.

One limitation of the measure is that it does not assess whether an agent improves over time. For this purpose, we propose the following measure.

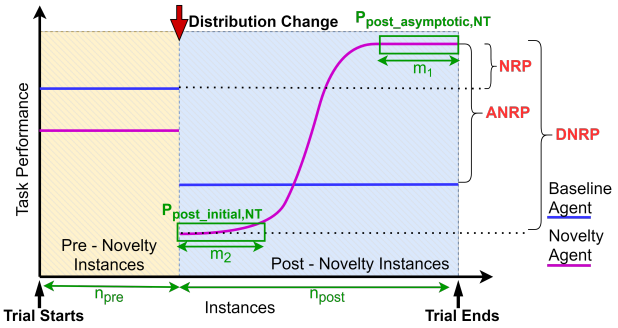


Figure 5: Overview of the novelty reaction measures.

Double-ended novelty reaction performance (*DNRP*): *DNRP* measures whether the agent improves its performance over a post-novelty sequence.

$$DNRP = \frac{1}{|T|} \sum_{t=1}^{|T|} DNRP_t \quad (12)$$

For each trial, *t*:

$$DNRP_t = \frac{P_{post\_asymptotic,NT,t}}{P_{post\_initial,NT,t} + P_{post\_asymptotic,NT,t}} \quad (13)$$

where,

$$P_{post\_initial,j,t} = \frac{1}{m_{2,t}} \sum_{i=n_{pre,t}}^{n_{pre,t}+m_{2,t}} TSP_{i,j,t} \quad (14)$$

$m_{2,t}$: Length of initial subsequence of the post-novelty instances in trail *t*

and $P_{post\_asymptotic,NT,t}$, $TSP_{i,j,t}$, *j*, and $n_{pre,t}$ are defined as in *ANRP*.

Similar to the *ANRP*, $m_{1,t}$ and $m_{2,t}$ should be adjusted based on the experimental setting and if $P_{post\_asymptotic,NT,t}$ = 0, we consider $DNRP_t$ to be zero as there is no performance.

If *DNRP* > 0.5, it implies that the agent has improved over the post-novelty sequence. If *DNRP* $\simeq$ 0.5, it implies that the agent did not adjust or it reacted to the novel distribution as soon as the problem distribution changed.

In summary, *NRP* depends on the three novelty distribution categories. To eliminate that, *ANRP* instead compares BL and NT on the same post novelty distribution. Finally, *DNRP* measures an agent's improvement in the post-novelty distribution. Collectively, *ANRP* and *DNRP* measures capture an agent's reaction ability (see Figure 5).

4 Demonstration in a Sample Domain

Our empirical study's protocol and evaluation measures are general and intended to work for every OWL task. We use the research clone of the popular physics-based puzzle game Angry Birds (Ferreira and Toledo 2014) as an example domain for introducing novel distributions. Angry Birds is a popular domain for developing and evaluating AI agents that operate in a simulated physical world, with a long-running AI competition held at IJCAI conferences (Renz et al. 2015). Angry Birds mimics real-world environments with physics concepts such as gravity, friction, and mass. This makes it

Case	Agent	ND	Accuracy	BA	Precision	Recall	F1	CDT	WDT	IDN
1	A	ND1	0.16	0.52	NaN	0.10	NaN	0.00	0.10	-
2	A	ND16	0.19	0.53	NaN	0.12	NaN	0.00	0.12	-
3	B	ND8	0.14	0.52	NaN	0.04	NaN	0.54	0.00	52.89
4	A	ND18	0.35	0.64	NaN	0.29	NaN	0.44	0.04	31.59
5	B	ND7	0.35	0.64	NaN	0.29	NaN	0.74	0.02	39.19
6	B	ND13	0.70	0.84	NaN	0.67	NaN	0.96	0.02	28.42
7	A	ND15	0.93	0.93	0.99	0.93	0.96	0.84	0.16	8.83
8	B	ND1	0.27	0.61	NaN	0.21	NaN	0.60	0.00	44.27
9	B	ND14	0.96	0.98	1.00	0.96	0.98	1.00	0.00	4.30
10	B	ND15	0.93	0.96	1.00	0.92	0.96	1.00	0.00	8.32

Table 2: Comparison of our proposed Novelty Detection measures with standard performance measures. ND refers to the novel distribution. While existing measures show drawbacks, measures we propose allow more intuitive performance comparisons for OWL tasks.

an ideal platform to add realistic novel problem distributions (Gamage et al. 2021). We have added such distributions that align with the formal theories on novelty mentioned in Section 2.1.

In Angry Birds, a problem instance is a game level. BL performance is measured on pre-novelty instances, where the performance task is to solve the game with maximum score. More than 60 agents have participated in the AIBirds competition in prior years (AIBirds 2021). Several competition winners are available in open source and can be used as baseline agents. In 2021, there was a new novelty track to the AIBirds competition to encourage the development of AI systems that can react to novel distributions as efficiently and as effectively as humans (AIBirds-NoveltyTrack 2021). We conducted our experiment with two agents designed to detect and respond well to such distributions, which we call Agents A and B. Our experiment contains 18 trial-sets (i.e., 18 novel distributions), where each trial-set contains 50 trials. Each trial contains 10-20 pre-novelty instances and 100 post-novelty instances. We use sample cases from the experiment to discuss our measures.

4.1 Novelty Detection

We compare the results of our experiment with common measures to justify the use of our proposed measures (see Table 2).

Impact of the number of pre-/post-novelty instances: Cases 1 and 2 represent measures collected from Agent A's performance with two novel problem distributions. The agent did not detect the novel distribution in any trial (CDT = 0%). However, the standard measures produce varying values due to the difference in the number of pre- and post-novelty instances per trial. Case 3 represents measures from Agent B collected when applied to novel distribution ND8. All standard measures produce lower values in comparison to Cases 1 and 2, falsely indicating a lower performance by the agent. However, the agent detected 54% of the trials correctly (CDT = 54%) and, interestingly, it recorded no wrongly detected trials (WDT = 0%). Furthermore, IDN indicates that the agent required 52.89 instances on average to detect this distribution change. For Cases 4 and 5, all of the standard measures produce the same values. However, CDT for Case 5 was 74% and was 44% for Case 4. While the standard measures instead indicate a similar performance due to the difference in number of pre- and post-novelty instances, CDT enables identifying whether agents detect the distribution change correctly.

Impact of miss detection and point of FP detection: For Case 6, Agent B detected 96% of the trials correctly but precision and F1 cannot be defined, as the agent did not detect the distribution change for some trials. This is an example where some standard measures cannot be defined, while CDT, WDT, and IDN convey important information concerning detection performance. In Case 7, all the standard measures report high values even though the agent falsely detected a distribution change (before the point of change) in 16% of the trials. That is, standard measures can falsely indicate a high performance because they are sensitive to when a FP detection occurs.

Performance comparison with proposed measures: Cases 1 and 8 exemplify that Agent B outperforms Agent A (for the CDT measure) for novel distribution ND1. Interestingly, Agent B did not detect a distribution change in the pre-novelty instances (WDT = 0%). Cases 9 and 10 depict examples for which the agent has equal CDT, and IDN contrasts their performance by measuring their timeliness of detection. As Agent B has lower IDN for novel distribution ND14 (Case 9) than ND15 (Case 10) with the same CDT, we can conclude that Agent B performs better for ND14. Another interesting scenario involves comparing two agents' performance with low CDT and low IDN versus high CDT and high IDN, respectively. Agents with high CDT are, in general, preferable as they have a greater potential to improve in the future by improving IDN.

In summary, our proposed measures successfully distinguished the agents' detection performance, independent of the number of pre- and post-novelty instances, trial order, and number of FP instances. This is not the case for standard measures, as demonstrated in the cases we described.

Case-Agent	ND	NRP	ANRP	DNRP
1 - A	ND2 : *bonus*	0.52	0.49	0.50
2 - B	ND2 : *bonus*	0.54	0.51	0.50
3 - B	ND14 : *bonus*	0.48	0.65	0.49
4 - B	ND1 : *neutral*	0.50	0.50	0.50
5 - B	ND18 : *neutral*	0.41	0.57	0.49
6 - A	ND15 : *hazard*	0.00	0.00	0.00
7 - B	ND15 : *hazard*	0.12	0.74	0.50
8 - A	ND4 : *bonus*	0.49	0.50	0.53

Table 3: Summary of novelty reaction measures. ND refers to the novel distribution. NRP can be used to compare agents only within a single novel distribution and depends on the novelty distribution category (bonus/hazard nature). ANRP extends this comparison by removing the positive or negative effect of novel distributions, and DNRP measures agent improvement on post-novelty problem sequences.

4.2 Novelty Reaction

In Angry Birds, task performance can be measured using the score at the end of each instance. We used the mean performance of three AIBirds competition agents, namely Data Lab, Naïve, and Eagle's Wing as the baseline agents (as there is no clear best agent in Angry Birds). Table 3 displays the novelty reaction performance measure results for representative cases from our experiment. The length of the final subsequence and the initial subsequence of the post-novelty instances is taken as 10 (m_1=m_2=*10*, which is 10% of post-novelty instances).

Cases 1 and 2 contrast the measures for Agents A and B for novel distribution ND2. As NRP indicates, Agent B outperforms A on post-novelty instances. As these two cases are from a bonus novel distribution category, it is expected that NRP > 0.5. However, ANRP highlights that Agent A has not adapted to the novel distribution (ANRP$<$0.5). Similarly, in Case 2, Agent B performs only slightly better than the BL for post-novelty instances (ANRP=0.51). Even though Case 3 is also from the bonus category, NRP indicates that Agent B has not attained BL's task performance on pre-novelty instances. However, ANRP=0.65 shows that it has adapted better to the novel distribution when compared to BL.

Cases 4 and 5 display the measures of agents operating in the neutral novel distribution category. For Case 4, Agent B's NRP ≈ 0.5, implying that Agent B novelty performance has reached baseline agent's pre-novelty performance. That is expected that novelty agents can only reach 0.5 as ND1 belongs to the neutral novelty category. However, ANRP = 0.5 indicates that Agent B is not better than the BL agent in ND1 and DNRP further indicates no improvement over time. For Case 5, although Agent B did not attain the performance of the BL in pre-novelty, ANRP = 0.57 > 0.5 confirms that B outperforms BL for post-novelty instances. Moreover, DNRP ≈ 0.5 indicates that the agent adapted to the novel distribution as soon as distribution changed.

Cases 6 and 7 display measures of agents operating in the hazard novel distribution category. Thus, NRP < 0.5 irrespective of how the agent performs. In Case 6, ANRP and DNRP are both zero, implying that agent A has not performed in the post-novelty instances. For Case 7, as ANRP = 0.74 > 0.5, we can conclude that Agent B has adapted to the novel distribution but DNRP ≈ 0.5 shows that there is no improvement during the post-novelty sequence.

Case 8 shows a case where an agent shows an improvement over the post-novelty problems (The mean-value 0.53$>$0.50 was statistically significant at 5% level of significance) even though ANRP is marginal.

These cases provide evidence that NRP only helps to compare agents within a single novel distribution, and is dependent on the nature of the novelty distribution category. i.e., for example, all novelty distributions in the hazard category would be less than 0.5 irrespective of the agent performance. ANRP helps to compare them by eliminating the positive or negative effect of a novel distribution on the achievable score, and DNRP expresses an agent's improvement. The two measures are needed to obtain a complete assessment of an agent's ability to react to a novel distribution. These measures can also be used to compare how an agent responds to different novel distributions.

5 Conclusion

Identifying the best evaluation measures for a given task is a long-standing challenge in AI research. Inspired by performance measures in other research fields, we presented performance evaluation measures to assess the capabilities of agents for detecting and reacting to novel problem distributions in open-world learning environments. Detection measures evaluate the correctness and timeliness of detection whilst the reaction measures determine whether an agent has successfully adapted to the novel distribution. The proposed measures are domain-independent, and using a sample domain we have demonstrated that they do not suffer from the same flaws as standard measures; they accurately and intuitively distinguish good versus poor agent performance and demonstrate how agents improve. Our measures can be used collectively to understand the agent performance in detection and reaction independent of the domain and independent of the novel distribution.

As part of our future work, we plan to extend these measures to assess the novelty characterization ability of agents (i.e., to evaluate whether an agent correctly detects what is novel in a problem distribution and whether an agent understands the impact caused by the novel distribution). With these measures, we believe we set a foundation to address a major concern for evaluating AI systems in a growing research area.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) and was accomplished under Cooperative Agreement Number W911NF-20-2-0002. The views and conclusions 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. The U.S. Government is authorized to reproduce and distribute reprints for Government

purposes notwithstanding any copyright notation herein.

References

Ade, R.; and Deshmukh, R. 2013. Methods for Incremental Learning: A Survey. *International Journal of Data Mining and Knowledge Management Process*, 3: 119–125.

AIBirds. 2021. Angry Birds AI Competition. http://aibirds.org/. [Accessed: August. 22, 2021].

AIBirds-NoveltyTrack. 2021. Angry Birds AI Competition, Novelty Track. http://aibirds.org/angry-birds-ai-competition/novelty-track.html. [Accessed: August. 23, 2021].

Aminikhanghahi, S.; and Cook, D. 2017. A Survey of Methods for Time Series Change Point Detection. *Knowledge and Information Systems*, 51.

Baker, A. 2020. Want to Teach An AI Novelty? First, Teach It Monopoly. Then Throw Out the Rules. https://viterbischool.usc.edu/news/2020/07/want-to-teach-an-ai-novelty-first-teach-it-monopoly-then-throw-out-the-rules/. [Accessed: August. 18, 2021].

Boult, T.; Grabowicz, P. A.; Prijatelj, D.; Stern, R.; Holder, L.; Alspector, J.; Jafarzadeh, M.; Ahmad, T.; Dhamija, A. R.; Cli; Cruz, S.; Shrivastava, A.; Vondrick, C.; and Scheirer, W. 2021. Towards a Unifying Framework for Formal Theories of Novelty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 15047–15052.

Caruana, R.; and Niculescu-Mizil, A. 2006. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006: 161–168.

Fawcett, T.; and Provost, F. 1999. Activity Monitoring: Noticing interesting changes in behavior. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ferreira, L.; and Toledo, C. 2014. A search-based approach for generating Angry Birds levels. In *2014 IEEE Conference on Computational Intelligence and Games*, 1–8.

Gamage, C.; Pinto, V.; Xue, C.; Stephenson, M.; Zhang, P.; and Renz, J. 2021. Novelty Generation Framework for AI Agents in Angry Birds Style Physics Games. In *2021 IEEE Conference of Games, COG 2021*.

Hernández-Orallo, J.; Flach, P.; and Ferri, C. 2012. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. *J. Mach. Learn. Res.*, 13(1): 2813–2869.

Hodge, V.; and Austin, J. 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22: 85–126.

Horner, K. 2020. Polycraft Team To Lay Groundwork for Smarter AI. https://news.utdallas.edu/science-technology/polycraft-world-darpa-2020/. [Accessed: Aug. 08, 2020].

Jafarzadeh, M.; Dhamija, A. R.; Cruz, S.; Li, C.; Ahmad, T.; and Boult, T. 2020. Open-World Learning Without Labels. *ArXiv*, abs/2011.12906.

Jiang, X.; Cooper, G.; and Neill, D. 2009. Generalized AMOC Curves For Evaluation and Improvement of Event Surveillance. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009: 281–5.

Khreich, W.; Granger, E.; Miri, A.; and Sabourin, R. 2012. A survey of techniques for incremental learning of HMM parameters. *Information Sciences*, 197: 105–130.

Klenk, M.; Piotrowski, W.; Stern, R.; Mohan, S.; and Kleer, J. d. 2020. Model-Based Novelty Adaptation for Open-World AI. https://advancesincognitivesystems.github.io/acs/data/ACS2020_paper_5.pdf.

Langley, P. 2020. Open-World Learning for Radically Autonomous Agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 13539–13543.

Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; and Zhang, G. 2018. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, PP: 1–1.

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

Marsland, S. 2001. Novelty Detection in Learning Systems. *Neural Computing Surveys*, 3.

Page, E. S. 1954. Continuous Inspection Schemes. *Biometrika*, 41(1-2): 100–115.

Pan, S. J.; and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Pears, R.; Sripirakas, S.; and Koh, Y. S. 2014. Detecting concept change in dynamic data streams. *Machine Learning*, 97.

Peng, X.; Balloch, J. C.; and Riedl, M. O. 2021. Detecting and Adapting to Novelty in Games. *arXiv e-prints*, arXiv–2106.

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

Renz, J.; Ge, X.; Gould, S.; and Zhang, P. 2015. The Angry Birds AI Competition. *AI Magazine*, 36: 85–87.

SAIL-ON. 2019. Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON). https://iresearch-cms.tau.ac.il/sites/resauth.tau.ac.il/files/DARPA\%20SAIL-ON_HR001119S0038.pdf. [Accessed: May. 23, 2021].

Schmitt, S. 2020. Kitware Wins DARPA Contract to Develop Artificial Intelligence Systems that Adapt to Novel Conditions. https://blog.kitware.com/. [Accessed: May. 23, 2021].

Sebastião, R.; and Gama, J. 2009. A Study on Change Detection Methods. *Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence, EPIA 2009*.

Senator, T. 2011. Transfer Learning: Progress and Potential. *AI Megazine*, 32: 84–86.

Senator, T. 2019. Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON).

https://www.darpa.mil/program/science-of-artificial-intelligence-and-learning-for-open-world-novelty. [Accessed: August. 29, 2021].

Straube, S.; and Krell, M. M. 2014. How to evaluate an agent's behavior to infrequent events?—Reliable performance estimation insensitive to class distribution. *Frontiers in Computational Neuroscience*, 8: 43.

Syed, N. A.; Liu, H.; and Sung, K. K. 1999. Handling Concept Drifts in Incremental Learning with Support Vector Machines. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, 317–321. New York, NY, USA: Association for Computing Machinery. ISBN 1581131437.

Wang, W.; Zheng, V. W.; Yu, H.; and Miao, C. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.*, 53(3).

Yan, S.; Xie, J.; and He, X. 2021a. DER: Dynamically Expandable Representation for Class Incremental Learning. arXiv:2103.16788.

Yan, S.; Xie, J.; and He, X. 2021b. DER: Dynamically Expandable Representation for Class Incremental Learning. *CoRR*, abs/2103.16788.

Yi, J.; Chen, Y.; Li, J.; Sett, S.; and Yan, T. 2013. Predictive Model Performance: Offline and Online Evaluations. In *ACM SIGKDD*.