EVALUATING AGENTS WITHOUT REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has enabled agents to solve challenging control tasks from raw image inputs. However, manually crafting reward functions can be time consuming, expensive, and prone to human error. Competing objectives have been proposed for agents to learn without external supervision, such as artificial curiosity, information gain, and empowerment. Estimating these objectives can be challenging and it remains unclear how well they reflect task rewards or human behavior. We study these objectives across seven agents and three Atari games. Retrospectively computing the objectives from the agent's lifetime of experience simplifies accurate estimation. We find that all three objectives correlate more strongly with human behavior than with the task reward. Moreover, task reward with curiosity better explains human behavior than task reward alone.

1 INTRODUCTION

Deep reinforcement learning (RL) has enable agents to solve complex tasks directly from high-dimensional image inputs, such as locomotion (Heess et al., 2017), robotic manipulation (Akkaya et al., 2019), and game playing (Mnih et al., 2015; Silver et al., 2017). However, many of these successes are built upon rich supervision in the form of manually defined reward functions. Unfortunately, designing informative reward functions is often expensive, time-consuming, and prone to human error (Krakovna et al., 2020). Furthermore, these difficulties increase with the complexity of the task of interest.

In contrast to many RL agents, natural agents generally learn without externally provided tasks, through intrinsic objectives. For example, children explore the world by crawling around and playing with objects they find. Inspired by this, the field of intrinsic motivation (Schmidhuber, 1991; Oudeyer et al., 2007) seeks mathematical objectives for RL agents that do not depend on a specific task and can be applicable to any unknown environment. We study three common types of intrinsic motivation:

- Curiosity encourages encountering rare sensory inputs, measured by a learned density model (Schmidhuber, 1990; Bellemare et al., 2016b; Pathak et al., 2017; Burda et al., 2018b).
- Empowerment measures the agent's influence it has over its sensory inputs or environment (Klyubin et al., 2005; Mohamed and Rezende, 2015; Karl et al., 2017).
- Information gain, or infogain for short, rewards the agent for discovering the rules of its environment (Lindley et al., 1956; Houthooft et al., 2016; Shyam et al., 2018; Sekar et al., 2020).

Metric	Reward Correlation
Task Reward	1.00
Human Similarity	0.60
Curiosity	0.35
Empowerment	0.30
Information Gain	0.28

Metric	Human Correlation
Task Reward	0.60
Human Similarity	1.00
Curiosity	0.86
Empowerment	0.57
Information Gain	0.68

Table 1: Correlation coefficients between each metric and task reward or human similarity. The 3 task-agnostic metrics correlate more strongly with human similarity than with task reward. This suggests that typical RL tasks may not be a sufficient proxy for intelligent behavior seen in humans playing the same games.

Despite the empirical success of intrinsic motivation for facilitating exploration (Bellemare et al., 2016b; Burda et al., 2018b; Ecoffet et al., 2019), it remains unclear which family of intrinsic objectives is best for a given scenario, for example when task rewards are sparse or unavailable, or when the goal is to behave similarly to human actors. Moreover, it is not clear whether different intrinsic objectives offer similar benefits in practice or are orthogonal and should be combined. To spur progress toward better understanding of intrinsic objectives, we empirically compare the three objective families in terms of their correlation with human behavior and with the task rewards of well-known Atari games.

The goal of this paper is to gain understanding rather than to propose a new intrinsic objective or exploration agent. Therefore, there is no need to estimate intrinsic objectives while the agents are learning, which often requires complicated approximations. Instead, we train several well-known RL agents on three Atari games and store their lifetime datasets of experience, resulting in a total of 2.1 billion time steps and about 9 terabytes of agent experience. From the dataset of each agent, we compute the human similarity, curiosity, empowerment, and infogain using simple estimators with clearly stated assumptions. We then analyze the correlations between these metrics to understand how they relate to another and how well they reflect task reward and human similarity.

The key findings of this paper are summarized as follows:

- Task reward does not fully capture correlations between RL agent behavior and human behavior; a combination of task reward and curiosity better accounts for human similarity. This suggests that common task rewards may not be a complete or optimal measure of agent intelligence.
- Simple implementations of curiosity, empowerment, and information gain correlate substantially with human similarity. This suggests that they can be used as task-agnostic evaluation metrics when human data and task rewards are unavailable.
- As a consequence of the these two findings, task-agnostic metrics can be used to measure a different component of agent behavior than is measured by the task rewards of the reinforcement learning environments considered in our study.
- Empowerment and information gain correlate strongly with each other, but to a lesser degree with curiosity. This suggests that optimizing curiosity together with either of the two other metrics could be beneficial for designing exploration methods.

This paper is structured as follows. Section 2 describes the games and agents used for the study. Section 3 details the experimental setup and estimators used to implement the metrics. Section 4 discusses quantitative and qualitative results. Section 5 summarizes key take-aways and recommendations.

2 Background

To validate the effectiveness of our metrics for task-agnostic evaluation across a wide spectrum of agent behavior, we retrospectively computed our metrics on the lifetime experience of well-known RL agents. Thus, we first collected datasets of a variety of agent behavior on which to compute and evaluate our metrics.

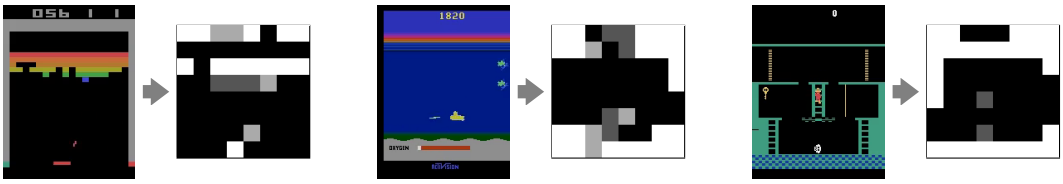
Environments We evaluated our agents in three different Atari environments provided by Atari Learning Environment (Bellemare et al., 2013): Breakout, Seaquest, and Montezuma's Revenge. Breakout and Seaquest are relatively simple reactive environments, while Montezuma's Revenge is a challenging platformer requiring long-term planning. We chose these three environments because they span a range of complexity, freedom, and difficulty, as detailed in Appendix C.

Agents The seven agent configurations represented in our dataset include three RL algorithms and two trivial agents for comparison. We selected RL agents spanning the range from extrinsic task reward only to intrinsic curiosity reward only. Additionally, we included random and no-op agents, two opposite extremes of trivial behavior. Our goal was to represent a wide range of behaviors: playing to achieve a high score, playing to explore the environment, and taking actions without regard to the environment. Specifically, we used the PPO agent (Schulman et al., 2017) trained to optimize task reward, and the RND (Burda et al., 2018b) and ICM (Pathak et al., 2017) exploration agents, which can incorporate both an intrinsic reward signal, and an extrinsic reward signal which can be enabled for task-specific behavior, or disabled for task-agnostic behavior. We evaluate RND and ICM in both of these configurations. Each agent is summarized in Appendix D.

3 Method

The goal of this paper is to evaluate agents using metrics other than task reward. For this, we collected 100 million frames on each of the three Atari environments with each of seven agents: random, no-op, PPO, and RND and ICM with and without task reward.

To evaluate the agents, we first preprocessed the data and then computed our metrics in aggregate over the entire lifetime of each agent/environment configuration (yielding one number per metric-agent-environment). In addition, we computed each of our metrics for each individual episode and conduct



(a) Breakout (b) Seaquest (c) Montezuma

Figure 1: Preprocessing used to assign input images to buckets. Similar to Go-Explore (Ecoffet et al., 2019), we resize the images to 8×8 pixels and discretize each of the resulting cells to one of 4 values. The examples show that this procedure preserves positions of objects in the game, such as the player, ball, fish, and skull. We enumerate the compactified images to represent each unique frame by an integer index to compute discrete probability tensors for the environments.

a preliminary per-episode analysis (Appendix E). In this section, we describe our preprocessing method and introduce our estimators for the five considered metrics.

3.1 Preprocessing

To efficiently compute the metrics in a robust manner, we discretize the agent's input images so that they can be represented by integer indices. This allows us to summarize each collected datasets as a probability tensor that holds the probability of each transition. Figure 1 visualizes the preprocessing for the three environments.

Discretization We first convert the RGB images to grayscale as they were seen by the agents. After that, we bilinearly resize them to 8×8 pixels. We discretize these low-resolution images to four possible values per pixel, with thresholds chosen as the brightness percentiles 25, 50, and 75 across all unique values of the corresponding pixel in the agent-environment configuration.

Aggregation The unique compactified images are enumerated and summarized into a tensor of counts for each agent and environment combination. For an image index $1 \leq i \leq |X|$, and action index $1 \leq j \leq |A|$, and a successor image index $1 \leq k \leq |X|$, the count tensor is defined as,

$$N_{ijk} \doteq \text{number of transitions from image bucket } i \text{ and action } j \text{ to image bucket } k. \tag{1}$$

Normalizing the count tensor N yields a probability tensor P that stores the probability of each transition in the agent's dataset. Under the assumption of a Markovian environment and agent, the probability tensor fully describes the statistics of the preprocessed dataset,

$$P \doteq \frac{1}{\sum_{ijk} N_{ijk}} N, \quad \text{so that} \quad \sum_{ijk} P_{ijk} = 1. \tag{2}$$

The probability tensor P describes the joint probability of transitions for each agent and environment and thus allows us to compute any marginals and conditionals needed to compute the metrics.

3.2 Metrics

We compare two task specific metrics, task reward and human similarity, as well as three task-agnostic metrics, curiosity, empowerment, and information gain. The task-agnostic metrics cover the different types of objectives identified by Hafner et al. (2020). We compute a single value of each of these metrics on each agent-environment dataset.

Task Reward The reward provided by reinforcement learning environments measures success at a specific task. The environments we use have only one predefined task each, despite the wide range of conceivable objectives in Montezuma's Revenge in particular. This is true of many RL environments, and limits one's ability to analyze the behavior of an agent in a general sense within one environment. While there are multi-task benchmarks, they often include a distinct environment for each task rather than multiple tasks in the same environment (Yu et al., 2019). This would make it difficult to evaluate the agent's ability to globally explore its environment independent of the task.

Human Similarity Task reward captures only the agent's success at the specific task defined via the reward function. This may not match up with a human observer's definition of intelligence. We suggest that a more general measure of intelligence may relate to similarity between the agent's behavior and human behavior in the same environment, i.e. using human behavior as a "ground truth". Hence, we propose a human similarity metric that approximates the overlap between the inputs observed by a human player and an RL agent. To compute this, we used the Atari-HEAD dataset (Zhang et al., 2019) and preprocessed the data as described in Section 3.1.

We compute human similarity as the Jaccard index, or intersection over union, between input images encountered in the human dataset and those encountered by the artificial agent. For this, we first compute the empirical visitation probabilities of inputs from the probability tensors P^{agent} and P^{human} of the artificial agent and the human player, respectively. The human similarity is then computed as the fraction of non-zero entries,

$$S \doteq \frac{\big|\big\{i : p(i) > 0 \wedge q(i) > 0\big\}\big|}{\big|\big\{i : p(i) > 0 \vee q(i) > 0\big\}\big|}, \quad \text{where} \quad p(i) \doteq \sum_{jk} P^{\text{agent}}_{ijk}, \quad q(i) \doteq \sum_{jk} P^{\text{human}}_{ijk}. \tag{3}$$

Note that while we use input images from recorded human behavior to compute human similarity, we are not able to compare the human and agent behavior directly, as the RL agents play in an environment with sticky actions, while the humans did not. More generally, human similarity would be challenging to compute in some environments, such as high-dimensional continuous control. Even where it is possible, large human datasets are expensive to collect. Thus, we consider three task-agnostic metrics, which do not require environment-specific engineering or human demonstrators.

Curiosity Curiosity measures how improbable the agent's inputs are under a trained density model (Schmidhuber, 1991). It has been used in RL to model an agent's success at exploring its environment, where a higher curiosity score implies a wider variety of states observed, for instance by Oudeyer et al. (2007); Bellemare et al. (2016a); Burda et al. (2018b).

In general, curiosity is the cross entropy of future inputs under a density model trained alongside the agent, but as we retrospectively compute the metric over the agent's lifetime, we only need to compute the probability vector over inputs once. Specifically, we use the marginal entropy over individual inputs, computed by summing the probability tensor over the second and third axes as,

$$C \doteq -\sum_{i} p(i) \log p(i), \quad \text{where} \quad p(i) \doteq \sum_{jk} P^{\text{agent}}_{ijk}. \tag{4}$$

Empowerment Empowerment measures the agent's influence over its environment (Klyubin et al., 2005), which has been applied to reinforcement learning by Mohamed and Rezende (2015); Karl et al. (2017). We use the interpretation of empowerment that measures the agent's realized influence on the world, rather than its potential influence, formalized as the mutual information between actions and sensory inputs (Salge et al., 2014; Hafner et al., 2020).

We compute empowerment mutual information as the difference between the entropy of actions given the preceding input, before and after observing the following input,

$$\begin{aligned} E &\doteq \bigg(\sum_{ijk} p(i,j,k) \log p(i,j,k)\bigg) - \bigg(\sum_{ij} q(i,j) \log q(i,j)\bigg), \\ \text{where} \quad p(i,j,k) &\doteq \frac{P_{ijk}}{\sum_{j'} P_{ij'k}}, \quad q(i,j) \doteq \frac{\sum_k P_{ijk}}{\sum_{j'k} P_{ij'k}}. \end{aligned} \tag{5}$$

Information gain Information gain is a measure of how much the agent learns from its observations Lindley et al. (1956). It is the mutual information between observations and the agent's representation of the environment. Information gain has led to successful exploration in reinforcement learning (Sun et al., 2011; Houthooft et al., 2016; Shyam et al., 2018).

To measure the amount of information gained, we need a way to represent the agent's knowledge about its environment. Preprocessing the agent's inputs into discrete classes enables us to represent its knowledge as a belief over the transition matrix M. The total information gain of over agent's lifetime is the entropy difference of its beliefs at the beginning and end of the dataset,

$$I \doteq \mathrm{H}\big[M\big] - \mathrm{H}\big[M \mid \text{dataset}\big] = \mathrm{E}\big[\log p(M \mid \text{dataset}) - \log p(M)\big]. \tag{6}$$

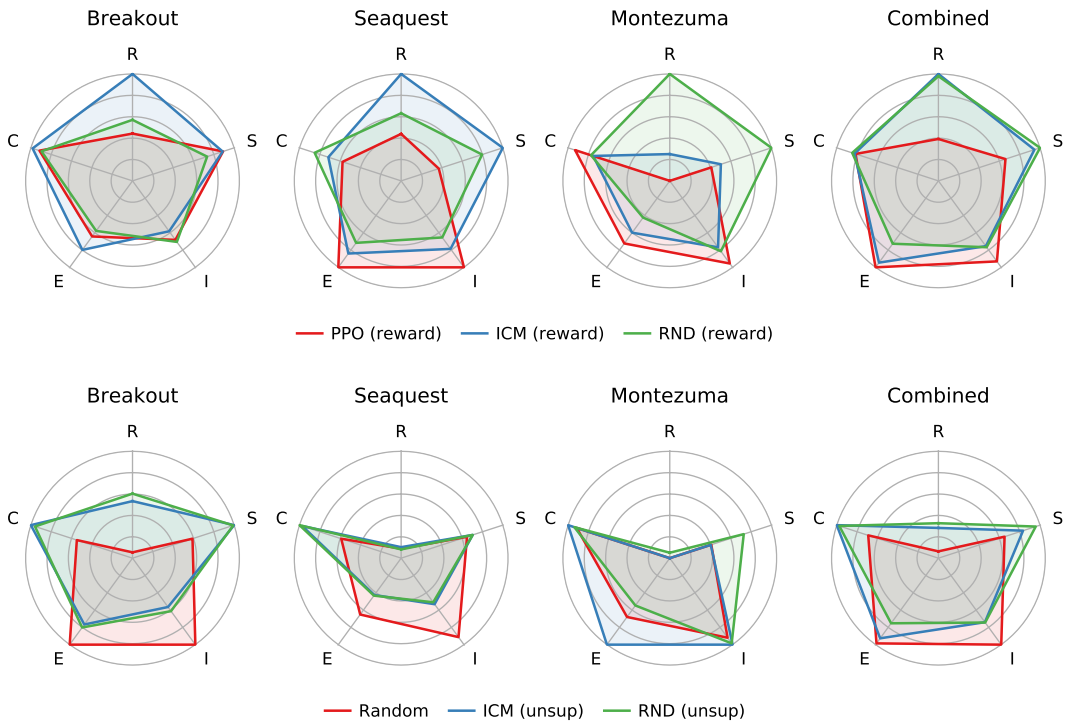


Figure 2: Overview of agent performance under the five metrics. Task Reward (R) and Human Similarity (S) are supervised metrics. Curiosity (C), Empowerment (E), and Information Gain (I) are task-agnostic metrics. The two task-specific exploration agents achieve the highest task reward and human similarity on average across environments. Task-agnostic ICM and RND achieve the highest curiosity value according to our metrics. More unexpectedly, we find that PPO and task-agnostic ICM achieve high empowerment, even in Montezuma where they achieve low task reward. The no-op agent achieves the lowest scores in all metrics does not show up in the normalized coordinates.

We choose the belief to be a vector of Dirichlet distributions as in (Sun et al., 2011; Friston et al., 2017), each of which has one concentration parameter, initialized at 1, for each possible subsequent input. For each transition in the dataset, the corresponding concentration parameter is incremented by 1. The entropy of a Dirichlet distribution is given by,

$$\mathrm{H}\big[D\big] = \log \mathrm{B}(\alpha) + \left(\sum_{k=1}^{|X|} \alpha - |X|\right)\psi\left(\sum_{k=1}^{|X|} \alpha\right) - \sum_{k=1}^{|X|}(\alpha_k - 1)\psi(\alpha_k), \tag{7}$$

where α is the vector of concentration parameters, ψ is the digamma function, and B is the incomplete beta function (Lin, 2016). The entropy of the distribution over M is then given by adding up the entropy of each of the Dirichlet distributions.

4 Analysis

We conduct a wide range of analyses to understand how the three task-agnostic metric relate to another and to the supervised metrics of task reward and human similarity. We first compare the agents included in our RL datasets based on their values of the two supervised and three task-agnostic metrics we consider. Next, we analyze correlations between our task-agnostic and supervised metrics. Finally, we discuss correlations between the three task-agnostic metrics themselves. Figure 3 shows correlation matrices of the five metrics, Table 2 contains the tables of all computed metric values, and further visualizations can be found in the supplementary material. We use the OpenAI implementations of ICM (Burda et al., 2018a) and RND (Burda et al., 2018b) and the Stable Baselines implementation of PPO (Hill et al., 2018). We will release the source code for replicating our analyses and the collected datasets and metrics upon publication to spur further work on unsupervised evaluation of RL agents.

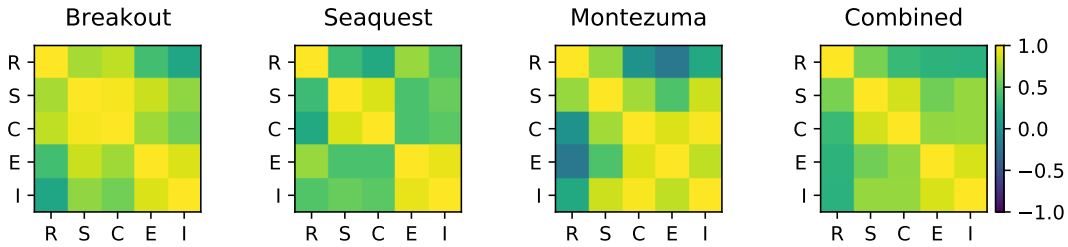


Figure 3: Pearson correlation coefficients between the five lifetime metrics considered in this study: Task Reward (R), Human Similarity (S), Curiosity (C), Empowerment (E), Information Gain (I). The metrics were computed for the lifetime of each agent and the correlation is taken across agents. Aggregated across environments, all metrics correlate positively. Human similarity correlates substantially with both task reward (0.60) and the task-agnostic metrics (0.86, 0.57, 0.68). Reward only correlates weakly with the task-agnostic metrics (0.35, 0.30, 0.28).

4.1 Evaluation of Agents

Task reward Comparing the mean episode score of task-specific RND and ICM in our agent datasets (Appendix B) with Taïga et al. (2020), we find similar performance of the two agents. Our agents perform better in Breakout, ICM has higher reward than RND in Seaquest which is the reverse in Taïga et al. (2020). We observe in Table 2 that task-specific RND or ICM achieves the highest task reward per time step in all environments, showing that agents benefit from explicit exploration objectives in all three environments.

Human similarity Human similarity is the highest for task-specific ICM and RND in Seaquest and Montezuma respectively, but for task-agnostic ICM in Breakout. This may be associated with the more reactive, less exploratory gameplay in Breakout. We qualitatively observe in Appendix E that in episodes selected for high human similarity, the agent makes progress in the game and tends to achieve a high task reward. We find that exploration agents achieve the highest human similarity in all three environments, as expected.

Curiosity We find that task-agnostic ICM and RND obtain the highest curiosity in all environments. These results may be related to the fact that ICM and RND maximize curiosity (Appendix D), and would suggest that using extrinsic reward "distracts" the task-specific agents from maximizing curiosity. The no-op agent is always the lowest, and in Breakout and Seaquest all RL agents achieve higher curiosity than the random agent. Individual episodes with high curiosity exhibit high task reward in Breakout and Seaquest (Appendix E).

Empowerment The agents exhibiting high and low empowerment vary more between environments than for curiosity. We observe that in all three environments, the random agent has relatively high empowerment; higher than all of the RL agents in Breakout (Table 2). This may be related to the fact that our simple preprocessing method is not semantically meaningful, i.e. it does not provide for generalization across similar but non-identical images. Thus, empowerment may not distinguish well between behavior in which the agent learns over time, and random behavior resulting in many subtly different episodes. This suggests that learning good input representations may key to exploration.

Information gain Information gain is highest for the random agent in Breakout, for PPO in Seaquest, and for the two configurations of RND in Montezuma. Like empowerment but unlike curiosity, it varies widely between environments.

4.2 Evaluation of Task-Agnostic Metrics

Correlation with task reward We find that task reward correlates relatively well (0.60) with human similarity. None of the task-agnostic metrics correlate strongly with reward, the highest being curiosity at 0.35. Considering that human similarity itself correlates well with reward, this suggests that curiosity and task reward capture different "components" of agent behavior. Providing further evidence for this, we normalized and added together task reward and curiosity and observed that this combined metric correlates more strongly with human similarity (0.89) than any individual metric.

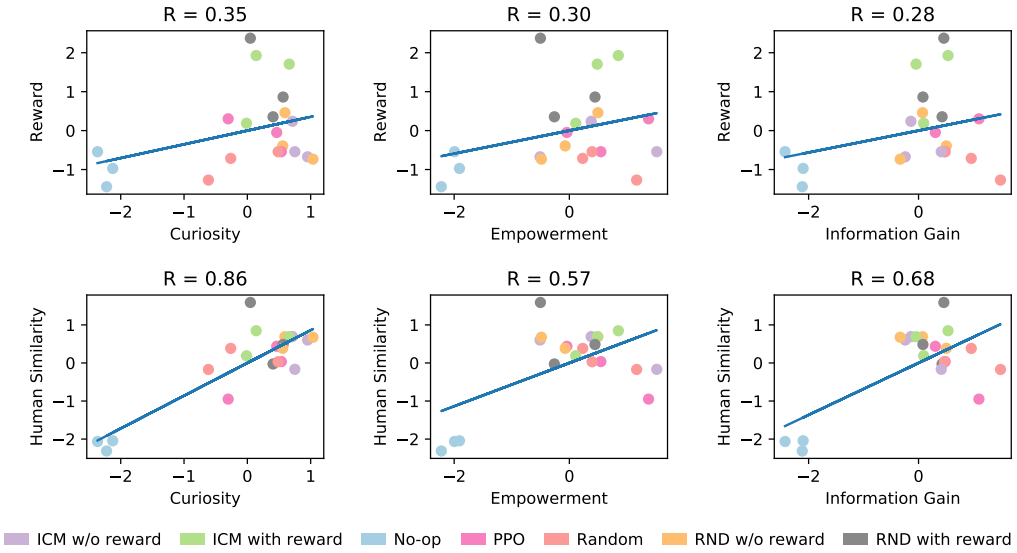


Figure 4: Scatter plots showing correlations between the three task-agnostic metrics (X axis) and the two task-specific metrics (Y axis), normalized for each environment separately. It is visually evident that the task-agnostic metrics show significant correlations with human similarity but are only weakly correlated with task reward. Additionally, note that the no-op configurations are clearly distinguished from the other agents along the curiosity, empowerment, infogain, and human similarity axes, while yielding similar task reward to the random agent and some RL agent configurations. This shows that curiosity, empowerment, and infogain capture aspects of behavior not captured by task reward, in particular distinguishing between the random and no-op agents that behave very qualitatively differently despite achieving similar task reward.

Correlation with human similarity Multiple metrics correlate well with human similarity, the strongest once again being curiosity at 0.86. Human similarity exhibits stronger correlations with the task-agnostic metrics we consider than does task reward. It is worth noting that curiosity correlates the most strongly with both task-specific metrics, as is visually evident in Figure 4. It would be worth considering correlations with human similarity on more open-ended environments in the future, where human players are not necessarily optimizing task reward. This would help determine whether the strong correlations observed with human similarity are a property of the task-agnostic metrics in general, or of the human behavior on Atari games specifically.

Comparison between environments In 17 out 18 cases in Figure 3, we find that the three task-agnostic metrics correlate positively with task reward and human similarity. The correlations are especially strong in Breakout and Seaquest. Curiosity correlates strongly with reward in Breakout (0.81). Similarly, both empowerment and information gain are correlated with reward in Seaquest (0.69 and 0.46 respectively). We attribute this to the fact that Breakout and Seaquest are reactive games that require fast paced action choices to achieve high task rewards. In Montezuma, the task-agnostic metrics correlate less strongly with reward; in fact, empowerment correlates negatively with reward (-0.20). We suggest this may be related to the open-endedness of the game, where the agent is free to take many different courses of action, some of which result in high empowerment, without obtaining any task reward.

Combining metrics To analyze whether a combination of task-agnostic metrics accounts better for task reward or human similarity than any individual metric, we sum up different task-agnostic metrics. No summation of two or three task-agnostic metrics correlates more strongly with human similarity or task reward than curiosity alone. Additionally, we find that a linear model of curiosity, empowerment, and infogain can predict task reward and human similarity with correlations of 0.36 and 0.89 respectively. These are only slightly better than the correlations with curiosity in both cases (0.35 and 0.86). This may be a reflection of the environments in our study, which feature equal amounts of stochasticity in all states. Empowerment and information gain are limited in overly

stochastic parts of the environment because they are lower when inputs are less predictable, which may not be necessary in our tasks.

4.3 Comparison among Task-Agnostic Metrics

Correlation among metrics In Figure 3, empowerment and information gain are shown to correlate strongly with one another (0.88). Curiosity also correlates positively with both of them, though less so (0.68 with empowerment and 0.68 with information gain). This suggests that curiosity explores in a different manner than empowerment and information gain. This finding could guide the design of future exploration methods by suggesting that, while combining metrics does not seem to be beneficial in our environments, curiosity and empowerment/information gain are distinct objectives and so combination of task-agnostic metrics could be applicable in other cases.

Qualitative comparison of individual episodes All previous results are based on our metrics calculated in aggregate over the entire lifetime of an agent, as in Table 2. However, we also qualitatively consider preliminary results regarding all five metrics calculated for each individual episode. Appendix E shows episodes with minimum and maximum values of each of our metrics.

5 Discussion

In this paper, we have collected large and diverse datasets of agent behavior, computed three task-agnostic metrics on each dataset, and analyzed the correlations of the task-agnostic objectives with task reward and with a human similarity metric. We have found that a simple probability tensor implementation of curiosity shows promise as a task-agnostic objective for RL agent evaluation, and that curiosity and task reward, when added together, correlate more strongly with human similarity than either does alone.

Reliability of results We are confident in proposing curiosity as a task-agnostic exploration metric, given that it correlates the most with task reward and human similarity. We are also confident in our human similarity metric as a supervised baseline. As the human data does not use sticky actions, we are limited to using a heuristic that does not rely on the action dynamics. Despite this, human similarity it correlates strongly with task reward as expected. While there are edge cases where individual episodes are qualitatively uninteresting but achieve high metric scores, shown in Appendix E, the metrics correlate strongly with human similarity aggregated over the agent lifetime.

Limitations and future work

- Our downscaling and discretization method is a simple and transparent preprocessing method, but may not be optimal. More semantically meaningful representations, potentially including deep learning embeddings, may have the potential to uncover additional correlations and/or increase the low overlap observed for our human similarity metric.
- The human dataset we used to is limited in quantity compared to agent data (roughly 250K frames per environment, versus 100M frames of agent data). This is a possible reason that our human similarity metric, overlap between agents and human inputs, is low; see Table 2. Access to more human data would be helpful for future work.
- When RL agents are trained to optimize extrinsic task reward, it is clear what task the agent is trying to accomplish. The same is not necessarily true of human data, especially in more open-ended games like Montezuma's Revenge, where players may choose to explore or pursue an objective other than that defined by the reward function. Use of a human similarity metric with respect to distinct human datasets with different tasks for the players could yield some insight on how closely the human concept of exploration aligns with the task in commonly used environments. An example of such a dataset is is the MineRL human dataset (Guss et al., 2019).

Summary of insights Task reward, while a useful measure of agent intelligence, may not be complete. We propose curiosity as a promising task-agnostic metric for agent evaluation, finding that it correlates strongly with human similarity, and that it measures a different component of agent behavior than task reward; a combination of task reward and curiosity accounts better for human similarity than either alone. We also find that empowerment and information gain correlate strongly with each other but to a lesser degree with curiosity, and thus recommend future research into combining curiosity with empowerment or information gain, in a variety of environments.

REFERENCES

I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Y. Aytar, T. Pfaff, D. Budden, T. Paine, Z. Wang, and N. de Freitas. Playing hard exploration games by watching youtube. In *Advances in Neural Information Processing Systems*, pages 2930–2941, 2018.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016a.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016b.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018a.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018b.

A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

K. J. Friston, M. Lin, C. D. Frith, G. Pezzulo, J. A. Hobson, and S. Ondobaka. Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683, 2017.

W. H. Guss, B. Houghton, N. Topin, P. Wang, C. Codel, M. Veloso, and R. Salakhutdinov. MineRL: A large-scale dataset of Minecraft demonstrations. *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. URL http://minerl.io.

D. Hafner, P. A. Ortega, J. Ba, T. Parr, K. Friston, and N. Heess. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*, 2020.

N. Heess, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami, M. Riedmiller, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu. Stable baselines. https://github.com/hill-a/stable-baselines, 2018.

R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

M. Karl, M. Soelch, P. Becker-Ehmck, D. Benbouzid, P. van der Smagt, and J. Bayer. Unsupervised real-time control through variational empowerment. *arXiv preprint arXiv:1710.05101*, 2017.

A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE, 2005.

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of ai ingenuity, 2020. URL https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

J. Lin. On the dirichlet distribution. Master's thesis, Queen's University, 2016. URL https://mast.queensu.ca/~communications/Papers/msc-jiayu-lin.pdf.

D. V. Lindley et al. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

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

S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 2125–2133, 2015.

P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

C. Salge, C. Glackin, and D. Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819, 2014.

J. Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. 1990.

J. Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pages 1458–1463. IEEE, 1991.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. *arXiv preprint arXiv:2005.05960*, 2020.

P. Shyam, W. Jaśkowski, and F. Gomez. Model-based active exploration. *arXiv preprint arXiv:1810.12162*, 2018.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354, 2017.

Y. Sun, F. Gomez, and J. Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*, pages 41–51. Springer, 2011.

A. A. Taïga, W. Fedus, M. C. Machado, A. C. Courville, and M. G. Bellemare. On bonus-based exploration methods in the arcade learning environment. 2020.

T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *arXiv preprint arXiv:1910.10897*, 2019.

R. Zhang, Z. Liu, L. Guan, L. Zhang, M. M. Hayhoe, and D. H. Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *ArXiv*, abs/1903.06754, 2019.

A LIFETIME METRICS

Breakout		Naive		Task-Agnostic		Task-Specific		
		No-op	Random	ICM	RND	PPO	ICM	RND
R	Task Reward	0.0000	0.0071	0.0695	0.0786	0.0576	**0.1302**	0.0743
S	Human Similarity	0.0000	0.0247	**0.0346**	0.0346	0.0317	0.0346	0.0263
C	Curiosity	0.0000	9.0696	**16.5545**	15.8827	15.1491	16.2690	14.8227
E	Empowerment	0.0000	**0.4844**	0.3716	0.3883	0.3114	0.3867	0.2804
I	Information Gain	0.0000	**0.8977**	0.4919	0.5447	0.6027	0.5154	0.6331

Seaquest		Naive		Task-Agnostic		Task-Specific		
		No-op	Random	ICM	RND	PPO	ICM	RND
R	Task Reward	0.0000	0.1546	0.1789	0.1424	0.7638	**1.7379**	1.0993
S	Human Similarity	0.0001	0.0031	0.0034	0.0035	0.0015	**0.0037**	0.0032
C	Curiosity	7.0172	12.9349	16.7889	**17.0601**	12.8069	14.2110	15.5612
E	Empowerment	0.0000	0.5822	0.3807	0.3871	**0.8934**	0.7507	0.6404
I	Information Gain	0.2024	0.6590	0.4798	0.4659	**0.6798**	0.5959	0.5278

Montezuma		Naive		Task-Agnostic		Task-Specific		
		No-op	Random	ICM	RND	PPO	ICM	RND
R	Task Reward	0.0000	0.0003	0.0000	0.2163	0.0003	1.0620	**4.2386**
S	Human Similarity	0.0001	0.0069	0.0063	0.0081	0.0070	0.0075	**0.0120**
C	Curiosity	3.1421	8.7690	**9.2900**	8.9128	8.8696	7.7834	7.9039
E	Empowerment	0.0000	0.2537	**0.3730**	0.2041	0.2699	0.2235	0.1584
I	Information Gain	0.0002	0.8977	0.8797	**0.9083**	0.9016	0.7809	0.8942

Table 2: Lifetime values of each metric for all agents and environments, with the highest value of each row in bold. In all three environments, the highest task reward is achieved by task-specific RND or ICM, which maximize both task reward and different implementations of curiosity; and the highest curiosity is achieved by task-agnostic RND or ICM, which is to be expected as these agents maximize curiosity alone. Agents with the highest human similarity, empowerment, and information gain vary by environment.

B EPISODE RETURNS

	Naive		Task-Agnostic		Task-Specific		
	No-op	Random	ICM	RND	PPO	ICM	RND
Breakout	0.0000	1.6513	92.9397	132.7004	105.6570	**240.3896**	147.6850
Seaquest	0.0000	98.9152	330.5743	349.1177	1713.8523	**4493.0141**	2700.6128
Montezuma	0.0000	0.6422	0.0130	489.0153	0.7339	238.7989	**5186.5680**

Table 3: For our analysis we normalized all metrics by the number of time steps in the agent's dataset. For comparison with prior work, this table shows the unnormalized episode returns of all agents. These verify that the agents were trained correctly.

C ENVIRONMENTS

We consider three Atari environments: Breakout, Seaquest, and Montezuma's Revenge. All three games are 2D environments with backgrounds fixed relative to the screen. The player is free to move around within the screen on one axis in Breakout, and two axes in the Seaquest and Montezuma. In Montezuma, the player can additionally navigate from one room into another.

Breakout Breakout is a game in which the agent controls a paddle at the bottom of the screen with the objective of bouncing a ball between the paddle and the blocks above, which disappear upon contact with the ball. The game is nearly deterministic: the only source of randomness other than sticky actions is the initial direction of the ball after starting the game or losing a life. Breakout is the simplest of our three environments, the player only being free to move a paddle in one dimension.

Seaquest Seaquest is a game in which the player controls a submarine, with the objective of defending oneself against sharks and other submarines which appear frequently and randomly at both sides of the screen. Additionally, the agent is tasked with picking up divers, which also appear at random, and bringing them to the top of the screen. Because of the random appearance of sharks and divers, the game can be difficult to predict, and made more so by sticky actions. It is more challenging that Breakout, as the player moves along two dimensions and enemies appear at random; however, the agent's task is reactive, with no long-term planning required.

Montezuma Montezuma's Revenge is a difficult platformer game with a large first level consisting of many rooms, which necessitates long-term planning. The player must navigate ladders, ropes, and various hazards such as moving skulls and lasers. Rewards are very sparse, and given only when the player completes an objective such as finding a key or opening a door, which often require complex and specific action sequences. For this reason, intrinsic rewards (Burda et al., 2018b) or human demonstrations (Aytar et al., 2018) are important to succeed at the game.

We follow the standard evaluation protocol (Machado et al., 2018). The games yield $210 \times 160 \times 3$ images, which are converted to grayscale and rescaled to 84×84 before being input to the agent. The agent chooses one of a set of discrete actions: 4 in Breakout, and 18 in Seaquest and Montezuma. While the effects of the actions in the original games are deterministic, we use "sticky actions", meaning that the agent's action is ignored with 25% chance and instead the previous action is repeated.

D AGENTS

Random An agent that uniformly samples random actions from the available action space.

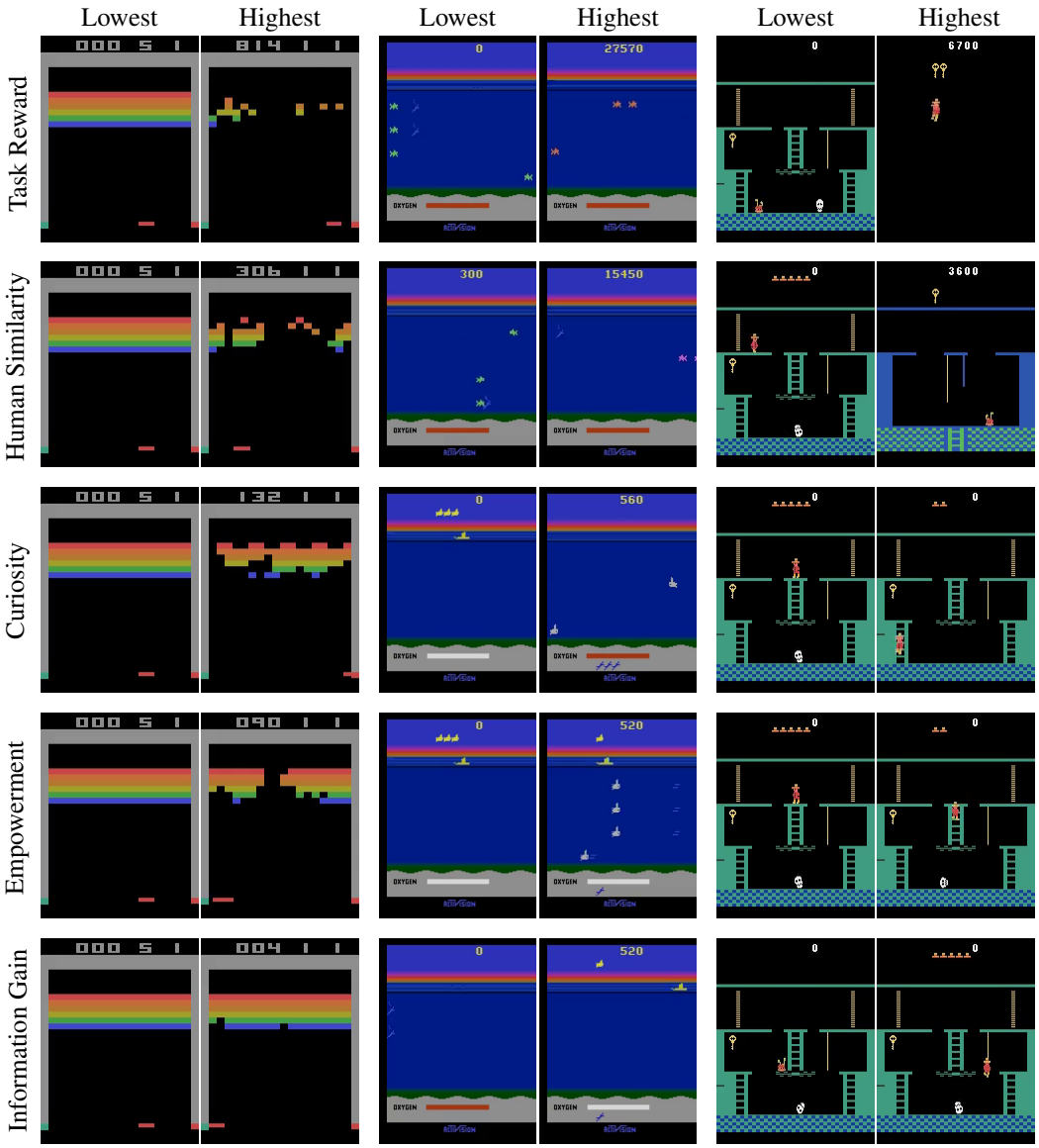
No-op The three environments we consider have a no-op action which does nothing, though the environment still continues to update. We consider an agent which always takes this no-op action.

PPO Proximal Policy Optimization (Schulman et al., 2017) trained on extrinsic rewards only. PPO is a commonly used policy gradient algorithm that optimizes the task reward on-policy. It optimizes task reward while preventing the policy from changing too much on each training step, so that the learning process is stable.

ICM Intrinsic Curiosity Module (Pathak et al., 2017) is an exploration agent that maximizes curiosity in addition to the task reward signal. Image embeddings are created that incorporate only those aspects of the image that can affect the agent, by training a network to predict the agent's action given the preceding and following input images. A second network is trained to predict an input image embedding given the preceding input and action. Its error is used as an intrinsic objective.

RND Random Network Distillation (Burda et al., 2018b) is an agent based on PPO that maximizes a curiosity objective, the prediction error of a model trained to distill a random function. It makes use of a randomly initialized and fixed neural network which predicts an embedding from the agent's input images. Another network is trained on the outputs of the random network, and its error, which models the agent's "familiarity" with the inputs in question, are used as an intrinsic objective.

E INDIVIDUAL EPISODES



(a) Breakout (b) Seaquest (c) Montezuma

Computing the evaluation metrics per-episode lets us visualize the episodes with lowest and highest values for each metric. The images show the last frame for each of these episodes. Note the task reward display in each image. The per-episode metrics differ from the per-dataset metrics used in the main text. This is most notable for the information gain, which when computed per-episode, does not account for incremental learning progress. Despite this, the the per-episode metrics favor episodes that are neither too predictable (no-op actions) nor too unpredictable (random actions), without requiring access to task rewards or human demonstrations. We further conclude that computing the metrics on the whole dataset leads to more robust evaluation scores.