Evaluating Agents without Rewards

Brendon Matusch **Jimmy Ba** [1] **Danijar Hafner** [1 2]

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 scale negatively 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).

Despite the empirical success of intrinsic motivation for facilitating exploration (Bellemare et al., 2016b; Burda et al.,

Metric	Reward Correlation
Task Reward	1.000
Human Similarity	0.696
Curiosity	0.353
Empowerment	0.296
Information Gain	0.235

Metric	Human Correlation
Human Similarity	1.000
Curiosity	0.792
Task Reward	0.696
Information Gain	0.590
Empowerment	0.525

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.

[1]University of Toronto [2]Google Brain. Correspondence to: Danijar Hafner <mail@danijar.com>.

Inductive Biases, Invariances and Generalization in RL (BIG) Workshop in the 37^{th} International Conference on Machine Learning, Vienna, Austria, 2020.

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 a human similarity metric that we have devised, 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 of experience as datasets, resulting in a total of 2.1 billion time steps. 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 an 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 RL environments we consider.
- 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.

2 Method

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 a dataset comprising 100 million timesteps of each of seven agents (Appendix B) training on each of three Atari environments: Breakout, Seaquest, and Montezuma's Revenge (Appendix A). These environments span a wide range of complexity, and provide a diverse sample of agent behavior on which to compute and evaluate our metrics.

To evaluate the agents, we first preprocessed the input images into categorical bins by downscaling them to 8×8 and discretizing to 2 bits per pixel, thresholding at quartiles over unique images for each agent individually (Appendix C). We then computed our metrics in aggregate over the entire lifetime of each agent/environment configuration, yielding one value per metric-agent-environment (Appendix E).

We computed our task-agnostic metrics from cooccurrence tensors generated from the preprocessed agent experience:

- Curiosity is computed as the marginal entropy of discretized inputs.
- Empowerment is computed as the difference between the entropy of actions given the preceding input, minus the entropy of actions given both the preceding and following inputs.
- Information gain is computed using a belief over possible transition matrices, modeled using a Dirichlet distribution over following inputs for each preceding input. It is the difference between the entropy of this belief at the beginning and end of the agent's lifetime.

For comparison, we consider mean task reward, and a human similarity metric calculated as the fraction of unique discretized input images found in the Atari-HEAD dataset of human experience (Zhang et al., 2019) that also occur in the agent's experience. All metrics are detailed in Appendix D.

3 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 1 shows correlation matrices of the five metrics, and tables of all metric values and further visualizations can be found in the supplementary material.

3.1 Evaluation of Agents

Task reward Comparing the mean episode score of task-specific RND and ICM in our agent datasets (Appendix F) with Taïga et al. (2020), we find similar performance of the two agents. Our agents perform better in Breakout, and ICM

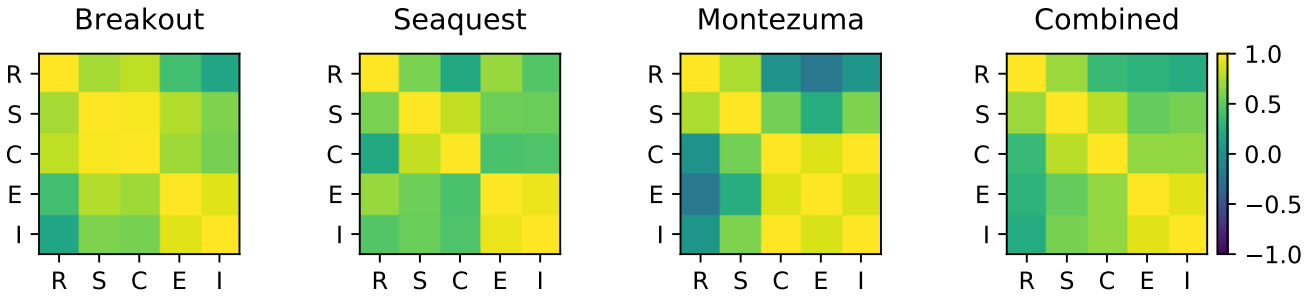

Figure 1: Pearson correlation coefficients between the five lifetime metrics considered in this study: Task Reward (R), Human Similarity (S), Curiosity (C), Empowerment (E), Information Gain (I). The metrics were computed for the lifetime of each agent and the correlation is taken across agents. Aggregated across environments, all metrics correlate positively. Human similarity correlates substantially with both task reward (0.696) and the task-agnostic metrics (0.792, 0.525, 0.590). Reward only correlates weakly with the task-agnostic metrics (0.353, 0.296, 0.235).

has higher reward than RND in Seaquest which is reversed in Taïga et al. (2020). We observe in Appendix E that task-specific RND or ICM achieves the highest task reward per time step in all environments, showing that agents benefit from exploration objectives in all three environments.

Human similarity We find that human similarity, more so than any task-agnostic metric, correlates strongly (0.696) with task reward. Additionally, we qualitatively observe in Appendix G that in episodes selected for high human similarity, the agent succeeds at 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. However, we observe low absolute values of human similarity in general, the highest being 0.0549.

Curiosity Curiosity correlates the most strongly with human similarity out of the three task-agnostic metrics by a significant margin (0.792, Table 1), and also has the greatest correlation with reward. The ICM and RND agents maximize curiosity (Appendix B). We find that task-agnostic ICM and RND obtain the highest curiosity in all environments. This would suggest that using extrinsic reward "distracts" the task-specific agents from maximizing curiosity.

Empowerment Empowerment correlates the most weakly with human similarity of the three task-agnostic metrics (0.525), and the correlation varies widely between environments (between 0.247 and 0.771). We observe that in all three environments, the random agent has relatively high empowerment; higher than all of the RL agents in Breakout (Appendix E). 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 be key to exploration.

Information gain Information gain correlates gain with human similarity slightly more strongly than empowerment (0.590). It shares the property of giving a high score to the random agent. In contrast to empowerment, correlations between information gain and human similarity vary less between environments; they are all weak positive correlations between 0.548 and 0.612.

3.2 Evaluation of Supervised Metrics

Correlation with task reward No metric other than human similarity correlates strongly with reward (the highest being curiosity once again, at 0.353). Considering that human similarity itself correlates well with reward, this suggests that curiosity and task reward capture different "components" of intelligence. 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 (0.905) than any individual metric alone.

Correlation with human similarity Multiple metrics correlate well with human similarity, the strongest being curiosity at 0.792. While human similarity exhibits stronger correlations in general with the metrics we consider, it is worth noting that the ranking of metrics by correlation is the same for reward as for human similarity: curiosity, empowerment, information gain. 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, to determine whether this is a property of the task-agnostic metrics themselves, or of the human behavior on Atari games specifically.

Differences between environments Curiosity correlates substantially with reward in Breakout (0.811), and both empowerment and information gain do likewise to a lesser degree in Seaquest (0.690 and 0.464 respectively). We attribute this to the fact that Breakout and Seaquest are reactive

environments, where taking appropriate and varied actions quickly in response to inputs is necessary to obtain task reward and avoid losing lives. In Montezuma, no metrics correlate well with reward; in fact, empowerment correlates negatively with reward (-0.199). We suggest this may be related to the open-endedness of Montezuma's Revenge, 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 We find that curiosity and empowerment normalized and added together correlates very slightly more strongly with task reward than curiosity alone (0.354 versus 0.353); but other than that, no addition of two or three task-agnostic metrics correlates more strongly with human similarity or task reward than curiosity alone. 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.

3.3 Comparison of Task-Agnostic Metrics

Correlation among metrics In Figure 1, empowerment and information gain are shown to correlate strongly with one another (0.907). Curiosity also correlates positively with both of them, though less so (0.677 with empowerment and 0.679 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 metrics calculated over the entire lifetime of an agent. However, we also qualitatively consider preliminary results regarding all five metrics calculated for each individual episode. Appendix G shows episodes with minimum and maximum values of each of our metrics.

4 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 a combination of curiosity and task reward correlates more strongly with human behavior than either alone. This work does not present any foreseeable direct societal consequence.

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. While it is a simple heuristic, 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 G, the metrics correlate strongly with human similarity aggregated over the agent lifetime.

Limitations and future work

- We consider only three Atari environments, which share many things in common (2D format, fixed surroundings on screen, etc.). It is possible that our downscaling and discretization method (Appendix C) would behave differently on environments without these properties.
- The human dataset we used is limited in quantity (~250K frames per environment). This is a possible reason that our human similarity metric, overlap of human inputs with agent inputs, is low. Access to more human data would be helpful for 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 human similarity overlap.
- 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 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.

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.

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-ingenuit

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.

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.