Evaluating Agents without Rewards

Brendon Matusch [1] **Jimmy Ba** [1 2] **Danijar Hafner** [1 2 3]

Abstract

Reinforcement learning has enabled agents to solve challenging tasks in unknown environments. However, manually crafting reward functions can be time consuming, expensive, and error prone to human error. Competing objectives have been proposed for agents to learn without external supervision, but it has been unclear how well they reflect task rewards or human behavior. To accelerate the development of intrinsic objectives, we retrospectively compute potential objectives on pre-collected datasets of agent behavior, rather than optimizing them online, and compare them by analyzing their correlations. We study input entropy, information gain, and empowerment across seven agents, three Atari games, and the 3D game Minecraft. We find that all three intrinsic objectives correlate more strongly with a human behavior similarity metric than with task reward. Moreover, input entropy and information gain correlate more strongly with human similarity than task reward does, suggesting the use of intrinsic objectives for designing agents that behave similarly to human players.

1 Introduction

Reinforcement learning (RL) has enabled 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 frequently learn without externally provided tasks, through intrinsic objectives. For example, children explore the world by crawling around and playing with objects they find. Inspired

[1]Vector Institute [2]University of Toronto [3]Google Brain. Correspondence to: Danijar Hafner <mail@danijar.com>.

Objective	Reward Correlation	Human Similarity Correlation
Task Reward	**1.00**	0.67
Human Similarity	0.67	**1.00**
Input Entropy	0.54	**0.89**
Information Gain	0.49	**0.79**
Empowerment	0.41	**0.66**

Table 1: Pearson correlation coefficients between each intrinsic objective and task reward or human similarity. The three intrinsic objectives correlate more strongly with human similarity than with task reward. This suggests that typical RL tasks may not be a sufficient proxy for the behavior that is seen in humans playing the same games.

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:

- Input entropy encourages encountering rare sensory inputs, measured by a learned density model (Schmidhuber, 1990; Bellemare et al., 2016b; Pathak et al., 2017; Burda et al., 2018).

- Information gain 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).

- Empowerment rewards the agent for maximizing the influence it has over its sensory inputs or environment (Klyubin et al., 2005; Mohamed and Rezende, 2015; Karl et al., 2017).

Despite the empirical success of intrinsic motivation for facilitating exploration (Bellemare et al., 2016b; Burda et al., 2018), 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 players. Designing intrinsic objectives that result in intelligent behavior across different environments is an important unsolved problem. Moreover, it is not clear whether different intrinsic objectives offer
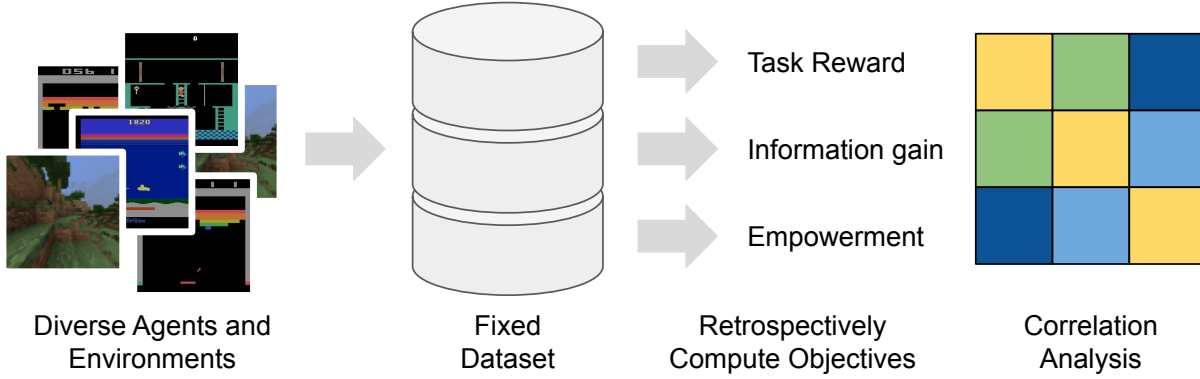
Figure 1: Training agents to evaluate different intrinsic objectives can be a slow and expensive process. To address this problem, we collect a diverse dataset of different environments and behaviors once and retrospectively compute agent objectives from it. We then analyze the correlations between intrinsic objectives and supervised objectives, such as task reward and human similarity. This speeds up the iteration time by letting us draw conclusions on the relationships between different intrinsic objectives without having to train a new agent for each of them.

similar benefits in practice or are orthogonal and should be combined. Progress toward answering these questions is hindered by slow iteration speeds because for each new intrinsic objective, one typically needs to design and train a new agent in one or more environments.

We propose the methodology of evaluating and comparing intrinsic objectives by correlation analysis on a fixed dataset, as shown in Figure 1. This frees us from having to train a separate agent for every objective considered, and alleviates the complexity associated with optimizing objectives online. To implement this methodology, we collect a diverse dataset of 26 agents in four complex environments to compare task reward, similarity to human players, as well as three representative intrinsic objectives. The key findings of the experimental study are summarized as follows:

- Input entropy and information gain correlate more strongly with human similarity than task reward does. The intrinsic objectives also correlate strongly with human similarity across all studied environments, while reward does not. To develop agents that behave similarly to human players, intrinsic objectives may thus be more relevant than typical task rewards.

- Simple implementations of input entropy, information gain, and empowerment based on image discretization correlate strongly with human similarity. Thus, they could lead to effective exploration when optimized online and could serve as evaluation metrics when task rewards and demonstrations are unavailable.

- Input entropy and information gain correlate strongly with each other, but to a lesser degree with empowerment. This suggests that optimizing empowerment together with either of the two other objectives could be beneficial for designing exploration methods.

2 Background

To validate the effectiveness of our intrinsic objectives across a wide spectrum of agent behavior, we retrospectively computed our objectives 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 compare our objectives.

Environments We chose three different Atari environments provided by Arcade Learning Environment (Bellemare et al., 2013): Breakout, Seaquest, and Montezuma's Revenge, and additionally the Minecraft Treechop environment provided by MineRL (Guss et al., 2019). Breakout and Seaquest are relatively simple reactive environments, while Montezuma is a challenging platformer requiring long-term planning. Treechop is a 3D environment in which the agent receives reward for breaking and collecting wood blocks, but has considerable freedom to explore the world. As detailed in Appendix F, we chose these four environments because they span a range of complexity, freedom, and difficulty.

Agents The 7 agents represented in our dataset include 3 learning algorithms and 2 trivial agents for comparison. We selected RL agents spanning the range from extrinsic task reward to intrinsic motivation reward. Additionally, we included random and no-op agents, two opposite extremes of naive 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., 2018) and ICM (Pathak et al., 2017) exploration agents that use PPO for policy optimization. We train the exploration agents once using only the intrinsic reward and once using both intrinsic and task rewards. The agents are described further in Appendix G.

2

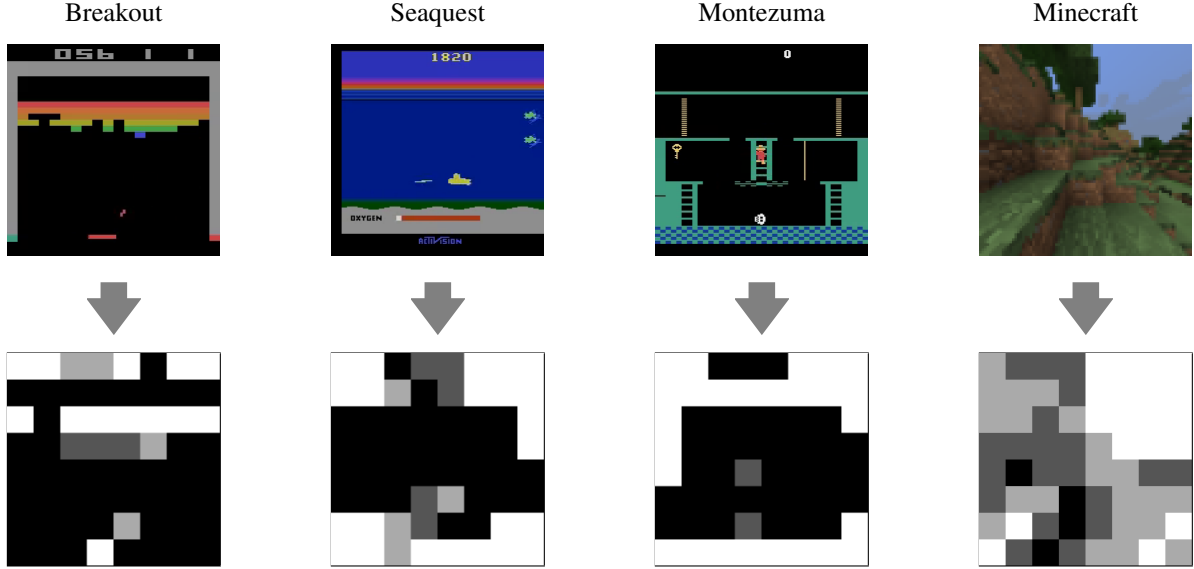Breakout   Seaquest   Montezuma   Minecraft

Figure 2: To make computing the agent objectives tractable and efficient, we preprocess the images in our collected datasets by discretizing them into 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 discretized images to represent each unique frame by an integer index to compute discrete probability tensors for the environments.

3 Method

To spur progress toward better understanding of intrinsic objectives, we empirically compared the three objective families in terms of their correlation with human behavior and with the task rewards of three Atari games and Minecraft. First, we trained several well-known RL agents on three Atari games and Minecraft and store their lifetime datasets of experience. Specifically, 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. Minecraft was evaluated for 12 million frames per agent because the simulation is slower than the Atari games, and five agents rather than seven were used, excluding both configurations of ICM. This resulted in a total of 2.1 billion time steps and about 9 terabytes of agent experience.

We preprocessed the collected experience datasets and computed human similarity, input entropy, empowerment, and information gain using simple estimators with clearly stated assumptions, in aggregate over each agent lifetime, yielding one number per objective-agent-environment. A table of all computed values is included in Table A.1. We then analyzed the correlations between the intrinsic objectives to understand how they relate to another and how well they reflect task reward and human similarity. We now describe the preprocessing and introduce our estimators for the objectives.

3.1 Preprocessing

To make the computation of the considered objectives tractable and efficient, we discretize the agent's input images so that they can be represented by bucket indices. This allows us to summarize each collected dataset as a sparse tensor that holds the counts of each possible transition tuple, from which we then compute the objective values.

Discretization As shown in Figure 2, we preprocess the images by first converting them from RGB 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 environment across all agents. We also considered choosing the thresholds based on each agent individually, as discussed in Appendix D. The unique discretized images are enumerated to represent each image by an integer index.

Aggregation For each pairing of agent and environment, we summarize the transitions preprocessed images and actions into a tensor of counts. For an image index $1 \leq i \leq |X|$, and action index $1 \leq j \leq |A|$, and a successor image index $1 \leq k \leq |X|$, where X is the set of inputs and A the set of actions, the count tensor N_{ijk} is defined as the number of transitions from image bucket i and action j to image bucket k.

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,

$$P \doteq N \Big/ \sum_{ijk} N_{ijk}. \qquad (1)$$

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 for computing the objectives.

3.2 Objectives

We compare two supervised objectives, task reward and human similarity, as well as three intrinsic objectives: input entropy, information gain, and empowerment. We compute a single value of each of these objectives on each agent-environment dataset.

Task reward The reward provided by RL environments measures success at a specific task. The environments we use have only one predefined task each, despite the wide range of conceivable objectives, especially in Montezuma's Revenge and Minecraft. This is true of many RL environments, and limits one's ability to analyze the behavior of an agent in a general sense. Multi-task benchmarks address this problem but often include a distinct environment for each task rather than multiple tasks within 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 the behavior of a human player who interacts with an environment. To capture this human-like aspect of behavior, we compute the similarity between the agent's behavior and human behavior in the same environment, that is, using human behavior as a "ground truth." Inspired by the inverse RL literature (Ziebart et al., 2008; Klein et al., 2012), we measure the overlap between human and agent observations. We leverage the Atari-HEAD (Zhang et al., 2019) and the MineRL Treechop (Guss et al., 2019) datasets and preprocess them the same way as the agent datasets. The Atari human datasets contain ~ 250K frames per environment, and ~ 460K frames are available for Minecraft.

We compute human similarity as the Jaccard index, also known as intersection over union, between the unique input images encountered in the human dataset and those encountered by the artificial agent. For this, we first compute the marginal input probabilities 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 a fraction of non-zero probability entries, where $\mathbb{1}(\cdot)$ denotes the indicator function that evaluates to 1 for true inputs and to 0 for false inputs,

$$\text{S} \doteq \sum_i \mathbb{1}(A_i > 0 \wedge B_i > 0) \Big/ \sum_i \mathbb{1}(A_i > 0 \vee B_i > 0)$$
$$A_i \doteq \sum_{jk} P_{ijk}^{\text{agent}}, \quad B_i \doteq \sum_{jk} P_{ijk}^{\text{human}}. \qquad (2)$$

Note that while we use input images from recorded human behavior to compute human similarity, we do not compare the human and agent actions directly, as the RL agents play in an environment with sticky actions, while the humans did not. Collecting human datasets is challenging in environments that are challenging for human players, such as high-dimensional continuous control. Thus, we consider three intrinsic objectives, which do not require environment-specific engineering or human demonstrators.

Input entropy The agent's input entropy in the environment measures how spread out its visitation distribution over inputs is. The input entropy tends to be larger the more inputs the agent has reached, and the more uniformly it visits them. In other words, input entropy measures how improbable individual inputs are under the input distribution of the agent. This idea has been used for exploration in RL, where the negative log-probability of inputs under a learned density model is used as exploration bonus (Schmidhuber, 1991; Oudeyer et al., 2007; Bellemare et al., 2016a; Burda et al., 2018). Because we compute the input entropy retrospectively, we have access to the agent's input distribution via the probability tensor that summarizes the agent's lifetime experience. We compute the the entropy over inputs x by marginalizing out actions and success inputs,

$$\text{C} \doteq \text{H}[x] = -\sum_i A_i \ln A_i, \quad A_i \doteq \sum_{jk} P_{ijk}. \qquad (3)$$

Empowerment Empowerment measures the agent's influence over its sensory inputs and thus the environment (Klyubin et al., 2005). It has been applied to RL by Mohamed and Rezende (2015); Salge et al. (2013); Karl et al. (2017); Leibfried et al. (2019); Zhao et al. (2020). Multiple variants exist in the literature, including potential empowerment where the agent aims to "have many options" and realized empowerment where the agent aims to "use many options." We consider the latter variant, that is measured as measured as the mutual information between the agents actions and resulting inputs, given current inputs (Salge et al., 2014; Hafner et al., 2020). We compute the one-step empowerment as the difference between the entropy of action a given the preceding input x, before and after observing
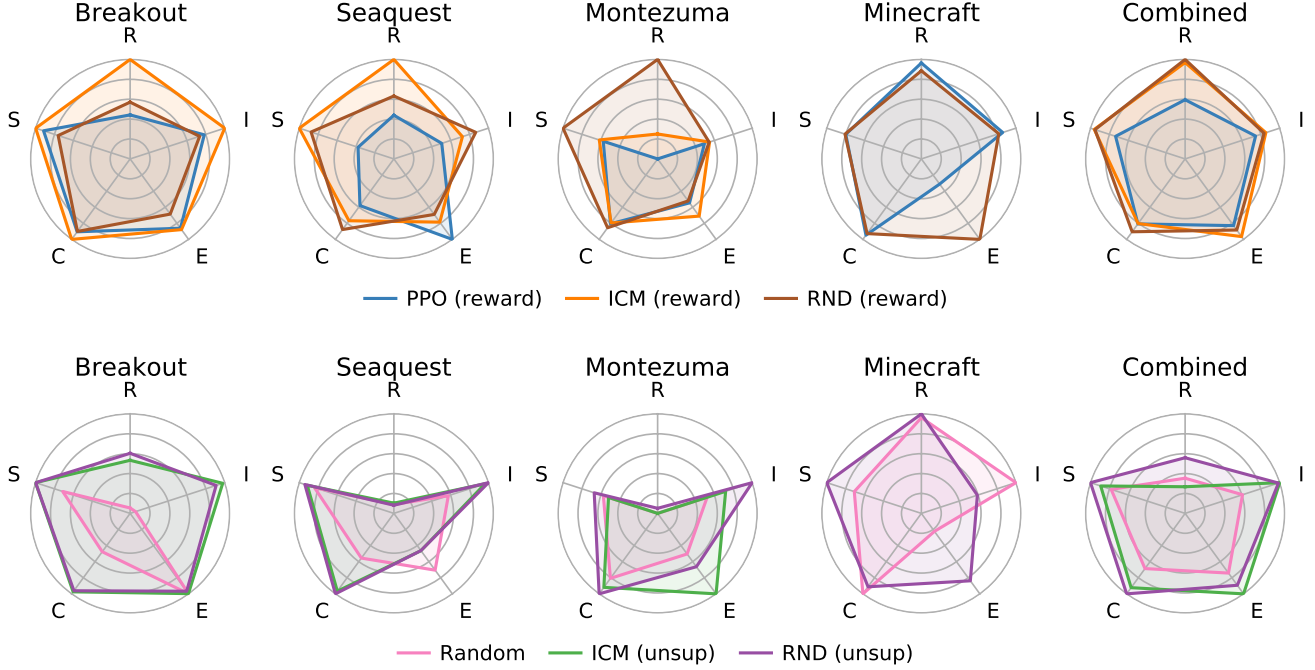
Figure 3: Objective values for agents that use task reward (top) and agents without access to task reward (bottom). The supervised objectives are task reward (R) and human similarity (S) and the intrinsic objectives are input entropy (C), empowerment (E), and information gain (I). The no-op agent achieves the lowest scores in all objectives does not show up in the normalized coordinates. The two exploration agents with access to task rewards achieve the highest task reward and human similarity across Atari environments, and RND without reward in Minecraft. ICM or RND each achieve the highest input entropy and information gain value in three out of four environments according to our objectives. Surprisingly, PPO and task-agnostic ICM achieve high empowerment, even in Montezuma where they achieve low task reward.

the following input x',

$$\mathrm{E} \doteq \mathrm{I}\big[x'; a \mid x\big] = \mathrm{H}\big[a \mid x\big] - \mathrm{H}\big[a \mid x, x'\big]$$

$$= \sum_{ijk} P_{ijk} \ln A_{ijk} - \sum_{ijk} P_{ijk} \ln B_{ij}, \qquad (4)$$

$$A_{ijk} = P_{ijk} \Big/ \sum_{i'k'} P_{i'jk'}, \quad B_{ij} = \sum_{k'} P_{ijk'} \Big/ \sum_{i'k'} P_{i'jk'}.$$

Information gain Information gain measures how many bits of information the agent has learned about the environment from its dataset of experience (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 RL (Sun et al., 2011; Houthooft et al., 2016; Shyam et al., 2018; Mirchev et al., 2018; Sekar et al., 2020).

To measure the amount of information gained, we need a way to choose a representation of the environment that summarizes the agent's knowledge. Preprocessing the inputs into discrete classes enables us to represent the knowledge as a belief over the true transition matrix. The total information gain of over agent's lifetime is the entropy difference of the prior belief over the transition matrix and the posterior

belief. We choose the beliefs to be Dirichlet distributions over successor inputs x' for every pairing of current inputs x and actions a (Sun et al., 2011; Friston et al., 2017),

$$\mathrm{I} \doteq \mathrm{I}\big[z; d\big] = \mathrm{H}\big[z\big] - \mathrm{H}\big[z \mid d\big]$$

$$= \sum_{ij} \mathrm{H}\big[\mathrm{Dir}(1)\big] - \sum_{ij} \mathrm{H}\big[\mathrm{Dir}(1 + N_{ij})\big]. \qquad (5)$$

The Dirichlet distribution is parameterized by a vector of counts, known as concentration parameter α. There is one Dirichlet distribution for each pairing of current input x and action a, with a concentration parameter of the length of possible successor inputs x'. We choose a uniform Dirichlet distribution as prior belief, corresponding to a vector of ones for its concentration parameter. The posterior belief is typically a Dirichlet belief that uses as concentration parameter the prior value of 1 plus the count vector N_{ij}.

Task Reward Correlation
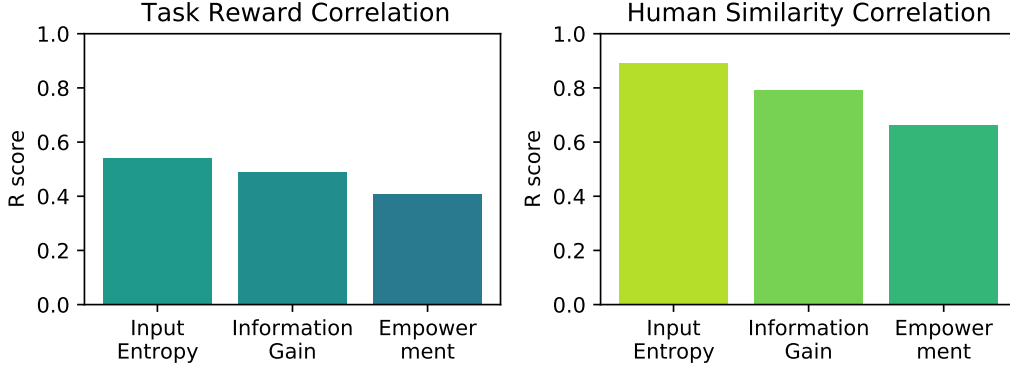
Human Similarity Correlation

Figure 4: Pearson correlation coefficients between the three considered intrinsic objectives and task reward (left) and similarity to human players (right). All considered intrinsic objectives correlate more strongly with human similarity than with the task rewards in the RL environments. This suggests that typical task rewards may be insufficient for evaluating exploration objectives when the goal is to produce behavior similar to human players. Interestingly, ranking the intrinsic objectives by their correlations with either task reward or human similarity gives the same ordering. The correlations were computed from $N = 26$ agent environment pairings and are statistically significant with $p < 0.05$.

The entropy of a Dirichlet distribution is computed from the vector of concentration parameters α using the digamma function $\psi(\cdot)$ the incomplete beta function $\mathrm{B}(\cdot)$ (Lin, 2016),

$$\mathrm{H}\big[\,\mathrm{Dir}(\alpha)\big] = \ln \mathrm{B}(\alpha) - \sum_k (\alpha_k - 1)\,\psi(\alpha_k)$$
$$+ \Big(\sum_k \alpha_k - |X|\Big)\,\psi\Big(\sum_k \alpha_k\Big). \quad (6)$$

In practice, we find it more effective to only consider unique transitions, and thus we use $1 + \mathrm{sign}(N_{ij})$ as the posterior concentration instead of the prior plus raw counts $1 + N_{ij}$. We conjecture that choice leads to a more meaningful information gain objective because the four environments are close to deterministic, with over 80% of input action pairs leading to a unique successor input in the datasets. Thus, experiencing a transition once establishes that the transition is possible, and revisiting it multiple times should not further increase the agent's knowledge. We compare alternatives to this choice in Appendix D.

4 Analysis

We conduct a correlation analysis to understand how the three intrinsic objectives relate to another and to the supervised objectives, task reward and human similarity. First, we compute all objectives for the agents and compare the agents by them. Second, we rank the intrinsic objectives by their correlations to task reward and human similarity. Third, we compare the correlations among the intrinsic objectives. We make the source code for replicating our analyses and the collected datasets available to support future work on evaluating agents without task rewards.[1]

4.1 Evaluation of Agents

We begin by computing and analyzing the objective values for individual agents, with the goal of intuitively understanding what specific agent behavior leads to high and low values of each objective. RND and ICM have the highest values of task reward and human similarity in all environments; that the agent with highest input entropy and is also the agent with highest information gain in all environments; and that empowerment varies widely by environment. The objective values are visualized in Figure 3 and their numbers are given in Appendix A.

Task reward The total episode rewards of the task-specific RND and ICM agents we trained are similar to those reported by Taiga et al. (2019). ICM achieve slightly higher reward than RND in Seaquest, while the opposite is true in Taiga et al. (2019). We observe in Appendix A that task-specific RND and ICM achieve the highest task reward per time step in the environments except in Minecraft, where task-agnostic RND performs best, showing that explicit exploration objectives are beneficial in all our considered 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 and task-agnostic RND in Minecraft. Exploration agents achieve the highest human similarity in all four environments. In environments other than Minecraft, the random agent has substantially lower human similarity than the other agents, and no-op is consistently the lowest across all environments.

Input entropy Task-agnostic ICM and RND obtain the highest input entropy in all environments, except in Minecraft where the random agent achieves the highest in-
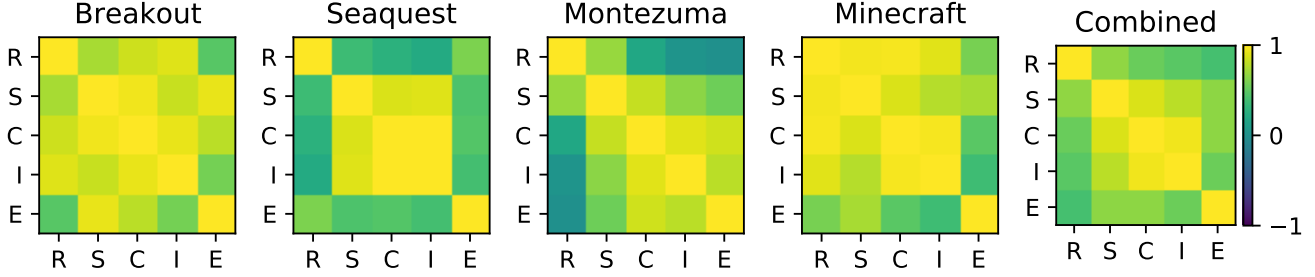
Figure 5: Correlation matrices for the objectives in our study: task reward (R), human similarity (S), input entropy (C), empowerment (E), information gain (I). The correlation is taken across agents. All objectives correlate positively but to different degrees. While human similarity and the three intrinsic objectives all correlate strongly with another, task reward only correlates weakly with them on Seaquest and Montezuma. As per-environment correlations are computed from 7 agents in Atari environments and 5 in Minecraft, they are not statistically significant and provided only for exploratory analysis. The numerical values for the correlation matrices are included in Table B.1.

put entropy. This suggests that using task reward "distracts" the task-specific agents from maximizing input entropy. The no-op agent has the lowest input entropy in all environments. The random agent achieves high input entropy in Minecraft, where many distinct inputs are easy to reach from the initial state, e.g. by moving the camera.

Information gain   Information gain is highest in Breakout for ICM with reward, in Seaquest and Montezuma for RND without reward, and in Minecraft for the random agent. We conjecture that the random agent would achieve lower information gain in Minecraft under a preprocessing scheme that groups semantically similar observations into the same bucket. In all four environments, the agent achieving the highest input entropy also achieves the highest information gain, implying that input entropy and information gain are closely related in practice.

Empowerment   The agents that have the most empowerment vary more across environments than they do for the other objectives. In Seaquest, where most meaningful actions are tied to reward, the PPO agent achieves highest empowerment. In Montezuma, where many actions influence the input in ways unrelated to the task, task-agnostic ICM achieves the highest empowerment. In Breakout, where almost all actions move the paddle and thus influence the input, the random agent achieves the highest empowerment.

4.2   Evaluation of Intrinsic Objectives

We evaluate the three intrinsic objectives based on their correlations with task reward and human similarity. The correlations aggregated across all environments are shown in Figure 4 and their numeric values are included in Table B.1. This constitutes our main result. All intrinsic objectives correlate more strongly with human similarity than with task reward. Moreover, ranking them by their correlations with either task reward or human similarity yields the same order. Finally, the intrinsic objectives correlate positively

with task reward and human similarity within individual environments in almost all cases.

By task reward and human similarity   All intrinsic objectives correlate substantially more with human similarity than with task reward. The correlations with human similarity are: input entropy (R=0.89), information gain (R=0.79), and empowerment (R=0.66). The correlations with task reward are: input entropy (R=0.54), information gain (R=0.49), and empowerment (R=0.41). This suggests that intrinsic objectives may capture more aspects of human behavior than typical task rewards, and may thus be the more promising approach when the goal is to design agents that behave similarly to humans.

Ranking of intrinsic objectives   Ranking the intrinsic objectives by their correlations with either task reward or human similarity yields in the same order. Input entropy correlates most strongly with task reward and human similarity, followed by information gain, and then empowerment. This suggests that exploration objectives such as input entropy and information gain are beneficial across the four environments, whereas task reward, which is not directly encouraging exploration, does not seem as important. Moreover, our results identify input entropy as a promising intrinsic objective that should work well across many scenarios.

Per-environment correlations   We additionally include per-environment correlations in Figure 5. Not only aggregated across environments, but also within each environment, the intrinsic objectives correlate positively with both task reward and human similarity in 23 out of 24 cases. The correlations between intrinsic objectives and human similarity are strong in all four environments (0.56≤R≤0.96), except for empowerment in Seaquest (R=0.44). In contrast, the correlations with task reward vary more across environments, being larger in Breakout (0.48≤R≤0.91) and Minecraft (0.59≤R≤0.98) but weaker in Seaquest (0.22≤R ≤0.61) and Montezuma (0.00≤R≤0.19).

4.3 Comparison among Objectives

The right-most correlation matrix in Figure 5 shows the correlations among all five objectives, which are statistically significant with $p < 0.05$. The numerical values are included in Table B.1. Task reward and human similarity correlate positively (R=0.67) but this correlation is weaker than those between human similarity and input entropy (R=0.89) or information gain (R=0.79). Input entropy and information gain correlate strongly (R=0.95), suggesting that they would have similar effects when optimized online. In contrast, empowerment correlates less strongly with the two objectives (R=0.66 and R=0.55). This suggests that empowerment measures a different component of agent behavior and combining it with either input entropy or information gain could be beneficial when designing novel exploration methods.

5 Discussion

In this paper, we have collected large and diverse datasets of agent behavior, computed three intrinsic objectives on each dataset, and analyzed the correlations of the intrinsic objectives with each other and with task reward and a human similarity objective. Our retrospective evaluation methodology enabled us to compare different intrinsic objectives while avoiding the complexity and typically slow iteration speed that come with online training.

Key findings   All studied intrinsic objectives correlate more strongly with human similarity than the task rewards do. These correlations hold consistently across all environments, whereas task reward and human similarity correlate strongly only in half of the environments. We thus recommend intrinsic objectives over task rewards when the goal is to design general agents that behave similarly to human players. Furthermore, we find that input entropy and information gain are similar objectives while empowerment may offer complementary benefits, and thus recommend future work on combining intrinsic objectives.

Future work   To assign the agent observations to buckets, we naively downscaled them, which is simple but does not account for the semantic similarity between images. We suggest learning the representations using deep neural networks as a direction for future work. A limitation of the human similarity values is that the human datasets are relatively small and it is unclear what instructions the human players received, which could affect how much the players focus on apparent tasks compared to open-ended behavior. Access to more human data and control over the instructions given to the players would be helpful for future work.

References

I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. Mc-Grew, 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, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

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.

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.

E. Klein, M. Geist, B. Piot, and O. Pietquin. Inverse reinforcement learning through structured classification. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1007–1015. Curran Associates, Inc., 2012.

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.

F. Leibfried, S. Pascual-Diaz, and J. Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. In *Advances in Neural Information Processing Systems*, pages 7869–7880, 2019.

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.

A. Mirchev, B. Kayalibay, M. Soelch, P. van der Smagt, and J. Bayer. Approximate bayesian inference in spatial environments. *arXiv preprint arXiv:1805.07206*, 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. Approximation of empowerment in the continuous domain. *Advances in Complex Systems*, 16(02n03):1250079, 2013.

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. Taiga, W. Fedus, M. C. Machado, A. Courville, and M. G. Bellemare. On bonus based exploration methods in the arcade learning environment. In *International Conference on Learning Representations*, 2019.

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.

R. Zhao, P. Abbeel, and S. Tiomkin. Efficient online estimation of empowerment for reinforcement learning. *arXiv preprint arXiv:2007.07356*, 2020.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.

A Objectives

Breakout

Symbol	Objective	No-op	Random	ICM_n	RND_n	PPO	ICM_r	RND_r
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	Input Entropy	0.0000	7.9303	16.2240	15.9964	15.0695	**16.6389**	14.9995
I	Information Gain	0.0000	0.0203	0.3458	0.3213	0.2770	**0.3533**	0.2596
E	Empowerment	0.0000	**0.4039**	0.4028	0.3907	0.3479	0.3548	0.2770

Seaquest

Symbol	Objective	No-op	Random	ICM_n	RND_n	PPO	ICM_r	RND_r
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	Input Entropy	6.7031	12.4245	16.6600	**17.0014**	12.6777	14.6161	15.7433
I	Information Gain	0.0000	0.2209	0.3700	**0.3835**	0.1941	0.2779	0.3319
E	Empowerment	0.0000	0.6110	0.4013	0.4008	**0.8661**	0.6806	0.5975

Montezuma

Symbol	Objective	No-op	Random	ICM_n	RND_n	PPO	ICM_r	RND_r
R	Task Reward	0.0000	0.0003	0.0000	0.2163	0.0003	1.0620	**4.2374**
S	Human Similarity	0.0001	0.0069	0.0063	0.0081	0.0070	0.0075	**0.0120**
C	Input Entropy	1.9008	7.1812	7.9123	**8.4405**	7.1757	7.1073	7.4942
I	Information Gain	0.0000	0.0120	0.0165	**0.0229**	0.0113	0.0126	0.0125
E	Empowerment	0.0000	0.1326	**0.2629**	0.1743	0.1436	0.1869	0.1373

Minecraft

Symbol	Objective	No-op	Random	ICM_n	RND_n	PPO	ICM_r	RND_r
R	Task Reward	0.0000	0.0012	—	**0.0013**	0.0012	—	0.0011
S	Human Similarity	0.000002	0.00003	—	**0.00004**	0.00003	—	0.00003
C	Input Entropy	9.4112	**16.2270**	—	14.8039	15.3758	—	15.0612
I	Information Gain	0.0004	**0.0583**	—	0.0345	0.0502	—	0.0477
E	Empowerment	0.0000	0.0770	—	0.2885	0.1077	—	**0.3444**

Table A.1: Lifetime values of each objective 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 input entropy; and the highest input entropy is achieved by task-agnostic RND or ICM, which is to be expected as these agents use a form of input entropy as their objective. Agents with the highest human similarity, empowerment, and information gain vary by environment. The agents achieve low reward in Minecraft because we only simulated 12M frames due to the slow simulator. We conjecture that the random agent achieves high input entropy and information gain in Minecraft because moving the camera randomly results in many visually different inputs.

B   Correlations

Breakout

	Task Reward	Human Similarity	Input Entropy	Information Gain	Empowerment
Task Reward	1.00	0.74	0.85	0.91	0.48
Human Similarity	0.74	1.00	0.96	0.83	0.93
Input Entropy	0.85	0.96	1.00	0.94	0.79
Information Gain	0.91	0.83	0.94	1.00	0.58
Empowerment	0.48	0.93	0.79	0.58	1.00

Seaquest

	Task Reward	Human Similarity	Input Entropy	Information Gain	Empowerment
Task Reward	1.00	0.37	0.28	0.22	0.61
Human Similarity	0.37	1.00	0.89	0.90	0.44
Input Entropy	0.28	0.89	1.00	1.00	0.46
Information Gain	0.22	0.90	1.00	1.00	0.40
Empowerment	0.61	0.44	0.46	0.40	1.00

Montezuma

	Task Reward	Human Similarity	Input Entropy	Information Gain	Empowerment
Task Reward	1.00	0.69	0.19	0.03	0.00
Human Similarity	0.69	1.00	0.83	0.65	0.56
Input Entropy	0.19	0.83	1.00	0.91	0.86
Information Gain	0.03	0.65	0.91	1.00	0.80
Empowerment	0.00	0.56	0.86	0.80	1.00

Minecraft

	Task Reward	Human Similarity	Input Entropy	Information Gain	Empowerment
Task Reward	1.00	0.96	0.98	0.91	0.59
Human Similarity	0.96	1.00	0.89	0.78	0.74
Input Entropy	0.98	0.89	1.00	0.98	0.49
Information Gain	0.91	0.78	0.98	1.00	0.38
Empowerment	0.59	0.74	0.49	0.38	1.00

Combined

	Task Reward	Human Similarity	Input Entropy	Information Gain	Empowerment
Task Reward	1.00	0.67	0.54	0.49	0.41
Human Similarity	0.67	1.00	0.89	0.79	0.66
Input Entropy	0.54	0.89	1.00	0.95	0.66
Information Gain	0.49	0.79	0.95	1.00	0.55
Empowerment	0.41	0.66	0.66	0.55	1.00

Table B.1: Pearson correlation coefficients for each environment and combined; see Figure 5.
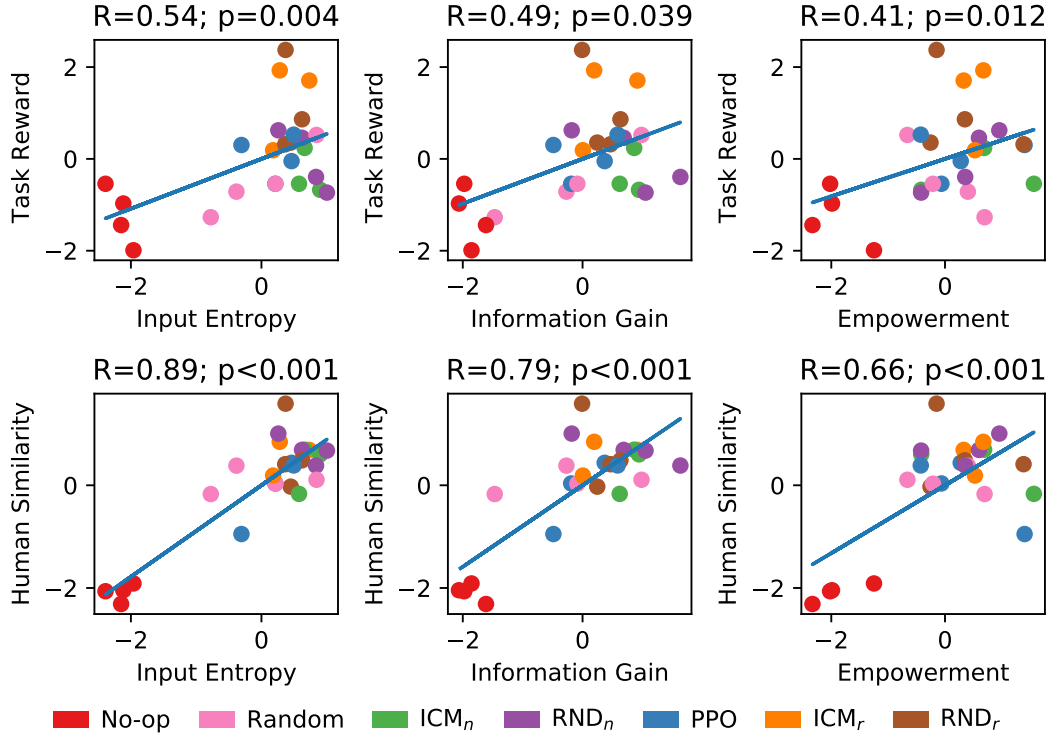
C   Scatter Plots

Figure C.1: Normalized objective values for all agents and environments. The plots visualize the correlations between three intrinsic objectives (X axis) and the two supervised objectives (Y axis). The intrinsic objectives show large correlations with human similarity but their correlations with task reward are weaker. The intrinsic objectives thus capture aspects of agent behavior that is not captured well by the task rewards. For example, they distinguish the qualitatively different behavior of no-op and random agents, even though they obtain similar task reward.

D Information Gain Variants

Information Gain Variant	Shared Discretizations		Unshared Discretizations	
	Task Reward	**Human Sim.**	**Task Reward**	**Human Sim.**
Dirichlet of transition counts	−0.44	−0.36	0.05	0.40
Dirichlet of unique transitions	0.49	0.79	0.35	0.71
Logarithm of transition counts	0.55	0.84	0.37	0.78
Square root of transition counts	0.52	0.84	0.38	0.78

Table D.1: Correlations between four information gain implementations and task reward and human similarity. We compare the implementations for two forms of preprocessing, using the same discretization levels across agents or using a different discretization for each agent based on its percentiles only. We find that the Dirichlet distribution of unique transitions and logarithm and square root of transition counts exhibit greater correlations when using shared discretizations, and that all of the said three methods correlate strongly with human similarity and more weakly with task reward.

E Human Similarity Variants

Human Similarity Variant	**Task Reward**	**Input Entropy**	**Information Gain**	**Empowerment**
Jaccard Similarity	0.67	0.89	0.79	0.66
Jensen-Shannon Divergence	0.26	0.77	0.67	0.66

Table E.1: Correlations between two human similarity implementations and the other objectives. We compare the Jaccard similarity (intersection over union) of the set of inputs visited by the human player and those visited by the RL agent, with the Jensen-Shannon divergence between the two sets. We find that Jaccard similarity correlates more strongly with task reward, slightly more strongly with input entropy and information gain, and near-equally with empowerment, as compared to Jensen-Shannon divergence. The two implementations have a correlation of 0.78 with each other.

F Environments

We consider four environments: Breakout, Seaquest, Montezuma's Revenge, and Minecraft Treechop. The first three are 2D Atari games 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. Minecraft is a 3D block-based game in which the player is free to explore, build, and mine; the Treechop task requires the agent to collect wood from trees to obtain task reward.

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 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., 2018) or human demonstrations (Aytar et al., 2018) have been important to succeed at the game.

Minecraft Treechop MineRL (Guss et al., 2019) is a set of environments in Minecraft, a block-based 3D game in which the player can explore, build, and mine within a procedurally generated world. The Treechop environment provides the player with an axe and restricts the action space such that the player can walk around and break blocks, being given task reward for breaking and collecting wood blocks from trees. Though the goal is clearly defined, there are a wide range of possible activities the agent can pursue.

We follow the standard Atari protocol (Machado et al., 2018). Atari games yield $210 \times 160 \times 3$ images, which we converte to grayscale and rescale to 84×84 for the agents; MineRL uses 64×64 images which are input to the agent directly. The agent chooses one of a set of discrete actions: 4 in Breakout, 18 in Seaquest and Montezuma, and 10 in Minecraft. 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.

G Agents

We selected seven agents to span a wide range of behaviors: no-op and random agents cover low-entropy and high-entropy extremes and the RL agents maximize task reward, intrinsic objectives, and combinations thereof. We use the following agent implementations:

- ICM: https://github.com/openai/large-scale-curiosity
- RND: https://github.com/openai/random-network-distillation
- PPO: https://github.com/hill-a/stable-baselines

Random An agent that uniformly samples random actions from the available action space. graphNo-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 quickly during each training step, stabilizing the learning process.

ICM Intrinsic Curiosity Module (Pathak et al., 2017) is an exploration agent seeks inputs that incur high prediction error of a forward model in a learned feature space. Image embeddings that focus on information relevant to the agent are learned by predicting the action from the current and next input. A second network is trained to predict the next embedding from the current embedding an action, and its prediction error serves as the intrinsic reward.

RND Random Network Distillation (Burda et al., 2018) is an exploration agent that seeks inputs that incur high prediction error of a model that predicts image embeddings. The image embeddings are generated by a randomly initialized and fixed neural network. The prediction error serves as the intrinsic reward.