Measuring and Modeling Physical Intrinsic Motivation

Julio Martinez[1,*], **Felix Binder**[2], **Haoliang Wang**[3], **Nicker Haber**[4,5], **Judith Fan**[1,3], **Daniel L. K. Yamins**[1,5]
[1]Department of Psychology, Stanford University
[2]Department of Cognitive Science, University of California San Diego
[3]Department of Psychology, University of California San Diego
[4]Graduate School of Education, Stanford University
[5]Department of Computer Science, Stanford University
[*]juliomz@stanford.edu

Abstract

Humans are interactive agents driven to seek out situations with interesting physical dynamics. Here we formalize the functional form of physical intrinsic motivation. We first collect ratings of how interesting humans find a variety of physics scenarios. We then model human interestingness responses by implementing various hypotheses of intrinsic motivation including models that rely on simple scene features to models that depend on forward physics prediction. We find that the single best predictor of human responses is adversarial reward, a model derived from physical prediction loss. We also find that simple scene feature models do not generalize their prediction of human responses across all scenarios. Finally, linearly combining the adversarial model with the number of collisions in a scene leads to the greatest improvement in predictivity of human responses, suggesting humans are driven towards scenarios that result in high information gain and physical activity.

Keywords: Intrinsic Motivation; Curiosity; Intuitive Physics; Information Seeking

Introduction

From infancy, humans exhibit strong intrinsic motivations to curiously explore their physical environments and play with the objects they encounter. Their exploration patterns are not random but follow an underlying systematic approach. More generally, curiosity has been described as an intrinsic motivation that aids in closing gaps in knowledge (Loewenstein, 1994; Oudeyer & Kaplan, 2009; Kidd & Hayden, 2015). Many works have contributed to the understanding of human preferences during free-play and self-directed learning. Previous work shows that children prefer stimuli of intermediate predictivity (Cubit, Canale, Handsman, Kidd, & Bennetto, 2021) and that the inability to predict an outcome of an action is a powerful driver that steers information seeking behavior (Markant & Gureckis, 2014). Self-directed learning allows humans to focus their efforts on useful information they do not yet posses resulting in highly selective but efficient information sampling strategies (Kidd, Piantadosi, & Aslin, 2012).

A core component of physical intrinsic motivation is how it influences the learning of a physical world model (WM). In learning physical WMs infants are faced with the need to observe a wide range of complex physical dynamics (Kim, Sano, De Freitas, Haber*, & Yamins*, 2020). Thus it is a natural hypothesis that in order to gather data to learn good WMs there must be a powerful exploration strategy.

In this work we seek to formalize physical intrinsic motivation – the problem of determining the *intrinsic reward function* (IRF) that directs exploration actions in humans during free-play in physics environments. To do this we directly probe "interestingness" as a proxy for intrinsic reward (IR) and evaluate various IRFs as conceptually distinct hypotheses. Our contributions are as follows: 1. We collect human interestingness ratings over a diverse range of 3D simulated videos. 2. We formulate the formal theory of intrinsically motivated agents with different IRFs. 3. We compare different IRFs and composites thereof in terms of their predictivity of human ratings and their generalization across scenario types.

Our core conclusions are that WM-based IRFs generalize better than simple scene features in predicting human responses across scenarios suggesting a promising direction to improve the current set of WM-based IRFs. Second, that humans are characterized in part by both information seeking and high activity seeking motivations – humans find as interesting the stimuli whose outcomes are hard to predict or fun to watch because of the many collisions observed.

Theoretical Framework

Intrinsically Motivated Agents We start by using a Reinforcement Learning (RL) framework as in (Haber, Mrowca, Wang, Fei-Fei, & Yamins, 2018; Kim et al., 2020) to define an intrinsically motivated agent who generates its own IR (Fig. 1a). At time t an RL agent makes an observation from state $s_t \in S$ and takes an action $a_t \in A$ which results in a state transition $s_{t+1} \in S$ (i.e. the physical dynamics). We say the agent is intrinsically motivated when it's IR r_t is based on its own observations of the environment or its predictions of state transitions.

The Policy Model (PM) is a function $PM : S \rightarrow A$ that maps the state to an action. The PM is built around a subcomponent called the IRF that generates the IRs. Given a history H_t of states, actions, and IRs, the agent is able to learn a policy that optimizes for total IR. The physical WM, parameterized by θ_{WM} (i.e. the agent's knowledge of the world), allows the agent to simulate forward in time the physical dynamics of its environment. Its functional form takes as input the state of the environment and an action, and predicts the next state $WM : S \times A \rightarrow S$. The IRF is a function that generates the total reward from a sequence of state action pairs and can potentially depend on the state of the environment, the action being taken, and the state of the agent's knowledge of how

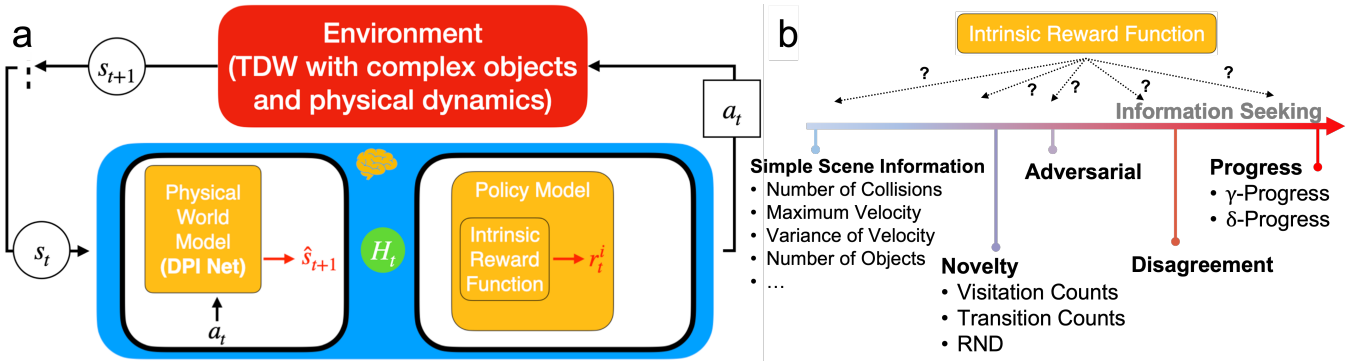

Figure 1: **Theoretical framework for an IR agent. (a)** Intrinsic reward RL agent acting in a simulated environment (TDW) that uses a WM for forward physics prediction, and a PM that optimizes an IR generated by an IRF. **(b)** A graphic describing the hypothesis space for IRFs, in the form of a continuum that progressively becomes more explicitly information seeking towards the right of the continuum.

the world evolves over time.

$$IRF : S \times A \times \theta_{WM} \rightarrow \mathbb{R}$$

Intrinsic Reward Functions We conceptualize the hypothesis space for IRFs by the degree to which they explicitly improve the WM i.e. by an information seeking continuum (Fig. 1b). On the left extreme of this continuum we have simple scene features, such as the number of object collisions, a data gathering signal which may or may not result in improving the WM. Progressing towards the right we have IRF classes that are increasingly designed to explicitly improve the WM. The *novelty* of the state (Tang et al., 2017) generates more reward for unfamiliar states effectively improving the WM by providing diverse training stimuli. *Adversarial* generates rewards proportional to the WM loss (Stadie, Levine, & Abbeel, 2015; Pathak, Agrawal, Efros, & Darrell, 2017). *Disagreement* generates rewards based on the variance in predictions from multiple WMs (Pathak, Gandhi, & Gupta, 2019) to minimize uncertainty across WM predictions. At the far right we have *Progress* which assigns higher reward to observations that reduce the WM loss (Ten, Kaushik, Oudeyer, & Gottlieb, 2021; Schmidhuber, 2010; Graves, Bellemare, Menick, Munos, & Kavukcuoglu, 2017; Achiam & Sastry, 2017). It is important to note that specific implementations of each IRF vary and may not preserve the order in which their conceptual counterpart appears on the continuum. To be explicit, we give details of the specific IRF implementations used for this work below.

Simple Scene Features based on the information readily available in the scene of each stimulus, as in (Kachergis et al., 2021; Holdaway et al., 2021), were used as candidate IRFs. These IRFs do not depend on a physical WM. They are defined as follows: *Position*: the x, y and z coordinates of all positions for all objects in a scene from which we compute the mean, variance, min, and max. *Velocity*: the x, y, and z coordinates of the velocities of all objects in a scene from which we compute the mean, variance, min, and max. *Variance of position*: the trace of the covariance of all object's positions from which we compute the initial value, mean, min, and max over several stimulus frames. *Variance of velocity*: the trace of the covariance of all object's velocities from which we compute the mean, min, max and initial value. *Number of collisions*: the number of collisions occurring over an entire stimulus. We also compute the initial number of collisions (how many objects are in contact that separate initially), mean, min, and max number of collisions over temporal intervals throughout the stimulus. *Number of objects*: the number of objects in a stimulus. *Number of object categories*: the total number of different types of objects in a stimulus. *Number of distractors*: total number of objects that are fixed and that do not interact with other objects.

Random Network Distillation (RND) is a novelty based IRF (Burda, Edwards, Storkey, & Klimov, 2018). In RND a randomly initialized target network, $\mu_{\theta_{target}}$ takes as input the state at time t, $s_t \in S$, and embeds it into a d-dimensional representation. A predictor network, $\mu_{\theta_{predictor}}$, randomly initialized using a different seed, is trained to minimize the RMSE between its prediction of the target network's embeddings. IR for a single state observation is equal to the RMSE between the predicted and target embeddings. RND does not rely on WM prediction, but does rely on the predictor network learning of the target network embeddings.

$$r^i_{t,rnd} = RMSE(\mu_{\theta_{target}}(s_t), \mu_{\theta_{predictor}}(s_t))$$

Adversarial reward, an approximation to surprisal (Achiam & Sastry, 2017) sets the IR for a single step equal to the WM loss for a k-step rollout. The k-step rollout is an autoregressive rollout where the WM first takes as input the current state s_t and predicts $\hat{s}_{t+1}$ (one step ahead). For a 2-step rollout, the WM feeds its 1-step prediction, $\hat{s}_{t+1}$ as input to make a 2-step rollout prediction $\hat{s}_{t+2}$, and so forth for a k-step rollout, $\hat{s}_{t+k}$. The MSE between the ground truth state at the kth step,s_{t+k} and the predicted $\hat{s}_{t+k}$ is the IR value.

$$r^i_{t,adversarial} = \mathcal{L}_{\theta_{WM}}(s_{t+k}, \hat{s}_{t+k}) = MSE(s_{t+k}, \hat{s}_{t+k})$$

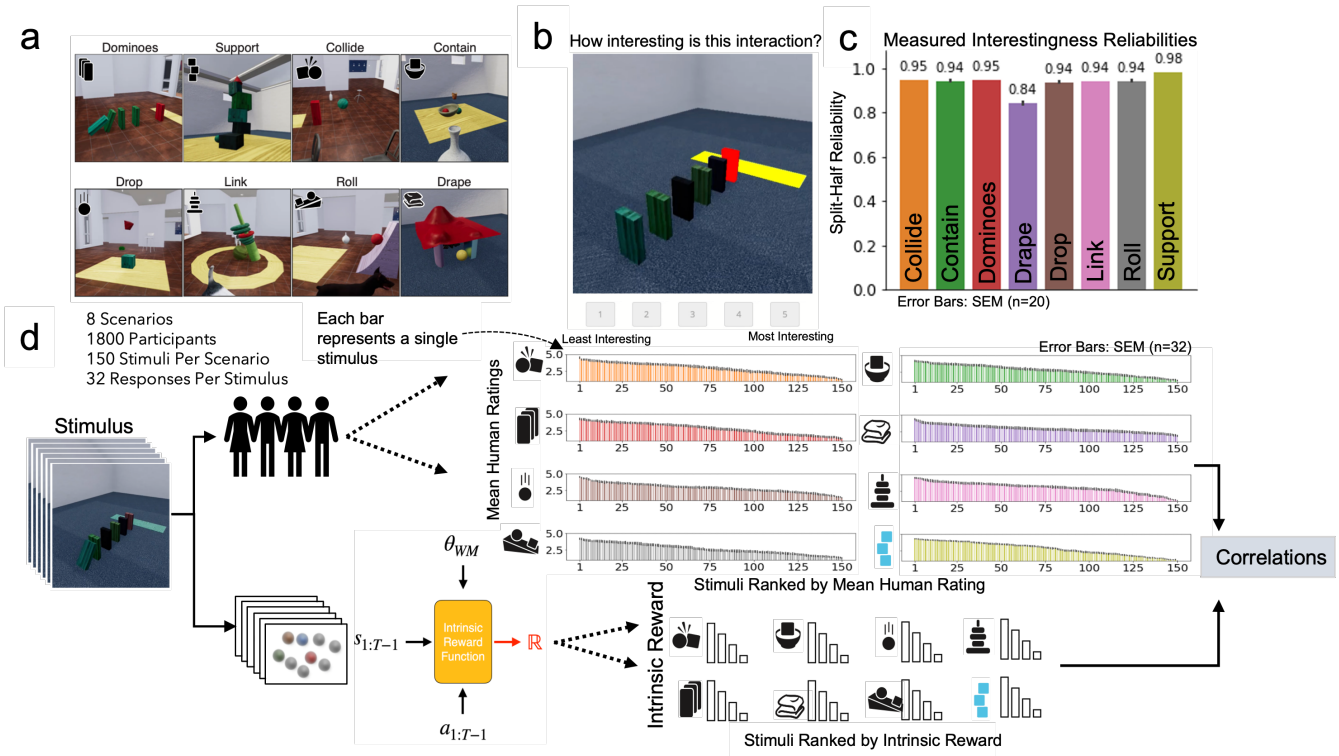

Figure 2: **Stimuli, Task Procedure and Modeling Pipeline.** **(a)** Stimulus examples of the 8 scenario categories from the Physion Benchmark dataset. Stimuli contain a diverse set of rigid and non-ridge body physical dynamics useful for probing and modeling physical intrinsic motivation. **(b)** In the experimental task, participants are shown the stimulus videos. After watching participants are asked to rate how interesting they find the video on a Likert scale from 1 to 5 (5 being most interesting). **(c)** The scenario mean split-half reliabilities for human ratings (n=20 split-halves for each stimulus and averaged across stimuli in each scenario). **(d)** The experimental and modeling pipeline that feeds stimulus to both humans and various IRFs. Mean human responses for each stimulus are correlated to IRF values to compare their predictivity on human interestingness.

Disagreement (Pathak et al., 2019) depends on multiple WMs each of which generates k-step rollouts resulting in predictions $\hat{s}_{t+k}^{\theta_{WM_1}}$, $\hat{s}_{t+k}^{\theta_{WM_2}}$, and $\hat{s}_{t+k}^{\theta_{WM_2}}$. The IR at time t is assigned the mean variance across each WM's k-step predictions made from time t.

$$r^i_{t,disagreement} = Mean(Var(\hat{s}_{t+k}^{\theta_{WM_1}}, \hat{s}_{t+k}^{\theta_{WM_2}}, \hat{s}_{t+k}^{\theta_{WM_3}}))$$

δ-*Progress* computes the difference between an old and new WM's loss of a k-step rollout prediction as shown below (Graves et al., 2017; Achiam & Sastry, 2017). The new WM is the current WM, or more generally, the WM trained after n number of training iterations parameterized by θ_{WM_n}. The old WM is δ training steps back, parameterized by $\theta_{WM_{n-\delta}}$.

$$r^i_{t,progress} = \mathcal{L}_{\theta_{WM_{n-\delta}}}(s_{t+k}, \hat{s}_{t+k}) - \mathcal{L}_{\theta_{WM_n}}(s_{t+k}, \hat{s}_{t+k})$$

Each IRF intermediately outputs an IR r^i_t for every transition in the environment. The IRF sums over the IRs of an entire stimulus of length T to compute the total IR, $\sum_{t=0}^{T-1} r^i_t \in \mathbb{R}$.

Choice of World Model In real life humans do not have access to the exact state of the world and instead infer it from perception through partial observation. Here we bypass perception, as it is out of scope for the problem under investigation, and directly feed our WMs an explicit state description as input. The state description is generated by converting the 3D state of the virtual environment to a 3D mesh from which object centric coordinates or particles are used to describe each object's position and velocity in the scene.

For physics prediction, we chose a model whose performance best represents human prediction accuracy, DPINet (Li, Wu, Tedrake, Tenenbaum, & Torralba, 2018). DPINet is part of a class of neural network scene graph physics models that learn physical prediction via local and hierarchical backpropagation (Li et al., 2018; Mrowca et al., 2018). DPINet showed the most human like accuracy and consistency on Physion when compared to other earlier versions of scene graph based models and several image based models, see Fig. 5 in (Bear et al., 2021). DPINet was pretrained on the Physion training set to reproduce the performance results previously stated (Bear et al., 2021). DPINet receives as input the state particle description, the positions and velocities of all the object particles, and makes a forward prediction of the positions and velocities of all object particles for the next time step.

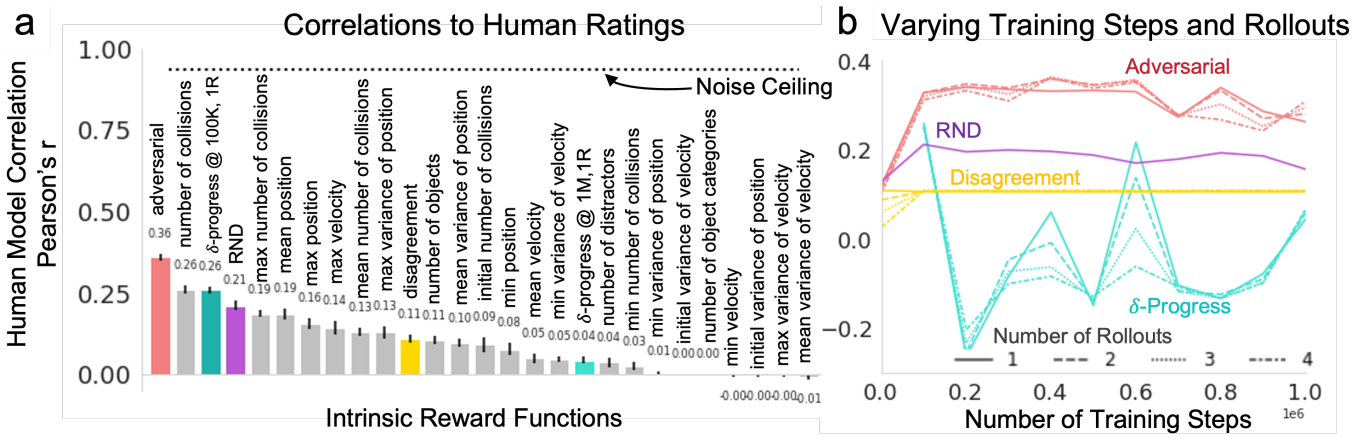

Figure 3: **Model-Human Correlation Results for Individual IRF Candidates. (a)** The correlations of IRFs to mean human ratings. Adversarial reward has the highest correlation to humans. δ-progress is shown twice, first at 0.1×10^6 training steps of DPINet and again at 1×10^6 training steps for DPINet. **(b)** We vary the amount of pretraining of our WM by varying the number of training steps for DPINet that adversarial, disagreement, and δ-progress depend on for WM prediction. We also vary the number of rollout steps DPINet makes for its predictions in each pretrained DPINet. Like DPINet we also vary the number of training steps for the RND predictor network. RND however does not depend on rollouts. As we vary the training, adversarial, disagreement, and RND are relatively stable after 0.1×10^6 time steps. δ-progress on the other hand is relatively unstable often changing from positive to negative correlation over many time steps. Adversarial, disagreement, and δ-progress do not change much as a function of the number of rollouts.

Experiment

For our experiment we gathered human interestingness ratings on a wide range of physics scenarios to serve as a target in modeling IR.

Participants We recruited 1,800 participants from Prolific to complete the task. Participants provided informed consent and were paid approximately $14 per hour.

Stimuli We use the Physion Benchmark dataset (Bear et al., 2021) generated from 3D simulated videos in ThreeDWorld (Gan et al., 2020). We repurpose Physion by changing the design variable to "interestingness" (instead of prediction accuracy) as a proxy for total IR by asking adult humans how interesting they find the stimulus. We chose these stimuli because they provide a diverse set of rigid and non-rigid body dynamics wherein to study physical intrinsic motivation. Physion is split into a training set, used to train WMs and IRFs, and a test for the experiment and for evaluating IRFs. Each set contains 8 scenarios: Collide, Contain, Dominoes, Drape, Drop, Link, Roll, and Support shown in Fig. 2a. There are 2000 and 150 videos for each scenario in the train and test sets respectively.

Task Procedure Participants were asked for interestingness ratings (Fig 2b) after observing the outcomes in each stimulus video in the Physion test set. Each of 150 trials began with a fixation cross, shown for a randomly sampled time between 500ms and 1500ms. Participants were then shown the first frame of the video for 2000ms after which the entire video was played. Once the video stopped playing, it was removed and the response buttons were enabled. The experiment moved to the next phase after participants selected an interestingness rating from 1 to 5. Participants were presented with the stimuli in a randomized sequence and were only allowed to take the task once. Data was excluded from participants that did not complete trials, whose exit survey indicated they did not understand the study, or failed to include at least one response for each endpoint of the scale.

Validation and Reliability We report the reliability between participants (mean of n=32 responses per stimulus) as the split-half reliability using Spearman Brown correction. We compute the average split-half reliability for a given stimulus using 20 split halves and average across stimulus within each scenario category. The cross scenario category mean split-half reliability was 0.935 with 0.002 SEM. All scenario category reliabilities are shown in shown in Fig. 2c.

Data Analysis Pipeline Comparing human action choices during free-play to the action choices of an RL agent is a natural modeling choice. However, doing so requires a PM to plan the optimal sequence of actions. Incorrectly choosing a suboptimal PM introduces an additional source of error in planning. We avoid this confound by using IRFs to instead directly generate total IR for our predefined stimulus and forgo using a PM for interactive RL. We compute Pearson r correlations to evaluate how predictive each IRF is to mean human ratings for each of the 150 rated stimuli in all 8 scenarios depicted in the upper half of Fig. 2d.

To compute IR from WM-based IRFs we pretrain DPINet for 1 million time steps on the training stimulus set from

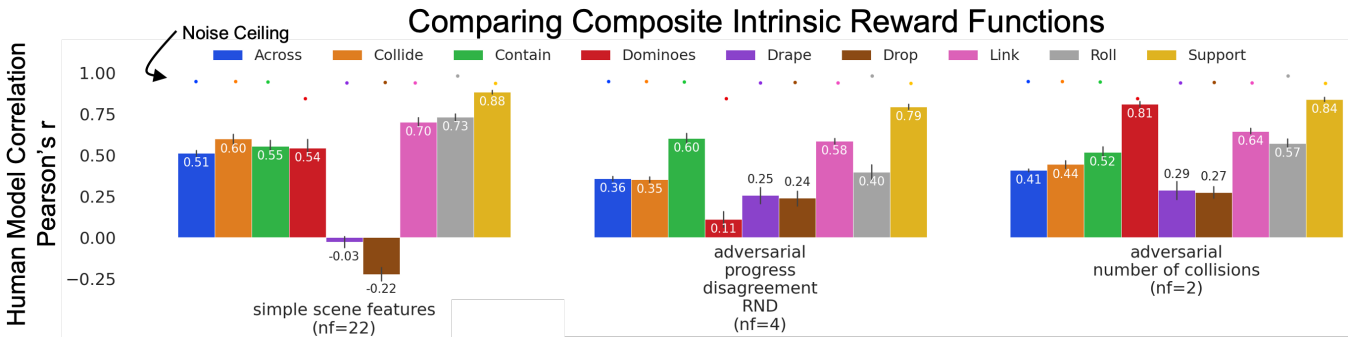

Figure 4: **Analysis of IRF Composites.** Three groups of bar plots correspond to three composite functions of IR. The first group on the left is a composite model using all simple scene features as predictors in linear regression. While correlation performance is 0.51 across scenarios, as indicated by the blue bar, Drop and Drape scenarios have small and negative correlation indicating an inability for linear combinations to generalize across scenarios well. The center group shows the results if a composite model of all WM-based IRFs and RND. Generalization is much better across scenarios although still below noise ceiling. The third group on the right shows a composite model of the adversarial IRF with number of collisions, the feature that when linearly combined with adversarial improves accuracy across scenarios the most.

Physion, resulting in similar loss from (Bear et al., 2021). Before training we extracted x, y, z particles representations for object positions and velocities at each frame of the stimuli. These object particle representations are fed as input into DPINet and the IRFs. We compute δ-progress, disagreement, and adversarial total reward at several training steps of DPINet $(0, 100 \times 10^3, 200 \times 10^3, ..., 1 \times 10^6)$. For each training step of DPINet, the WM-based IRFs were computed based on 1, 2, 3, or 4-step rollouts to test whether simulating further forward in time resulted in significant differences in IR prediction. For disagreement we pretrained DPINet models with 3 initialization seeds. For δ-progress δ =50,000 gave best results over a grid search on δ. For RND, we modified only the final layer of DPINet to output a 200 dimensional state embedding. This modified DPINet served as the architecture for our RND predictor and target networks (both architectures were identical with different initialization seeds). The predictor network was pretrained for the same number of time steps as DPINet.

In addition to evaluating individual IRFs, we trained linear regression models to predict human ratings using L1 regularization and leave-one-out cross validation to find the best fits for composite IRFs. We evaluated all model predictions by correlating them to mean human ratings. 10 random splits from the stimulus set were drawn and for each split 80% was used as regression model training data. The remainder 20% stimuli was used for evaluating predictions. The same test splits were used to evaluate the individual IRFs.

Results

Explicitly world-model-based IRFs tend to better explain human interestingness ratings than simple scene features, with adversarial loss achieving the best overall match. Evaluating the predictivity of single IRFs across all scenario types (Fig. 3a), we found a range of predictivity levels, with several simple scene features indistinguishable from 0 in their predictivity. In contrast, explicitly WM-based IRFs included 3 of the top 4 best predictors. Adversarial loss achieved the best predictivity, with an average cross-scenario correlation of 0.360 (0.39% of noise ceiling of 0.935).

World-model-based intrinsic reward functions are stable across rollouts and, mostly, across training steps. As shown in Fig. 3b, correlations of all IRF models were stable with respect to rollout length. Adversarial, progress, and RND provide stable correlations to human ratings after 100×10^3 training steps, with relative rankings maintained throughout. δ-Progress IRF was an exception, changing significantly from positive to negative correlation across training steps. (For this reason, Fig. 3a shows δ-progress results at both the maximally-correlated and final timesteps.) All IRF models included here are based on DPINet (or modified DPINet for RND). In future work we will measure the effects of using other WM architectures including pixel-based forward predictors such as TECO (Yan, Hafner, James, & Abbeel, 2022) and MCVD (Voleti, Jolicoeur-Martineau, & Pal, 2022).

Simple scene features do not generalize across scenarios but world model based intrinsic reward functions do. Linearly combining groups of simple scene features as predictors improves overall predictivity of human ratings, but reduces predictivity in specific scenarios, particularly Drop and Drape (Fig. 4, left). Interestingly, this outcome reproduces previous work from (Holdaway et al., 2021; Kachergis et al., 2021) where the Support scenario was predicted well using scene features, but the Drop scenario was not. Looking at the relationships of individual simple scene features to each scenario, we find that the cause of this outcome is the change of sign in relationship across scenarios (columns of Fig. 5). This indicates that, to the extent that simple scene features "explain" interestingness, they do so in a highly scenario-type

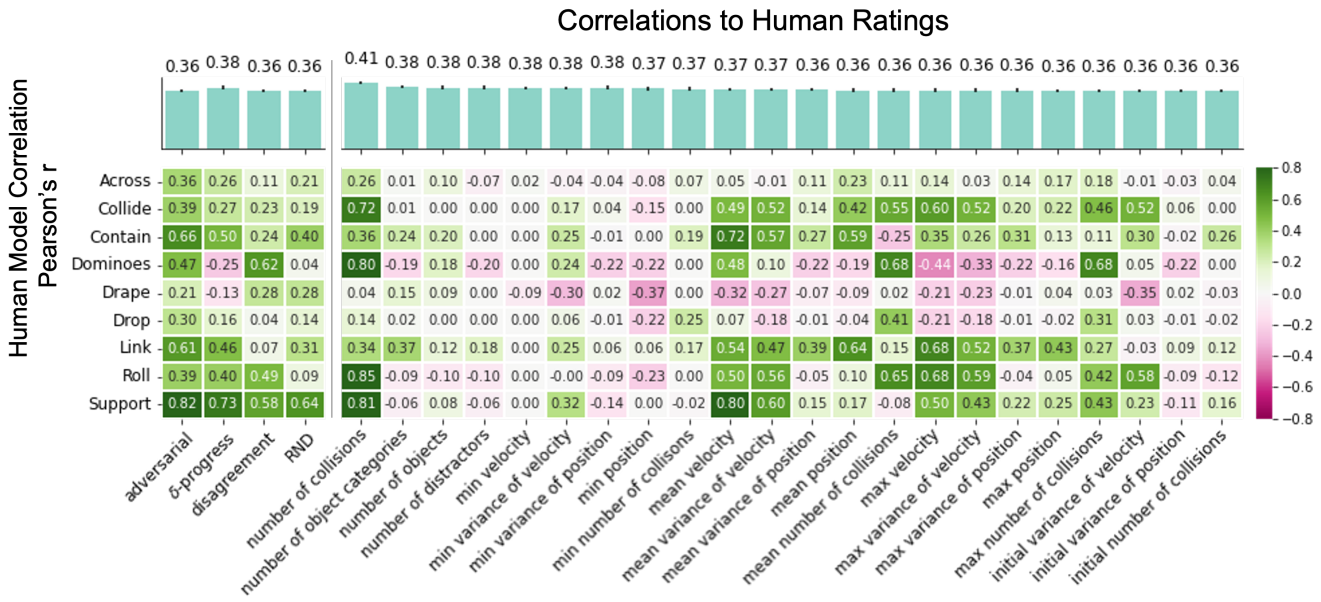

Figure 5: **Bottom Panel: Investigating per scenario results for each candidate IRF.** Each box represents the correlation of the indicated IRF (columns) for the indicated scenario (rows). All green within a column indicates the given IRF has the same sign across all scenarios; Columns mixed with green and pink indicate inconsistent correlation signs across scenarios and thus lack of generalization.**Top Panel: Combining adversarial IRF with other IRFs.** Each blue bar represents the predictive accuracy of the two-feature model linearly combining the best single-feature IRF (adversarial) with the indicated feature. Bars are ordered by complementarity with adversarial.

specific manner, and do not generalize. On the other hand, inspecting per-scenario results from the explicitly WM-based IRFs shows that the explanatory direction is consistent across all scenarios (Fig. 4, center), indicating improved generalization. Correlation signs are consistently positive across scenarios for all world-model-based IRFs (Fig. 5).

All intrinsic reward function correlations to human ratings are far below noise ceiling. Despite some explanatory power in several IRF candidate models, all are far below noise ceiling both across scenarios (dashed line in Fig. 3a) and for most individual scenarios (dots over each column in Fig. 4).

Adversarial reward and number of collisions are the most complimentary intrinsic reward components. To determine complementarity among our IRF candidates, we built linear combinations of pairs of IRFs. Starting with the best matching adversarial IRF, we found that the most complementary feature was a WM independent IRF, the number of collisions (Fig. 5, blue bars).

Discussion

In this work, we measured human interestingness judgements and assessed several IRFs on their predictivity of human ratings. We observed that IRFs explicitly involving assessment of WM knowledge tended to be better predictors of human data than simple scene features, with adversarial loss having the best overall correlation. This suggests humans are at least partly motivated by "information seeking" goals in our experimental setting. Explicit WM-based IRFs also generalized across different types of physical scenarios, while simple scene features did not. Perhaps most saliently, all IRFs tested in this work are well below the human experimental noise ceiling. This encourages us to make further improvements for WM-based IRFs that can account for intrinsic motivation over a wide range of physical scenarios.

While WM-based IRFs add complexity in modeling forward dynamics, simple scene features require pre-specifying only the relevant features for any new scenario a human might encounter. However, we also observed that the most complementary predictor to adversarial reward was the raw number of collisions in a given scenario, particularly in scenarios like Dominoes where object collisions are especially salient. This suggests participants sometimes valued scenarios as interesting based on the amount of activity rather than explicitly for information gain. We plan to explore simple scene features that more directly relate to physical activity and that may potentially yield further performance improvements when integrated into a WM-based IRF.

Finally, we also seek to further investigate if this bifurcation in explanatory modes (i.e. the level of physical activity vs information gain) is an artifact of our experimental design (e.g. because it measures interestingness in non-interactive displays rather than real interactive settings), or a core feature of intrinsic motivation that a better IRF model will need to explain.

Acknowledgments

This work was supported by the following grants: Simons Foundation grant 543061 (D.L.K.Y), National Science Foundation CAREER grant 1844724 (D.L.K.Y), Office of Naval Research grant S5122 (D.L.K.Y.), Stanford University Human-Centered Artificial Intelligence Inaugural Fellowship.

References

Achiam, J., & Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

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

Cubit, L. S., Canale, R., Handsman, R., Kidd, C., & Bennetto, L. (2021). Visual attention preference for intermediate predictability in young children. *Child development*, *92*(2), 691–703.

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., ... others (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

Graves, A., Bellemare, M. G., Menick, J., Munos, R., & Kavukcuoglu, K. (2017). Automated curriculum learning for neural networks. In *international conference on machine learning* (pp. 1311–1320).

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F., & Yamins, D. L. (2018). Learning to play with intrinsically-motivated, self-aware agents. *Advances in neural information processing systems*, *31*.

Holdaway, C., Bear, D. M., Radwan, S. F., Frank, M. C., Yamins, D. L., & Fan, J. E. (2021). Measuring and predicting variation in the interestingness of physical structures. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Kachergis, G., Radwan, S. F., Long, B., Fan, J. E., Lingelbach, M., Bear, D. M., ... Frank, M. C. (2021). Predicting children's and adults' preferences in physical interactions via physics simulation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kim, K., Sano, M., De Freitas, J., Haber*, N., & Yamins*, D. (2020). Active world model learning with progress curiosity. In *International conference on machine learning* (pp. 5306–5315).

Li, Y., Wu, J., Tedrake, R., Tenenbaum, J. B., & Torralba, A. (2018). Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, *116*(1), 75.

Markant, D., & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. *Advances in neural information processing systems*, *31*.

Oudeyer, P.-Y., & Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 6.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning* (pp. 2778–2787).

Pathak, D., Gandhi, D., & Gupta, A. (2019). Self-supervised exploration via disagreement. In *International conference on machine learning* (pp. 5062–5071).

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, *2*(3), 230–247.

Stadie, B. C., Levine, S., & Abbeel, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., ... Abbeel, P. (2017). A study of count-based exploration for deep reinforcement learning. In *31st conference on neural information processing systems, long beach, ca, dec* (pp. 4–9).

Ten, A., Kaushik, P., Oudeyer, P.-Y., & Gottlieb, J. (2021). Humans monitor learning progress in curiosity-driven exploration. *Nature communications*, *12*(1), 5972.

Voleti, V., Jolicoeur-Martineau, A., & Pal, C. (2022). Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*.

Yan, W., Hafner, D., James, S., & Abbeel, P. (2022). Temporally consistent video transformer for long-term video prediction. *arXiv preprint arXiv:2210.02396*.