What can AI Learn from Human Exploration?
Intrinsically-Motivated Humans and Agents in Open-World Exploration

Yuqing Du[*†], **Eliza Kosoy**[*‡], **Alyssa L. Dayan**[*†], **Maria Rufova**[†], **Pieter Abbeel**[†], **Alison Gopnik**[‡]

Abstract

What drives exploration? Understanding intrinsic motivation is a long-standing question in both cognitive science and artificial intelligence (AI); numerous exploration objectives have been proposed and tested in human experiments and used to train reinforcement learning (RL) agents. However, experiments in the former are often in simplistic environments that do not capture the complexity of real world exploration. On the other hand, experiments in the latter use more complex environments, yet the trained RL agents fail to come close to human exploration efficiency. To study this gap, we propose a framework for directly comparing human and agent exploration in an open-ended environment, Crafter [23]. We study how well commonly-proposed information theoretic intrinsic objectives relate to actual human and agent behaviours, finding that human and intrinsically-motivated RL agent exploration success consistently show positive correlation with Entropy and Empowerment. However, only human exploration shows significant correlation with Information Gain. In a preliminary analysis of verbalizations, we find that children's verbalizations of goals show a strong positive correlation with Empowerment, suggesting that goal-setting may be an important aspect of efficient exploration.

1 Introduction

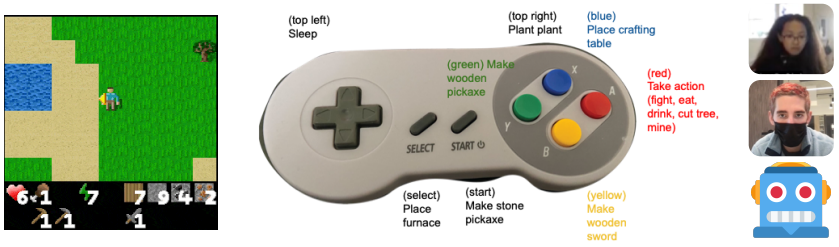

Figure 1: Left: Example screen from Crafter [23]. The player is at the center of the screen; the yellow arrow shows which direction they are facing. Their health, food, water and energy status are at the bottom left, the raw materials they have collected are at the bottom right, and the tools built so far are in the bottom row. Middle: Actions available to the human participants and RL agents. Right: We compare behaviours of children, adults, and RL agents.

Humans often explore new environments remarkably effectively, even in the total absence of external rewards [40]. There have been many attempts to formalize the natural curiosity of humans, but empirical studies are limited in that they tend to put humans in simple and unrealistic environments and look at at limited range of exploratory behaviours, e.g. [8]. Tracking the complexities of more open-ended and spontaneous exploration has proven challenging. Kosoy et. al have begun designing more complex unified online environments which allow spontaneous exploration [16, 17],

[*]These authors contributed equally to this work. [†]UC Berkeley EECS. [‡]UC Berkeley Psychology.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

but have still primarily focused on simple tasks. Other work such as [38] are limited by what can be clicked within the realm of a specific game. Our goal is to study human exploration in a more complex setting that can also help us develop more effective AI agents. Reinforcement learning (RL) agents must actively collect meaningful experience to find optimal behaviours in initially unknown environments. To facilitate this, existing works have proposed various objectives that guide exploration by approximating some notion of novelty or curiosity, with some commonly-used concepts being count-based state-visitation [42, 39], prediction error [44, 53], state novelty [9, 67], skill learning [18, 32], or information gain [27, 45]; see [2, 50] for surveys. However, general intrinsically-motivated RL agents still don't come close to human-level sample efficiency, these intrinsic rewards sometimes produce counterproductive behavior [9], and it remains unclear if engineered intrinsic rewards are truly aligned with human exploration. Few studies use human exploration as a basis for agent exploration—some examples include illuminating key differences between human and agent priors [14], or objectives such as [44] being loosely inspired by how children exploration is driven by curiosity and seeking novelty.

Motivated by the gap between human and agent exploration, we study the behavior of humans and agents in the same environment—Crafter [23], a Minecraft-like complex, open-ended environment. We collect play data from both children and adults, emphasizing that the child behavioural data is important for gaining insights into fundamental untrained exploration capacities of humans. We propose five ways of scoring exploration in the game and analyze how well the exploration performance of humans and agents correlate with commonly-used information theoretic objectives that have also been used to explain exploration motivations: Entropy, Information Gain, and Empowerment. We find that human and agent exploration performance is consistently positively correlated with Entropy and Empowerment, but only human performance significantly correlates with Information Gain. Information Gain is also the only objective where both adults and children perform significantly better than the agents; this is despite one type of agent being explicitly trained to optimize an approximation of Information Gain, and the overall exploration scores of humans and agents spanning a similar range (although humans are far more sample efficient). This suggests that for agents to exhibit more human-like and sample efficient exploration, it may be worth exploring the design of intrinsic reward functions for agents that are better aligned with Information Gain.

We also record and transcribe human utterances during play, and in a preliminary analysis find a significant positive correlation between children's frequency of verbalizing goals and their Empowerment, supporting previous work in psychology which has suggested that self-talk could play an important role in children's creative problem-solving [33, 47, 64] and in AI which suggests goal-generation aids exploration in agents [13, 10, 28, 11].

2 Related Work

We discuss the related works in more depth in Appendix A. Exploration strategies in AI range in complexity from occasionally acting randomly (*e.g.*, ϵ-greedy) to optimizing complex intrinsic reward functions. Intrinsic rewards are often motivated by objectives such as increasing entropy, information gain, and/or empowerment. Similarly, a range of objectives underlying human exploration has been studied in the cognitive science literature [62, 38], including perceived novelty [59, 6, 58, 49] which simply suggests that humans are drawn to stimuli that appear more novel, expected learning progress [3, 61, 43] which is the idea that people find it intrinsically rewarding to improve their performance, information gain [36, 52, 1] where the driver of exploration is to gather maximal information about the environment, (or even more specifically the possibility of learning causal relations [59]) maximizing empowerment [30, 22, 8] and totally random exploration more common in younger children [41].

3 Environment and Data Collection

Open-Ended Environment: Crafter. Motivated by the lack of human exploration studies in rich and open-ended environments, we conduct our comparisons in Crafter [23]. Similar to Minecraft, Crafter comprises of exploration challenges in both the breadth and depth of activities to explore. The player controls a character in a procedurally generated world containing various resources that can be collected and used to replenish health or build tools (Figure 1). Players are able to explore a breadth of skills: collecting food and water, sleeping, and avoiding or killing enemies, as well as

depth of skills: crafting increasingly complex tools. This gives rise to the achievement tree shown in Figure 5. See Appendix B for more details on Crafter.

Participants and Agent Training. In this pre-registered, IRB approved study (AsPredicted reference: 92521), we introduced children and adults to the novel "Crafter" game. We recruited 51 children between the ages of 6-10 years (Mean age: 8.6 years, Female: 19, Male: 32) from the Bay Area Discovery Museum (BADM), as well as 24 adults from the University of California, Berkeley campus ages 18-25 years (Mean age 24.8, Female: 10, Male: 14). No direction was given about the game in order to encourage open-ended play, and participants were allowed to play for up to 20 minutes. As baselines we train three RL agents and use one random agent. Specifically, we compare against state-of-the-art intrinsic RL objectives: NovelD [67] and APT [37]. NovelD incentivizes information gain by providing a large intrinsic reward at the boundary between explored and unexplored regions, using RND [9] as a measure of state novelty. APT uses a particle-based entropy estimator [57] to reward the agent for maximizing state entropy in a learned representation space. As a measure of best-case performance we also train an agent with the game extrinsic reward function, which reveals the possible set of achievements. See Appendix D for details.

4 Results and Analyses

As there is no single objective measure for "good exploration", we propose five exploration scores for Crafter (Table 1). Humans and trained agents generally attain comparable exploration scores by the end of their learning, although humans are many orders of magnitude more efficient. Details about the scores and summary statistics can be found in Appendix C.

Analyzing Information Theoretic Objectives.

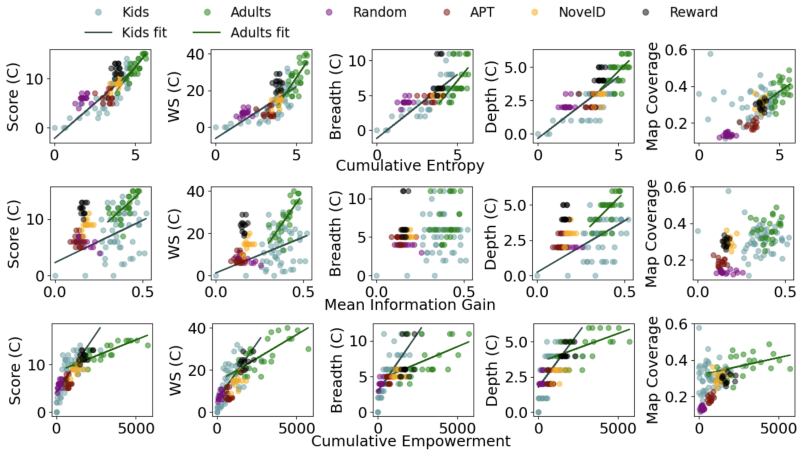

Figure 2: Information theoretic objectives vs. exploration scores. We note a consistent positive correlation between the human participants' achievement-based exploration scores and their attainment of the information theoretic intrinsic objectives. This same trend appears to hold across the population of agents for Entropy and Empowerment.

We verify whether objectives proposed as intrinsic motivation functions are indeed significantly correlated with human and agent exploration. We focus on Entropy, Information Gain, and Empowerment, each of which have been proposed as motivations for exploration in both the AI and cognitive science literature (see Section 2). For each objective, we look at the *cumulative experience* of each person or agent over all episodes, rather than calculating them independently per episode (see Appendix E for details). In Figure 2, we plot each information theoretic objective against the proposed exploration scores and compute a least squares linear fit for the humans, plotting the fit where significant ($p < 0.05$). We find consistent positive correlation (though stronger for adults) between humans' exploration scores and their information theoretic objectives. The same trend appears to hold across the entire agent population for Entropy and Empowerment. We do not see similar correlations for map coverage, suggesting that the information theoretic objectives are more aligned with exploration in the skill space rather than just physical exploration of the map space.

Interestingly, Information Gain is the only metric with no significant correlation with exploration score among agents, and where both adults and children perform significantly better than all agents (see also Fig. 6). Plotting the behavior over time (Fig. 3), we observe that all humans and trained agents increase their overall Entropy and Empowerment throughout learning, while Random quickly stagnates. Information Gain per episode decreases as the most accessible observations get exhausted, but humans maintain it at a significantly higher level than all the agents throughout their learning.

These results suggest exploration in both humans and agents may be advantageously guided by information-theoretic intrinsic motivation, and it may be worth exploring better information-gain related objectives for training agents.

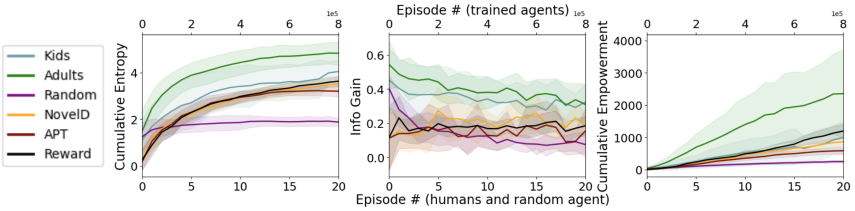

Figure 3: Mean and standard deviation of information theoretic objectives over time.

Analyzing Verbalizations. We investigate whether self-talk might help exploration by examining the relationship between verbalizations and the intrinsic objectives. Following prior works using LLMs for summarizing human data [51], we take transcriptions from each participant and use ChatGPT (`gpt-3.5-turbo`) to classify whether each utterance expresses a question and/or a goal. See Appendix F for details. The children talked significantly more during gameplay than the adults (averaging 240 vs 160 words per session, despite adults often playing for longer). To account for different play durations, we normalize the number of utterances by the total number of timesteps played. Our exploratory analysis found, among just the child participants, a significant correlation between the fraction of verbalizations expressing goals or questions, and the cumulative Empowerment (see Figure 12), with goals exhibiting by far the highest correlation ($r^2 = 0.28, p = 0.005$ unadjusted). This corroborates with prior findings in psychology that self-talk can help direct and focus problem solving in children, especially by focusing their behavior in a goal-directed manner, and findings in AI that agents that generate goals may explore more effectively [33, 28, 13]. Inferring which exploration motivations are implied by the choice of verbalized goal is important future work.

5 Conclusions and Limitations

Our goal in this work is to develop an understanding of human exploratory behaviours in an open-ended environment. To this end, we propose a framework for studying human and agent behaviours in a shared, open-ended environment within Crafter, with various scores for measuring exploration quality. We find that human and agent exploration success consistently correlates with Entropy and Empowerment. On the other hand, we find only human exploration correlates significantly with Information Gain. Although there is a wide spread in the kids' behavior, Information Gain is also the only objective where kids significantly outperform agents. This suggests that perhaps humans make use of Information Gain to guide exploration more effectively than current RL agents are able to, and developing better Information Gain related intrinsic reward functions may be worth exploring. In some preliminary analyses of verbalizations, we find that goal-based utterances in children are significantly correlated with Empowerment. This suggests that goal-setting may be an important component of exploration, with further verbalization analyses left for future work.

Data-set Release Upon publication we plan to release our data-set of human play data and transcripts. It will be available for download and we hope it will be a useful resource for future research.

Limitations. Limitations of our work include the small sample size, limiting broader conclusions about human exploration, as well as the limited number of RL baselines evaluated, which we aim to address in future work. We also note that the analyses on verbalizations were exploratory, and need to be confirmed with a larger sample in a preregistered study format. That said, we hope this work inspires interest in the intersection between cognitive science and AI, laying ground for future work that can collect larger datasets in richer and more naturalistic settings.

References

[1] C. Addyman and D. Mareschal. Local redundancy governs infants' spontaneous orienting to visual-temporal sequences. *Child development*, 84(4):1137–1144, 2013.

[2] A. Aubret, L. Matignon, and S. Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[3] G. Baldassarre, T. Stafford, M. Mirolli, P. Redgrave, R. M. Ryan, and A. Barto. Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Frontiers in psychology*, 5:985, 2014.

[4] A. F. Baranes, P.-Y. Oudeyer, and J. Gottlieb. The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in neuroscience*, 8:317, 2014.

[5] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[6] D. E. Berlyne. Novelty and curiosity as determinants of exploratory behaviour. *British journal of psychology*, 41(1):68, 1950.

[7] F. Brändle, L. J. Stocks, J. Tenenbaum, S. J. Gershman, and E. Schulz. Intrinsically motivated exploration as empowerment. 2022.

[8] F. Brändle, L. J. Stocks, J. B. Tenenbaum, S. J. Gershman, and E. Schulz. Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, pages 1–9, 2023.

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

[10] C. Colas, T. Karch, N. Lair, J.-M. Dussoux, C. Moulin-Frier, P. Dominey, and P.-Y. Oudeyer. Language as a cognitive tool to imagine goals in curiosity driven exploration, 2020.

[11] C. Colas, T. Karch, O. Sigaud, and P.-Y. Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

[12] C. Cook, N. D. Goodman, and L. E. Schulz. Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 2011.

[13] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas. Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692*, 2023.

[14] R. Dubey, P. Agrawal, D. Pathak, T. L. Griffiths, and A. A. Efros. Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*, 2018.

[15] F. Dupuis, W. Yu, and F. M. Willems. Blahut-arimoto algorithms for computing channel capacity and rate-distortion with side information. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 179. IEEE, 2004.

[16] K. E. Exploring exploration: Comparing children with rl agents in unified environments. *rXiv preprint arXiv:2005.02880*, 2020.

[17] K. E. Learning causal overhypotheses through exploration in children and computational models. *PMLR*, 2022.

[18] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[19] S. J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

[20] J. C. Gidley Larson and Y. Suchy. The contribution of verbalization to action. *Psychological Research*, 79:590–608, 2015.

[21] G. Granato, A. M. Borghi, and G. Baldassarre. A computational model of language functions in flexible goal-directed behaviour. *Scientific reports*, 10(1):21623, 2020.

[22] K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

[23] D. Hafner. Benchmarking the spectrum of agent capabilities. *arXiv preprint arXiv:2109.06780*, 2021.

[24] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[25] E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[26] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[27] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[28] E. S. Hu, R. Chang, O. Rybkin, and D. Jayaraman. Planning goals for exploration. *arXiv preprint arXiv:2303.13002*, 2023.

[29] A. S. Klyubin, D. Polani, and C. L. Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer, 2005.

[30] 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.

[31] E. Kosoy, J. Collins, D. M. Chan, S. Huang, D. Pathak, P. Agrawal, J. Canny, A. Gopnik, and J. B. Hamrick. Exploring exploration: Comparing children with rl agents in unified environments. *arXiv preprint arXiv:2005.02880*, 2020.

[32] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

[33] S. W. F. Lee. Exploring seven-to eight-year-olds' use of self-talk strategies. *Early Child Development and Care*, 2011.

[34] C. H. Legare. Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child development*, 2012.

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

[36] E. G. Liquin, F. Callaway, and T. Lombrozo. Developmental change in what elicits curiosity. In *Proceedings of the annual meeting of the cognitive science society*, volume 43, 2021.

[37] H. Liu and P. Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

[38] P. M. The elaboration of exploratory play. *Phil. Trans. R. Soc*, 2020.

[39] M. C. Machado, M. G. Bellemare, and M. Bowling. Count-based exploration with the successor representation. *arXiv preprint arXiv:1807.11622*, 2018.

[40] B. Matusch, J. Ba, and D. Hafner. Evaluating agents without rewards. *arXiv preprint arXiv:2012.11538*, 2020.

[41] B. Meder, C. M. Wu, E. Schulz, and A. Ruggeri. Development of directed and random exploration in children. *Developmental science*, 24(4):e13095, 2021.

[42] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2721–2730. JMLR. org, 2017.

[43] P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2007.

[44] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[45] D. Pathak, D. Gandhi, and A. Gupta. Self-supervised exploration via disagreement. *arXiv preprint arXiv:1906.04161*, 2019.

[46] J. Piaget. Children's philosophies. a handbook of child psychology. *A handbook of child psychology*, 1933.

[47] J. Piaget. *Language and Thought of the Child: Selected Works vol 5*. Routledge, 2005.

[48] S. Pitis, H. Chan, S. Zhao, B. Stadie, and J. Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*, pages 7750–7761. PMLR, 2020.

[49] F. Poli, M. Meyer, R. B. Mars, and S. Hunnius. Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119, 2022.

[50] R. Portelas, C. Colas, L. Weng, K. Hofmann, and P.-Y. Oudeyer. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

[51] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. Robertson, and J. J. Van Bavel. Gpt is an effective tool for multilingual psychological text analysis. 2023.

[52] A. Ruggeri, M. Pelz, E. Schulz, et al. Toddlers search longer when there is more information to be gained. 2021.

[53] J. Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.

[54] L. Schulz. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, 2012.

[55] L. Schulz and E. B. Bonawitz. Serious fun: Preschoolers play more when evidence is confounded. *Developmental Psychology,*, 2007.

[56] D. H. Schunk. Verbalization and children's self-regulated learning. *Contemporary Educational Psychology*, 11(4):347–369, 1986.

[57] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4):301–321, 2003.

[58] C. D. Smock and B. G. Holt. Children's reactions to novelty: An experimental study of" curiosity motivation". *Child Development*, pages 631–642, 1962.

[59] F. Taffoni, E. Tamilia, V. Focaroli, D. Formica, L. Ricci, G. Di Pino, G. Baldassarre, M. Mirolli, E. Guglielmelli, and F. Keller. Development of goal-directed action selection guided by intrinsic motivations: an experiment with children. *Experimental brain research*, 232:2167–2177, 2014.

[60] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. X. Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.

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

[62] A. Ten, P.-Y. Oudeyer, and C. Moulin-Frier. Curiosity-driven exploration. *The Drive for Knowledge: The Science of Human Information Seeking*, page 53, 2022.

[63] P. A. Tsividis, J. Loula, J. Burga, N. Foss, A. Campero, T. Pouncy, S. J. Gershman, and J. B. Tenenbaum. Human-level reinforcement learning through theory-based modeling, exploration, and planning. *arXiv preprint arXiv:2107.12544*, 2021.

[64] L. Vygotsky. Tool and symbol in child development. *The vygotsky reader*, 1994.

[65] L. S. Vygotsky. *Thought and language*. MIT press, 1962.

[66] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021.

[67] T. Zhang, H. Xu, X. Wang, Y. Wu, K. Keutzer, J. E. Gonzalez, and Y. Tian. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34: 25217–25230, 2021.

A Related Work

Exploration in AI. Exploration strategies in AI range in complexity from occasionally taking random actions (*e.g.*, ϵ-greedy) to optimizing complex intrinsic reward functions. Intrinsic rewards are often motivated by objectives such as: increasing entropy, information gain, and/or empowerment. State entropy maximization objectives use the intuition that exploration is motivated by visiting diverse states. Information gain measures the amount of information gained about the environment [35], where exploration is motivated by reducing surprise, developing a better understanding of environment dynamics. Empowerment measures of the number of available options [30, 29], such that maximizing empowerment encourages exploration that increases the agent's control.

Prior work approximating these intrinsic objectives for agent training include count-based exploration bonuses [60, 5], entropy-maximization [25, 37, 66], curiosity-based approaches that encourage agents to take actions that are maximally informative about the environment; for example, by rewarding states or transitions that the agent can not yet predict well [53, 44, 9, 67], leveraging Bayesian networks [27], or network ensembles [45]. Empowerment-based objectives can learn behaviours that have measurable influence over the environment [22, 18, 29]. However, even with sophisticated exploration objectives, RL agents often lag far behind human sample efficiency [40, 24]. In the rare cases that human-level exploration is achieved, this is done by a painstaking amount of hard-coded structure [63].

One explanation for the gap between human and agent behaviour is the vast prior knowledge that humans have due to their life experience. [14] find that removing visual priors greatly reduces human abilities, while agents are unimpacted. [13] propose using large language models as a fuzzy repository of human knowledge as a way of incorporating human priors in RL exploration. That said, prior knowledge by itself is useless without an exploration objective. Our work aims to understand which *objectives* motivate human exploration, and how that can inform intrinsic rewards for agents.

Exploration in Cognitive Science. Dating back to Piaget [46], developmental researchers have conceived of children as active and curious learners who are intrinsically motivated to explore the world in systematic and rational ways [55, 12, 34, 54]; see Schulz [54] for a review. As in the AI literature, there is just as much variety in proposed objectives underlying human curiosity and exploration [62, 38], including perceived novelty [59, 6, 58, 49] which simply suggests that humans are drawn to stimuli that appear more novel, expected learning progress [3, 61, 43] which is the idea that people find it intrinsically rewarding to improve their performance, information gain [36, 52, 1] where the driver of exploration is to gather maximal information about the environment, (or even more specifically the possibility of learning causal relations [59]) maximizing empowerment [7, 8] and totally random exploration more common in younger children [41].

Although humans have been found to be sensitive to many of the above exploration objectives, evidence on what exactly people base their exploration on is inconsistent [49]. Many papers propose that humans are driven by a combination of objectives, such as a desire for both knowledge of task space and competence across that space [4], or that humans find it rewarding to perform above a certain level while simultaneously making substantial learning progress [61]. This mirrors how RL agents also commonly maximize the weighted sum of multiple objective functions, the weighting of which is also often changed during the course of the agent's lifetime of environment interactions [48, 63]. Studies thus far have been limited to highly simplistic and unrealistic environments, typically stateless or with only a couple different states [19] or where participants are asked to choose from a limited set of options [4, 61]. At the more naturalistic end are 3D maze environments where participants can move around [31], but these are still quite limited with navigation being the only available task. Our hope is that studying human exploration in a richer environment can help shed light onto which exploration objectives people actually use and why they are so effective.

Language and Exploration. We also partially use utterances to understand human exploration in this work; while we are not aware of existing work on verbalizations and intrinsic motivation, there are some works on verbalization and problem solving. It has been found across a wide range of studies that verbalization or private speech can be helpful for understanding situations and surmounting difficulties, for instance by focusing attention on important features and discarding irrelevant ones [56, 21], or assisting with coding and retention of information [20]. This suggests that participants with more utterances might explore better because they can better process the flow of information from their environment. Furthermore, it was found that overt verbalization (i.e. thinking aloud) is

especially common for younger children (ages 6-7), and particularly when encountering obstacles [65]. This could suggest a stronger correlation between success in exploration and frequency of utterances for children, but not adults.

B Environment Details

Environment Control. The available actions a either move the player or enable interactions with the environment. Interactions only affect the cell that the player is directly facing, with the "do" action being the most versatile (used to eat, drink, cut tree, mine, fight). Seven additional actions execute a unique action. Note that we remove three of the most complex actions from the original game as they were not feasible to fit on a conventional game controller (see Figure 1, centre). Interaction actions have no effect if the player lacks sufficient prerequisites (*e.g.* `place crafting table` only works if the player has sufficient wood in their inventory). While the original Crafter work contained expert human data, we focus on collecting play data from adults and children who are fully unfamiliar with the game in a reward-free setting so we can observe how they explore in an unknown environment.

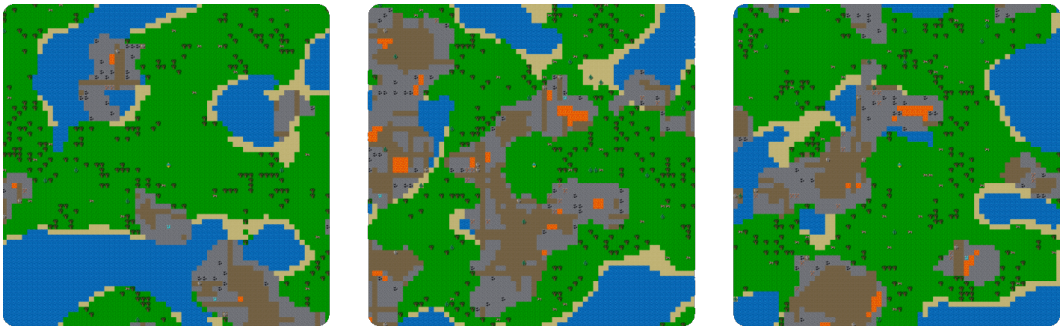

Figure 4: Examples of procedurally generated world maps for Crafter.

We also modify the game to make it easier for human play. First, we slightly lengthen the fraction of an in-game day that is spent in daylight by changing the daylight function from $1 - |\cos(\pi x)|^3$ to $1 - |\cos(\pi x)|^{12}$. We also add an explicit 'Game Over' screen when an episode ends so participants are aware of episode transitions. Lastly, we slightly prune the action space as described in Figure 5 to fit the available actions onto the handheld controller.

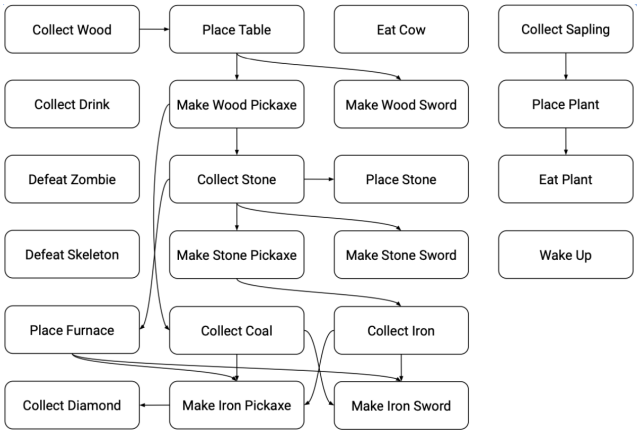

Figure 5: Dependency tree of all the achievements that can be unlocked in Crafter (Figure 4, [23]). Due to our action pruning, `Collect Diamond`, `Make Iron Pickaxe`, `Make Iron Sword`, `Make Stone Sword`, `Place Stone` are not achievable in our setting.

C Summary Statistics

C.1 Overview of Exploration Scores

As there is no single objective measure for "good exploration", we construct five exploration scores for Crafter (Table 1). As the difficulty and number of achievements unlocked is a simple measure of how well a player explores semantically meaningful state changes in Crafter, four measures are achievement-based. The last one, map coverage, is based on task-agnostic physical exploration.

Exploration Score	Definition
Achievement Score	Number of unique achievements (cells of Figure 5) unlocked throughout gameplay.
Weighted Achievement Score	Same as score, but accounts for task complexity by weighting each achieved task by its level in the skill tree (*i.e.* deeper tasks contribute more to the score).
Breadth Score	A measure of how broadly the player has explored the task space by calculating how much progress has been made in a breadth-first traversal of the skill tree (*i.e.* we only count tasks up to and including the first incomplete level of the tree).
Depth Score	A measure of how deeply the player has explored the task space by returning the depth of the deepest task achieved.
Map Coverage Score	Percentage of the game map covered.

Table 1: Description of exploration scores used in our study.

C.2 Summary Statistics

We present summary statistics across all human and agent data. First, Figure 6 shows summary histograms for each measure, showing the normalized density of people and agents on the proposed exploration scores. We find that adults generally score better than children, and there is a wider diversity of performance among both children and adults than any individual agent condition. However, we note that the overall spread of human and agent performances are similar–*i.e.*, it is not the case that humans greatly outperform the agents or vice versa.

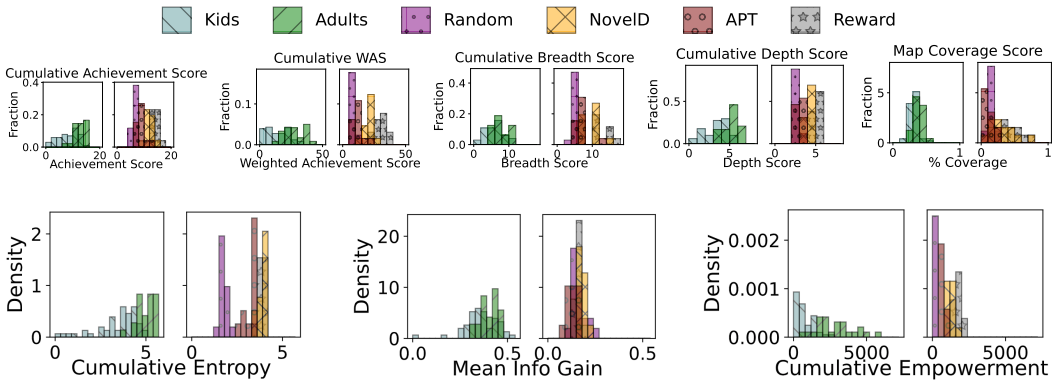

Figure 6: Density histograms for each exploration score and information theoretic objective. Cumulative measures are cumulative across all episodes, while Map Coverage averages across episodes. Left plot for each measure shows human performance, right plot shows agent performance.

Next, we look at how exploration progression over time differs between humans and agents. Figure 7 looks at a random subset of participants and trained agents, plotting the total number of unique achievements they have ever unlocked over time, while Figure 8 plots the mean and standard deviation of three exploration scores aggregated across all agents in each population. Again, adults on average score higher than children. There is also larger variation in children gameplay, with many children who quit playing early, including those who were making rapid progress. Agents are much less sample efficient, reaching similar performance only after over 1000× the number of environment interactions used by humans.

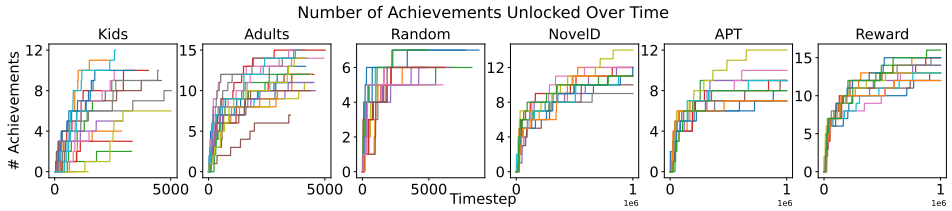

Figure 7: The total number of unique achievements unlocked over time.

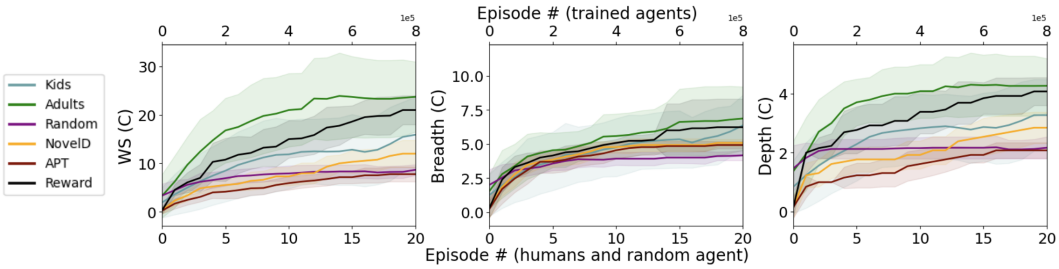

Figure 8: Mean and standard deviation of exploration scores over time.

Finally, we examine whether exploration is more focused on breadth or depth. Figure 9 shows a normalized 2D histogram of exploration breadth and depth through the achievement tree. Notably, children show the clearest correlation between the breadth and depth of exploration, whereas all other groups have participants or agents that are more focused on either breadth or depth. This suggests that children may play in a way that explores diverse and increasingly complex skills similarly.

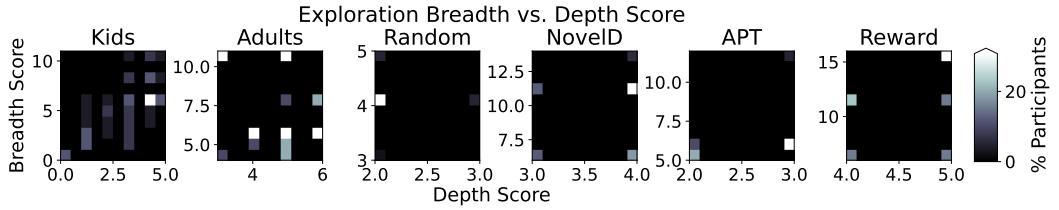

Figure 9: Exploration Breadth vs. Depth through the achievement tree.

D Data Collection Procedure

Participants. We recruited 51 children between the ages of 6-10 years (Mean age: 8.6 years, Female: 19, Male: 32) from the Bay Area Discovery Museum (BADM), as well as 24 adults from the University of California, Berkeley campus ages 18-25 years (Mean age 24.8, Female: 10, Male: 14). Children were tested in a quiet corner at the BADM and adults were tested in a private testing room at Berkeley. The experimenter sat next to the participant at a table in front of a laptop and a small gaming controller. Per IRB guidelines the study lasted a maximum of 20 minutes. Participants who were not able to complete at least one full game round were excluded. We found that 80% of children had video game experience with 64.7% having Minecraft-specific experience, and 79.1% of adults had video game experience with 54.1% having played Minecraft previously.

Data Collection Procedure. In this pre-registered, IRB approved study (AsPredicted reference: 92521), we introduced children and adults to the novel "Crafter" game. We allowed participants to play for up to 20 minutes. Participants were first shown a short tutorial video explaining what each controller button did (Figure 1) and then allowed to play for up to 20 minutes, with the option to quit early. During this time period, the game automatically restarts a new episode any time the player died to a Game Over screen. Participants were not shown any score or given any objective–which was reflected in the wide variation in responses to the question about the point of the game, ranging from

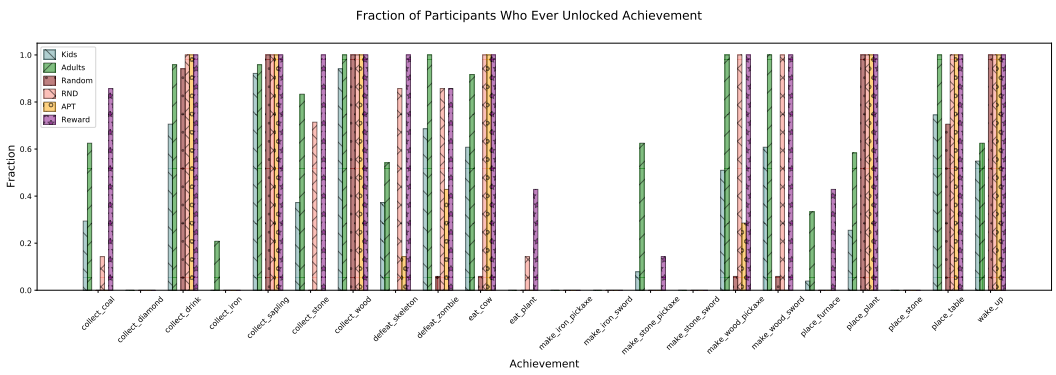

Figure 10: Bar plot of all possible achievements, showing the fraction of each set of participants or agents that unlocked each achievement at least once.

"just have fun" or "try not to rage quit" to "killing the skeletons" or "don't die". All actions taken and the complete world state was recorded for every timestep while playing, along with audio from the participant which was later transcribed manually with timestamps. Due to a lack of consent for audio recording for all participants, this resulted in transcripts from 35 children and 22 adults.

Agent Training Procedure We train three RL agents and use one random agent as baselines. The random agent samples `noop` 47.5% of the time in order to match the average reaction time of the human players, and uniformly samples all available actions otherwise. For trained agents, we compare against state-of-the-art intrinsic RL objectives: NovelD [67] and APT [37]. NovelD incentivizes information gain by providing a large intrinsic reward at the boundary between explored and unexplored regions, using RND [9] as a measure of state novelty. APT uses a particle-based entropy estimator [57] to reward the agent for maximizing state entropy in an abstract representation space. As a measure of best-case performance we also train an agent with the game extrinsic reward function, which reveals the possible set of achievements. The game reward function provides a sparse reward of 1 every time a new achievement is unlocked alongside a small health-based reward every time the agent is hurt or healed. The agent policy input is a simplified semantic representation of the game: the material in the cell the agent is facing, the status, and the inventory. All RL agents are trained with Rainbow DQN [26] for one million timesteps, with ϵ-greedy exploration decaying ϵ from 1 to 0.01 on a linear schedule over 250000 timesteps. We report 12 seeds for each agent.

E Information Theoretic Metrics vs. Exploration Scores

Rather than computing these functions on raw pixel inputs, we construct a state representation s that inherently imbues some prior knowledge by combining the following: the semantic label of the cell the player is currently facing, the contents of their inventory, and the increase in their status from the previous state, if any. This captures aspects of the environment that the player is most likely to be paying attention to and has direct control over, while aiming to avoid meaningless increases in the objectives (*e.g.*, visually novel configurations of the procedurally generated map that are not semantically novel). We use this representation to construct transition tables of each participant's and agent's behaviours, mapping each transition (s, a, s') to the number of times it was experienced.

Entropy. The entropy of the distribution of states visited throughout play is given by

$$\text{Entropy} = \sum_s -p(s)\log(p(s)) \tag{1}$$

Concretely, we compute $p(s)$ as $N_s / \sum_s N_s$, where N_s is the number of times a transition in the transition table started with s. We report the cumulative entropy over the person or agent's total experience. This can be interpreted as a measure of the diversity of all visited states.

Information Gain. We measure the total information gain from all experiences of taking action a from state s as the log count of the total number of times that transition has been made [40].

$$IG(s, a) = \log(1 + N_{(s,a)}), \tag{2}$$

where $N_{(s,a)}$ is the number of times that same action a has been taken given being in state s. We then report the average amount of information gained per transition (accounting for all past experiences) as the overall information gain of a player's experience.

$$\text{Information Gain} = \frac{\sum_{(s,a)} IG(s,a)}{\sum_{(s,a)} N_{(s,a)}} \tag{3}$$

This can be interpreted as a measure of novel transitions encountered, such that the player acquires less information each time they take the same action in a known state. We use the log-count approximation as prior work has found it to perform similarly to more complex measures [40]. A similar alternative is to use the square root instead of the logarithm [8].

Empowerment. Empowerment is defined as the channel capacity of the agent's actuation channel [29]. We compute one-step empowerment as most impactful actions in Crafter are single-step:

$$\text{Empowerment} = \max_{p(a)} I(s';a) = \max_{p(a)} \sum_{\mathcal{A},S} p(s'|a)p(a) \log \frac{p(s'|a)}{\sum_{\mathcal{A}} p(s'|a)p(a)} \tag{4}$$

We use the Blahut-Arimoto [15] algorithm to approximate the channel capacity, as proposed in [29]. We report the cumulative empowerment over the person or agent's total experience. This can be interpreted as a measure of the amount of control the agent has over visited states, or the amount of information the agent could inject into the environment.

F Utterance Analyses

Although children tended to quit playing much earlier than adults, the children also tended to talk more than the adults–with an average of about 240 words uttered during their entire session compared to the adults' average of about 160. Following prior works using LLMs for summarizing human data [51], we take transcriptions from each child participant and use ChatGPT (`gpt-3.5-turbo`) to classify if each utterance is a question, expresses a goal, curiosity, or a realization. To improve accuracy, we also ask the LLM to generate reasoning before making each classification.

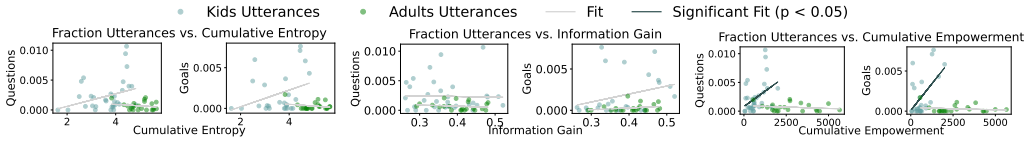

Figure 11: Fraction of verbalized questions and goals vs. cumulative entropy, information gain, and empowerment for adults and children. We find that the relationship between the fraction of uttered goals and empowerment has the highest correlation and largest significance ($r^2 = 0.28, p = 0.005$ unadjusted).

Questions	Questions about the game, such as "How do I move?" or "What is that skeleton doing?"
Goals	Stated goals for the game, such as "I need to get some water."

Table 2: Classes of verbalizations analyzed in our study.

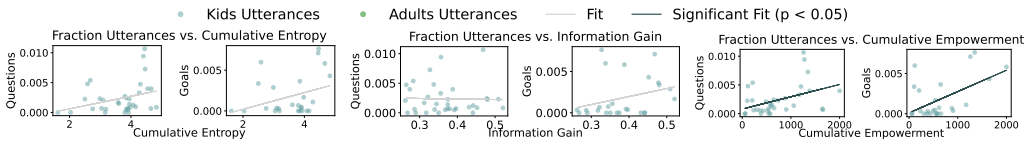

Figure 12: Fraction of verbalized questions and goals vs. cumulative entropy, information gain, and empowerment for children. We find that the relationship between the fraction of uttered goals and empowerment has the highest correlation and largest significance ($r^2 = 0.28, p = 0.005$ unadjusted). No significant relationship was found in the adult data (full plots in Figure 11).