From Child's Play to AI: Insights into Automated Causal Curriculum Learning

Annya Dahmani*† **Eunice Yiu***† **Tabitha Edith Lee**‡ **Nan Rosemary Ke**§

Oliver Kroemer‡ **Alison Gopnik**†

Abstract

We study how reinforcement learning algorithms and children develop a causal curriculum to achieve a challenging goal that is not solvable at first. Adopting the Procgen environments that include various challenging tasks, we found that 5- to 7-year-old children actively used their current level competence to determine their next step in the curriculum and made improvements to their performance during this process as a result. This suggests that children treat their level competence as an intrinsic reward, and are motivated to master easier levels in order to do better at the more difficult one, even without explicit reward. To evaluate RL agents, we exposed them to the same demanding Procgen environments as children and employed several curriculum learning methodologies. Our results demonstrate that RL agents that emulate children by incorporating level competence as an additional reward signal exhibit greater stability and are more likely to converge during training, compared to RL agents that are solely reliant on extrinsic reward signals for game-solving. Curriculum learning may also offer a significant reduction in the number of frames needed to solve a target environment. Taken together, our human-inspired findings suggest a potential path forward for addressing catastrophic forgetting or domain shift during curriculum learning in RL agents.

1 Introduction

Humans are exceptionally remarkable learners, especially when they are faced with challenging tasks. We possess the capacity to craft personalized curricula that shape our experiences in ways that maximize our acquisition of new knowledge and skills (1, 2, 3, 4, 5, 6, 7, 8, 9). Similarly, reinforcement learning (RL) agents also rely on curriculum-based learning to accomplish challenging tasks (10, 11, 12, 13, 14).

When designing a curriculum, it is essential to strike a balance between exploitation (leveraging existing skills and information for rewards) and exploration (discovering new skills and information to enhance decision-making). Studies have demonstrated that humans begin to master this balance between exploration and exploitation from early childhood (1, 15, 16, 17, 18). The question arises: How do humans learn to reason about their own learned capabilities and use this information to bootstrap the future knowledge they need to learn to address their current limitations? Causal learning may be crucial for enabling us to efficiently explore various levels of task difficulty and complexity within an environment (19, 20, 21). In particular, first mastering the causal relations that are involved in a simpler task can allow agents to solve more complex tasks that involve similar

*These authors contributed equally to this work. Correspondence: {adahmani,ey242}@berkeley.edu
†Department of Psychology, University of California, Berkeley, Berkeley, CA, USA
‡The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
§Google DeepMind, London, UK

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

relations (12, 19, 20, 21, 22). This ability contrasts with the abilities of even the most advanced RL agents. We use causal inference and monitor our competence and learning progress to help guide our exploration, rather than randomly varying policies and observing the results. Causal models are well-designed precisely to afford a wide range of novel actions and interventions on the world (23, 24, 25, 26). The ability to collect data from causal interventions can allow an agent to construct a new causal model, leverage that model to make decisions, and repeat this process for improvement. This may be key to improving the performance of RL agents in the future (19, 20, 21, 22). See Appendix Sec. A.1 for related works.

Machines may likewise benefit from structuring their learning through a causal curriculum, improving the speed of convergence and boosting generalization through sequencing training data, developing hierarchical causal models and self-assessment (10, 11, 12, 13, 14, 27). Curriculum learning (10, 11, 12, 27, 28) is relevant and beneficial for a broad range of applications from computer vision (27, 29, 30) and natural language processing (31, 32) to reinforcement learning (13, 14, 33). For instance, successful learning in neural networks has resulted from a curriculum that starts small (34). In the case of RL, while the advantages of improving learning progress (35) through a curriculum are generally recognized (14, 36), it is still unclear how to develop or select a curriculum in the automated and spontaneous way that humans do (28). Tasks are often specified a priori from human domain knowledge, and it is not necessarily clear in what sequence the tasks should be visited within a curriculum.

Our research poses a crucial question that applies to both humans and RL agents: When we approach a goal that is too complex and challenging to achieve outright, how do we evaluate our existing skills and knowledge, and then craft a curriculum of more manageable "sub-tasks"? This curated sequence of sub-tasks serves as a means to construct a causal model, ultimately aiding us in reaching the ultimate goal. This line of work bears interests on both RL and cognitive science. On the cognitive science side, we seek to understand how children autonomously develop a curriculum to attain a goal that is difficult to accomplish at first. In particular, we hypothesize that children are intrinsically motivated to monitor their competence and proceed to acquire higher levels of skill accordingly, even without explicit rewards, and find evidence that this is true. By putting children and RL agents on a level playing field, we also provide a benchmark and point of comparison for human curriculum learning against curriculum learning in RL agents. On the RL front, our initial goal is to evaluate RL agents through a predefined curriculum inspired by recent research (33). We subsequently use this as a baseline against which we assess RL agents that include level competence as an auxiliary reward, inspired by our results with children. Our findings indicate that incorporating this reward significantly enhances the pace of learning and improves convergence. We discover something new about children; that is, we show that children use intrinsic rewards based on level competence. We then show that designing RL agents in a similar way leads to great progress and improvement. We outline the definitions of the terms we used in the paper here in Table 1:

Term	Definition
Level Competence	An indicator of success in an episode of game play on a specific level, reaching 100% when the specific level is successfully solved in that episode
Global Competence	A measure of success in an episode of game play with respect to the target level, reaching 100% if the target level is successfully solved in that episode
Level Advancement	Difference between current level competence and initial level competence within a particular level in the curriculum; it is a measure of how much competence increases or decreases in the same level
Global Advancement	Difference between current global competence and initial global competence across levels in the curriculum; it is a measure of how much competence increases or decreases with respect to achieving the target level
Auxiliary Reward	An internal reward used by the RL agent that augments the extrinisic reward

Table 1: Definitions of the terms we used in this paper

2 Automated Curriculum Learning in Children

Procgen, as introduced by 37, represents a procedurally generated environment that develops a wide range of RL games with varying levels of difficulty. To systematically analyze curriculum learning for both human players and RL agents, we selected a subset of Procgen environments and tailored them by adjusting game difficulty along a single parameter or variable. Our experiments focused on three distinct games: Leaper, Climber, and Heist (Appendix 2, Figure 2). We ask how children scaffold their own learning to solve a difficult level of a game. Our primary focus is to explore whether children employ a systematic and rational approach in constructing a learning curriculum for themselves. To assess this, we observe children engaging with Procgen games (37) of varying difficulty levels and analyze their automated decisions in curriculum development.

2.1 Methods

We have gathered data from a cohort of 22 children (10 females, 12 males), with a mean age of 5.55 years ($\sigma = 0.74$ years) at public museums in the Bay Area, California, USA. Participants were told that they would be rewarded with a sticker if they won a target game level. The target level was set to be too challenging to win in a single attempt - it required some form of multi-step, curriculum-like learning. Children were then given the opportunity to autonomously select different difficulty levels of the game (Levels 1-4, with Level 3 being the target game level). Children could play either until they won the target level, or for up to a total of 10 rounds, whichever came first (see Appendix C.2 for details).

2.2 Automated Curriculum Learning Results

Overall, children made an average global advancement toward the goal of $\mu = 58.4\%$ with a standard error $SE = 11.8\%$ towards the goal (this is computed as the difference in competence between the initial attempt and the final attempt at the target level), which was significantly different from 0% ($t(15) = 4.96$, $p < .001$). This was even though they did not necessarily make positive level advancement between the first and last attempt at the same level ($t(17) = .29$, $p = .78$ in a one-sample t-test compared to $\mu = 0$). After failing the initial goal level, 72.7% participants started their curriculum by selecting an easier level: 9 chose Level 1, 7 chose Level 2, 5 chose Level 3, 1 (of the 7 presented with Level 4) chose Level 4 - the even more difficult level.

Next, we examined the change in levels selected by children across their automated curricula, given their competence on their current level (see Figure 5 in Appendix C.3). A change in level of 0 indicates that the participant chose to remain on the same level, a change in level of -1 indicates that they chose the next easiest level, and so on. We found that children were more likely to remain on the same level or go to easier levels if their progress was low, and were more likely to move to more challenging levels if their competence was high. More specifically, level change (z-scored) was positively predicted by level competence (z-scored) in a linear mixed effects model with participant as a random effect, $\beta = .37$, $t(75) = 3.38$, $p < .01$. Thus, children, like adults in free exploration (8), used competence information to avoid overly easy tasks and advance to more difficult levels that were closer to the target level in their curriculum. However, children neither used their global advancement (z-scored) to guide their level change in the curriculum (z-scored) ($\beta = .029$, $t(77) = -.26$, $p = .80$), nor did they use their level advancement (z-scored) to guide their level change (z-scored) ($\beta = .070$, $t(50) = .49$, $p = .63$). One possible reason is because children were not told to freely explore all levels, but were given the extrinsic goal of solving a difficult level (Level 3). Thus, selecting more difficult levels was appealing even when children did not make much level advancement. This is evident in Figure 8 in Appendix C.3.

Furthermore, we also found that children demonstrated a causal understanding of their curriculum learning. We introduced 7 children to Level 4 which was even more difficult than the target Level 3. If children were selecting their curricula at random, all four levels should be equally likely to be selected, resulting in a 25% chance of selecting Level 4. However, children only selected Level 4 9.68% of the time. This suggests that children recognized that spending extra effort on Level 4 would not cause them to win a reward.

3 Hand-designed Curriculum Learning in RL Agents

3.1 Formulation

Reinforcement learning (RL) agents are trained to solve tasks modeled as a Markov Decision Process (MDP) (38, 39). In this formulation, $M = \langle \mathcal{S}, \mathcal{A}, T, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, T is the transition function, $R \in \mathcal{R} : \mathcal{S} \to \mathbb{R}$ is the reward function, and γ is the discount factor. Within each MDP, the agent acts according to policy $\pi : \mathcal{S} \to \mathcal{A}$, which concerns low-level control. For this work, Proximal Policy Optimization (PPO) (40) is used to train the policy π.

During the training of the low-level control policy, the agent experiences different distributions of MDPs depending on the agent's curriculum. Specifically, the distribution of MDPs is provided by a high-level curriculum selection function $\phi : \Theta \to M$, where Θ are environment parameters that specifies the "difficulty" of the tasks. The goal of the agent is to learn a policy π that can solve a complex target task M_t, analogous to needing to solve the goal level as in Sec. 2. Tasks are considered solved when the episode reward R exceeds a predetermined threshold, R_S. For the purposes of monitoring agent learning, we assume there exists a function $\Lambda : \mathcal{S} \to \mathbb{R}$, called level competence, that describes how far the agent is into the task. Level competence is bounded from 0 to 1.

3.2 Methods

We assess curriculum learning in RL agents through a hand-designed curriculum on the Procgen game of Leaper. Task difficulty Θ is varied within the curriculum by parameterizing the number of water lanes, from 1 (easiest) to 5 (hardest). The goal level M_t has 5 water lanes, consistent with Level 3 (Sec. 2). The agent starts at 1 water lane and advances to a more complex level (i.e., increased water lanes) when the training reward exceeds a threshold. The level competence Λ for Leaper is the vertical lane that the agent has reached, normalized by the total number of lanes between the start and the finish line. Additional method details are available in Appendix D.1.

3.3 Baseline Curriculum Learning Results

We conduct six trials of training the agent using the hand-designed curriculum as specified by ϕ. Representative results are summarized in Figs. 9-10 in App. D.2. Overall, results are poor. None of the six trials successfully finished training. All trials experienced training divergence, where the agent experiences catastrophic forgetting, leading to a permanent regression of reward to zero and a corresponding decrease in level competence. Figure 9 shows a representative time history of an agent exhibiting training divergence by catastrophic forgetting. It has been shown that agents can experience catastrophic forgetting in continual learning problems with distribution shifts (41, 42, 43, 44). In our problem, our hand-designed curriculum induces intentional distribution shifts in order to train the agent in tasks of increasing difficulty. Behaviorally, when this divergence arises, the agent completely forgets the ability to cross lanes, thereby losing the reward signal in such a sparse rewards problem. Note that, for the representative result in Fig. 9c, after divergence, the agent never exceeds 40% mean global competence, which is approximately where the first lane occurs in the distribution of tasks of this difficulty (4.25 water lanes). Without the reward signal, the agent does not receive feedback on which actions are good to take. Although for simpler environments it may be possible to recover, as the difficulty is smaller and random actions may find the goal, for harder environments with multiple lanes, this divergence is unrecoverable. The average number of water lanes at time of divergence is 3.6667 lanes, with the minimum being 2.6875 lanes.

Figure 10 shows the relationship between mean episode training reward and mean episode level competence. Prior to training divergence, level competence is a proxy for reward, although the specific relationship depends on the difficulty of the task.

3.4 Curriculum Learning with Auxiliary Rewards

Motivated by the human capability for learning (i.e., achieving competence) despite not necessarily completing a level, we conducted an additional experiment where the agent uses level competence as an auxiliary reward. As before, six trials were conducted. Specifically, the auxiliary reward function is $\mathcal{R}_i = 2\Lambda/100$, which is calculated at the termination of an episode. For this experiment, we assume the agent has access to level competence, leaving how it should be computed for future work.

When training using auxiliary rewards, results were markedly improved. Five of the six trials were able to successfully finish training by solving the goal level via generalization. On average, generalization occurred at 3.353 million frames. Figure 11 in App. D.2 shows a representative time history for training using level competence as an auxiliary reward where generalization occurred. Although the remaining trial did not solve the target environment, training reached the maximum number of frames allowed without experiencing divergence.

The experimental results with using an auxiliary reward for level competence suggest this prevents training divergence. We confirmed this by examining one particular trial, shown in Fig. 12 in App. D.2, in which the agent *recovers* from catastrophic forgetting using the auxiliary reward of level competence. A difficulty level increase from 4 to 4.0625 water lanes around 4.141 million frames induces catastrophic forgetting, eventually leading to zero (extrinsic) reward and the complete loss of ability to cross lanes around 4.5 million frames. However, unlike the baseline experiments, the agent still obtains an auxiliary reward. Although the auxiliary reward decreases with level competence, it does not reach zero. Despite it being small, the learning signal is sufficient: the agent can still receive feedback for how to increase competence. This leads to a gradual restoration of agent capability, starting with an increase in level competence around 4.75 million frames. Eventually, this restoration yields a gradual increase of extrinsic reward, which is only obtained if the agent can solve the task. Although the restoration is not quick, taking about 1.5 million frames, the agent can nonetheless recover from catastrophic forgetting that would have otherwise led to training divergence.

3.5 Additional Baselines and Comparisons

In addition to Secs. 3.3 and 3.4, we conducted two additional experiments with a stochastic curriculum. Additionally, we conducted a random level baseline that is described in App. D.3.

Stochastic curriculum: selection of random levels In this experiment, 30% of the time, levels are chosen to be selected from the distribution of possible levels, from 1 water lane to 5 water lanes. In the remaining 70%, levels are determined based on the agent's current progress in the curriculum (as in Secs. 3.3 and 3.4). This experiment was conducted six times, and the agent was only trained using the extrinsic reward. In general, this strategy performs poorly. Four trials experienced training divergence. In only one trial was training still active at the end of the trial (getting to 3.0 water lanes). The average progression of the agent was 2.0313 water lanes, significantly less than in Sec. 3.3.

Stochastic curriculum: selection of previously solved levels This experiment is similar to the previous stochastic curriculum experiment, except that instead of selecting from any level, previously solved levels are chosen. This experiment is conducted six times with only the extrinsic reward being used. None of the six trials experienced catastrophic forgetting. However, the agent does not progress through the curriculum as quickly as with an auxiliary reward (Sec. 3.4). The average progression in the curriculum was 2.9583 water lanes. Of the six trials, none could solve the goal level, and only one trial reached 4 water lanes. Therefore, learning and leveraging level competence as an auxiliary reward appears promising as not only a way to prevent forgetting, but also to advance the curriculum.

4 General Discussion

Level competence as an auxiliary reward signal to overcome catastrophic forgetting While children are given the extrinsic reward of a sticker if and only if they solve the goal level of the game, they can visually observe their success in each game play and so assess their competence at each level. Seeing their competence allows for dense reward signals that are set by their performance. They further use this signal to determine the next step to take in the curriculum, leading to an average of 58.4% competence made towards the goal at the end of their automated curriculum learning. Similarly, level competence is a crucial and beneficial signal for RL agents. Our RL training demonstrates that exploiting level competence as an auxiliary reward signal reduces the odds of divergence and catastrophic forgetting.

Future directions Building on our finding that children employed level competence as a metric to assess advancements in curriculum-based learning, our RL experiments incorporating level competence as an auxiliary signal yielded notably improved results. This compelling evidence underscores the pivotal role of level competence as an incentive to overcome distribution shifts induced by a curriculum. These findings suggest that extracting level competence information from high-dimensional image inputs and using it as a reward mechanism has the potential to significantly enhance the efficiency of RL agents in curriculum learning. In fact, it might potentially enable them to autonomously acquire proficiency in curriculum learning. In addition, we hope to further explore whether *automated* curriculum learning necessarily outperforms curriculum learning in a sequence based on strictly incremental difficulty or in a random sequence (45, 46) among both children and reinforcement learning agents.

References

[1] Alison Gopnik, Andrew N Meltzoff, and Patricia K Kuhl. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co, 1999.

[2] Laura Schulz. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, 16(7):382–389, 2012.

[3] Henk G Schmidt, Sofie MM Loyens, Tamara Van Gog, and Fred Paas. Problem-based learning is compatible with human cognitive architecture: Commentary on kirschner, sweller, and. *Educational psychologist*, 42(2):91–97, 2007.

[4] Sofie MM Loyens, Joshua Magda, and Remy MJP Rikers. Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational psychology review*, 20:411–427, 2008.

[5] Louis Alfieri, Patricia J Brooks, Naomi J Aldrich, and Harriet R Tenenbaum. Does discovery-based instruction enhance learning? *Journal of educational psychology*, 103(1):1, 2011.

[6] Faisal Khan, Bilge Mutlu, and Jerry Zhu. How do humans teach: On curriculum learning and teaching dimension. *Advances in neural information processing systems*, 24, 2011.

[7] F Poli, G Serino, RB Mars, and S Hunnius. Infants tailor their attention to maximize learning. *Science advances*, 6(39):eabb5053, 2020.

[8] Alexandr Ten, Pramod Kaushik, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature communications*, 12(1):5972, 2021.

[9] Adrien F Baranes, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in neuroscience*, 8:317, 2014.

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[11] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

[12] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528. PMLR, 2018.

[13] Allan Jabri, Kyle Hsu, Abhishek Gupta, Ben Eysenbach, Sergey Levine, and Chelsea Finn. Unsupervised curricula for visual meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1):7382–7431, 2020.

[15] Eliza Kosoy, David M Chan, Adrian Liu, Jasmine Collins, Bryanna Kaufmann, Sandy Han Huang, Jessica B Hamrick, John Canny, Nan Rosemary Ke, and Alison Gopnik. Towards understanding how machines can learn causal overhypotheses. *arXiv preprint arXiv:2206.08353*, 2022.

[16] Emily Sumner, Amy X Li, Andrew Perfors, Brett Hayes, Danielle Navarro, and Barbara W Sarnecka. The exploration advantage: Children's instinct to explore allows them to find information that adults miss. 2019.

[17] Emily S Sumner, Mark Steyvers, and Barbara W Sarnecka. It's not the treasure, it's the hunt: Children are more explorative on an explore/exploit task than adults. In *CogSci*, pages 2891–2897, 2019.

[18] Emily G Liquin and Alison Gopnik. Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, 218:104940, 2022.

[19] Jane X Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charlie Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, et al. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents. *arXiv preprint arXiv:2102.02926*, 2021.

[20] Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

[21] Chentian Jiang, Nan Rosemary Ke, and Hado van Hasselt. Learning how to infer partial mdps for in-context adaptation and exploration. *arXiv preprint arXiv:2302.04250*, 2023.

[22] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[23] Judea Pearl. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431, 2000.

[24] Clark Glymour, Peter Spirtes, and Richard Scheines. Causal inference. *Erkenntnis*, 35(1-3): 151–189, 1991.

[25] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[26] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

[28] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr, 2017.

[29] Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 105–121. Springer, 2016.

[30] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166, 2021.

[31] Conglong Li, Minjia Zhang, and Yuxiong He. Curriculum learning: A regularization method for efficient and stable billion-scale gpt model pre-training. 2021.

[32] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

[33] Qiyang Li, Yuexiang Zhai, Yi Ma, and Sergey Levine. Understanding the complexity gains of single-task rl with a curriculum. In *International Conference on Machine Learning*, pages 20412–20451. PMLR, 2023.

[34] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[35] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.

[36] Clément Romac, Rémy Portelas, Katja Hofmann, and Pierre-Yves Oudeyer. Teachmyagent: a benchmark for automatic curriculum learning in deep rl. In *International Conference on Machine Learning*, pages 9052–9063. PMLR, 2021.

[37] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020.

[38] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[39] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[41] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[42] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[43] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

[44] Cuong V Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

[45] Richard C Atkinson. Optimizing the learning of a second-language vocabulary. *Journal of experimental psychology*, 96(1):124, 1972.

[46] Nate Kornell and Janet Metcalfe. Study efficacy and the region of proximal learning framework. *Journal of experimental psychology: Learning, Memory, and Cognition*, 32(3):609, 2006.

[47] Laura S Cubit, Rebecca Canale, Rebecca Handsman, Celeste Kidd, and Loisa Bennetto. Visual attention preference for intermediate predictability in young children. *Child development*, 92 (2):691–703, 2021.

[48] Celeste Kidd, Steven T Piantadosi, and Richard N Aslin. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, 7 (5):e36399, 2012.

[49] Celeste Kidd, Steven T Piantadosi, and Richard N Aslin. The goldilocks effect in infant auditory attention. *Child development*, 85(5):1795–1804, 2014.

[50] Lev S Vgotsky. Mind in society: The development of higher psychological processes. *London: Harvard University Press*.

[51] Julia A Leonard, Skyler R Cordrey, Hunter Z Liu, and Allyson P Mackey. Young children calibrate effort based on the trajectory of their performance. *Developmental Psychology*, 59(3): 609, 2023.

[52] Azzurra Ruggeri and Tania Lombrozo. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216, 2015.

[53] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR, 2018.

[54] Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33:13049–13061, 2020.

[55] Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-guided adversarial environment design. *Advances in Neural Information Processing Systems*, 34:1884–1897, 2021.

[56] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning*, pages 17473–17498. PMLR, 2022.

[57] Kevin P Murphy. Active learning of causal bayes net structure. 2001.

[58] Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pages 863–869. Citeseer, 2001.

[59] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

[60] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *arXiv preprint arXiv:1206.3250*, 2012.

[61] Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. Active structure learning of causal dags via directed clique tree. *arXiv preprint arXiv:2011.00641*, 2020.

[62] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733. PMLR, 2018.

[63] AmirEmad Ghassami, Saber Salehkaleybar, and Negar Kiyavash. Interventional experiment design for causal structure learning. *arXiv preprint arXiv:1910.05651*, 2019.

[64] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

[65] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR, 2017.

[66] Erik M Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. *arXiv preprint arXiv:1810.11867*, 2018.

[67] Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*, 2021.

A Related Works

Our work relates to prior work in cognitive science, curriculum learning, and active causal learning.

Automated Learning in Children Related work in cognitive science suggests that child learning resonates with the Goldilocks principle: children opt for information that is neither too easy nor too complex, but “just right” and moderately predictable (47, 48, 49). Furthermore, children as learners seem capable of monitoring the "zone of proximity" between their current capabilities and the goal at hand; enabling them to progress from what they cannot do to what they can learn to do with the interventions of an adult or a teacher (50). Infants allocate their visual attention based on surprise, predictability and learning progress of the environmental structure (7). 4- to 6-year-old children use their improvement over time to decide whether to persist on a challenging goal on their own (51). And by age 7, children ask questions that yield high information gain and more constraint-seeking questions when the problems are difficult (52). However, there are no systematic studies showing that young children can indeed construct an appropriate curriculum in order to master complex goal-directed tasks, nor studies that would allow a comparison between children and artificial agents.

Curriculum Learning 10 introduced the concept of curriculum learning and proposed that an effective approach to learning involves the provision of examples that strike a balance between being neither overly simple nor excessively challenging. This concept is further supported by theoretical proofs in reinforcement learning (33). Various metrics have been proposed to measure task difficulty, including the transferability of models trained on other tasks to the current task (53), complexity-driven progress, and loss-driven progress (28). 11,9 proposed a framework for automatic curriculum learning through self-play, while 12 proposed a framework utilizing automatic goal generation. 54 introduced the Unsupervised Environment Design paradigm, formalizing variation in environments by parameters, along with the PAIRED algorithm that produced an implicit curriculum. Prioritized Level Replay (55) also implicitly learns a curriculum for training an RL agent not through environment adaptation, but through judicious selection of past levels to replay based on learning potential. Later, 55 unified these two concepts under the Dual Curriculum Design framework with Robust PLR as a representative algorithm, and 56 introduced ACCEL, an evolutationary-based approach for editing environments to form a curriculum. Instead of proposing a new automatic curriculum learning framework, we draw inspiration from child development studies. We use level competence as an auxiliary reward within a curriculum inspired by 31. Our results show that employing auxiliary rewards mitigates RL agent catastrophic forgetting, leading to faster and improved convergence.

Active Causal Learning In RL curriculum development, we perform environmental interventions to enable swift achievement of learning objectives, resembling active causal learning (57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67). However, curriculum learning in RL differs from active causal learning in that it dispenses with explicit causal structure learning and prioritizes targeted causal interventions, disregarding non-essential variables or edges.

B Procgen Environments

There are 12 different Procgen environments. Figure 1 taken from 37 showing all available Procgen environments.

B.1 Our Adapted Procgen Environments

The goal of each game is provided in Table 3. Each game includes three levels: Level 1 is the easiest level, whereas Level 3 is the most difficult level. Participants were given the goal to win Level 3. In addition, a level more difficult than Level 3, Level 4, was presented to a random subset of the participants. Participants did not need to learn Level 4 to achieve their goal of winning Level 3. Table 2 describes the levels of the different games. We also include visual examples of the levels in Figure 3 for the game Leaper across its two different axes, log lanes and car lanes.

C Automated Curriculum Learning in Children

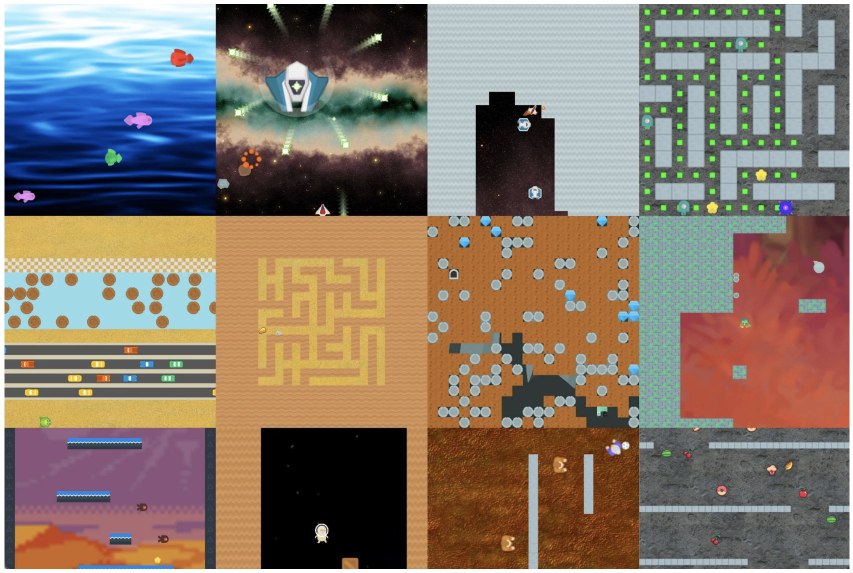
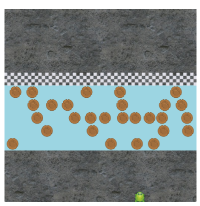
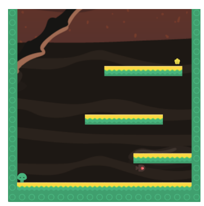


Figure 1: An illustration of all Procgen environments

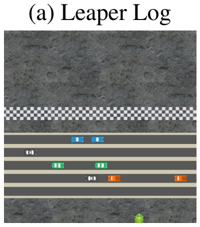


(a) Leaper Log

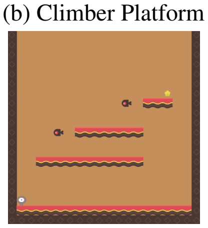


(b) Climber Platform

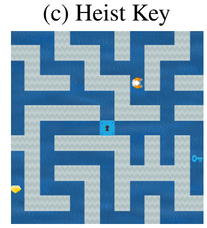
(c) Heist Key



(d) Leaper Car



(e) Climber Enemy

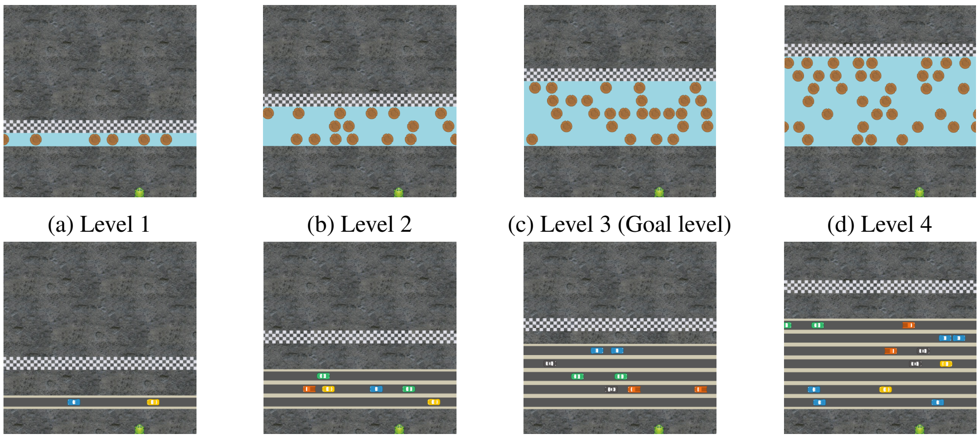


(f) Heist Size

Figure 2: We selected 3 Procgen environments with 2 different axes each and varied the level of difficulty across the axes within a game. (a)-(f) show the goal levels for the selected Procgen games.

Table 2: Description of levels used in selected Procgen games of Leaper, Climber, and Heist. These games were chosen to vary certain aspects of each game based on a particular axis. The levels increase in difficulty from Level 1 through Level 4. The goal level to complete is Level 3; Level 4 is the most challenging level, which is not necessarily needed to be solved to complete Level 3.

Game	Axis	Level 1	Level 2	Level 3 (Goal level)	Level 4
Leaper	Log	1 log lane	3 log lanes	5 log lanes	7 log lanes
Leaper	Car	1 car lane	3 car lanes	5 car lanes	7 car lanes
Climber	Enemy	0 enemies	1 enemy	2 enemies	3 enemies
Climber	Platform	1 platform	2 platforms	3 platforms	4 platforms
Heist	Size	small	medium	large	extra large
Heist	Key	1 key/lock	2 keys/locks	3 keys/locks	4 keys/locks



(a) Level 1 (b) Level 2 (c) Level 3 (Goal level) (d) Level 4

(e) Level 1 (f) Level 2 (g) Level 3 (Goal level) (h) Level 4

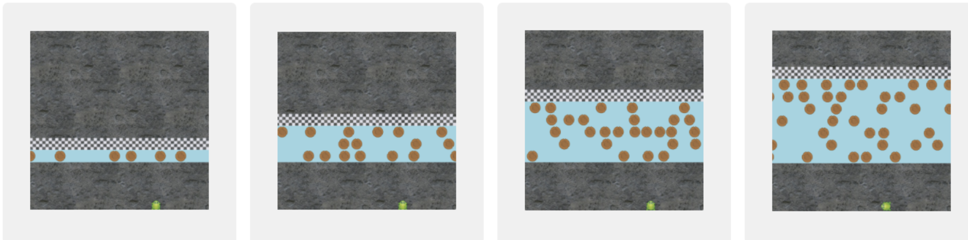
Figure 3: (a)-(d) Levels of difficulty for the game Leaper with the number of log lanes as the difficulty axis. (e)-(f) Levels of difficulty for the Procgen game Leaper with the number of car lanes as the difficulty axis.

Game	Goal
Leaper Log	Cross the finish line and avoid going in the water as log lanes increase.
Leaper Car	Cross the finish line and avoid getting hit by a car as car lanes increase.
Climber Enemy	Reach the jewel and avoid enemies as number of enemies increase.
Climber Platform	Reach the jewel and avoid a single enemy as number of platforms increase.
Heist Size	Reach the final jewel by getting key and unlocking a lock as size of maze increases.
Heist Key	Reach the final jewel by getting key and unlocking locks as number of locks and keys increase.

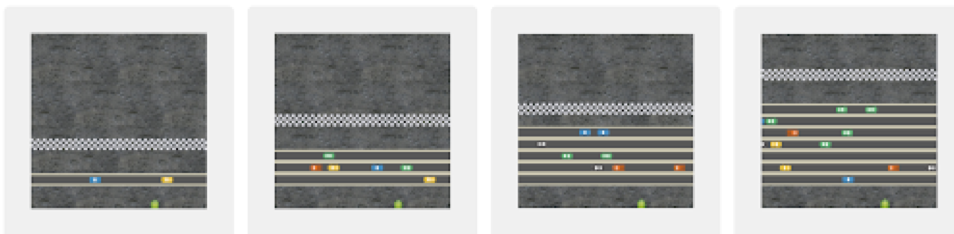
Table 3: We provide the goals for each of the Procgen games. Participants were never told the rules of the game and had to learn how to win the game through their own learning.

C.1 Procgen Difficulty Levels

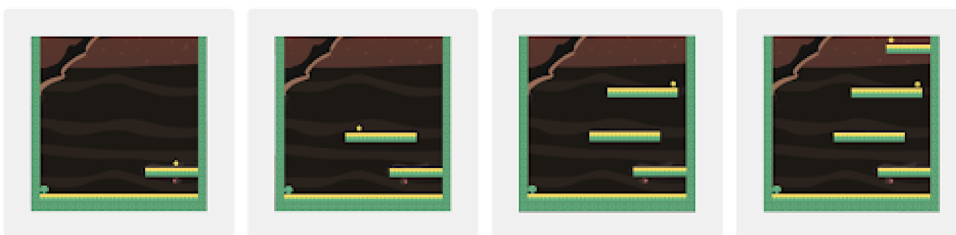
We provide the levels that were shown to participants in the different Procgen games. There were four levels in varying difficulty, with Level 3 being the goal level.



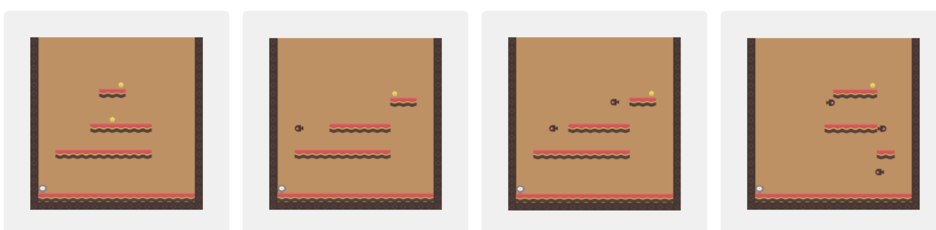
(a) Levels 1-4 for Leaper Log



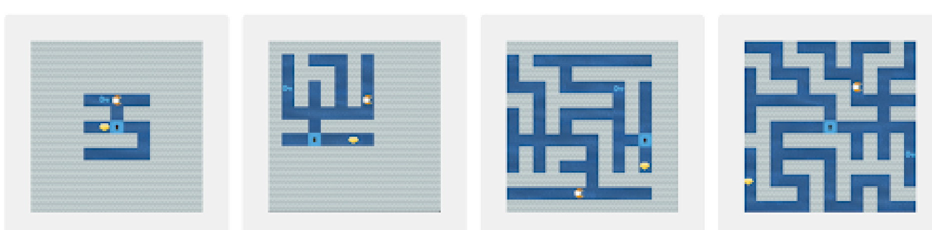
(b) Levels 1-4 for Leaper Car



(c) Levels 1-4 for Climber Platform



(d) Levels 1-4 for Climber Enemy



(e) Levels 1-4 for Heist Size

C.2 Experimental Procedure

The study was performed on a computer. Participants were randomly assigned to play one of the four Procgen games. Participants first underwent a familiarization trial where they practiced exploring an empty environment of the game with a video-game controller (e.g., they explored Leaper without any lanes or obstacles between the starting point and the finishing line).

Next, participants were told to play the goal level of the game in which they would be rewarded a sticker if they won. The rules of the game were not revealed to the participants. Since the experiment aimed to measure curriculum learning in a case where the goal was too challenging to be attained outright, participants must fail the goal level to continue with the experiment. If a participant passed the goal level, experimenters reassigned them to play a different Procgen game. After the participant failed the goal level, the experimenter restated to them that the goal was to win that particular level, and once they did, they would get a sticker. The experimenter asked if the participant could tell them how the game worked and what the participant would have to do to get the sticker.

Then, participants were asked which level of the game they wanted to play next and why. They were shown images of varying difficulty levels that quantitatively varied along a single game axis on a tablet, similarly to the ones in Figure 3. Specifically, they were told the increasing variable of the axis along the levels but were not explicitly told the relative difficulty of each level. Participants were presented with a total of three levels of difficulty: the goal level and two levels that were incrementally easier than the goal level. A subset of 7 participants were further presented with an additional level that was unnecessarily more difficult than the goal level.

Participants received verbal and visual feedback about their performance after playing their selected levels each time. After every other trial, they were reminded to focus on the goal level in order to win a sticker.

This procedure continued until the participant passed the goal sticker level or up to a total of 10 trials, whichever was earlier. Participants who did not pass the goal level by the tenth trial were invited to play the goal level again and then the experiment concluded.

C.3 Results

Figures 5 and 6 show the level adjustments children made when selections of a level that is unnecessarily challenging beyond the goal level are not or are included. In the latter case, since there is a total of 4 difficulty levels in this case, the maximum possible absolute level change is 3. Whereas one participant made a level adjustment from Level 4 to Level 1, no participant made a level adjustment from Level 1 to Level 4. Figure 7 shows the level adjustments made within each level based on children's current level competence.

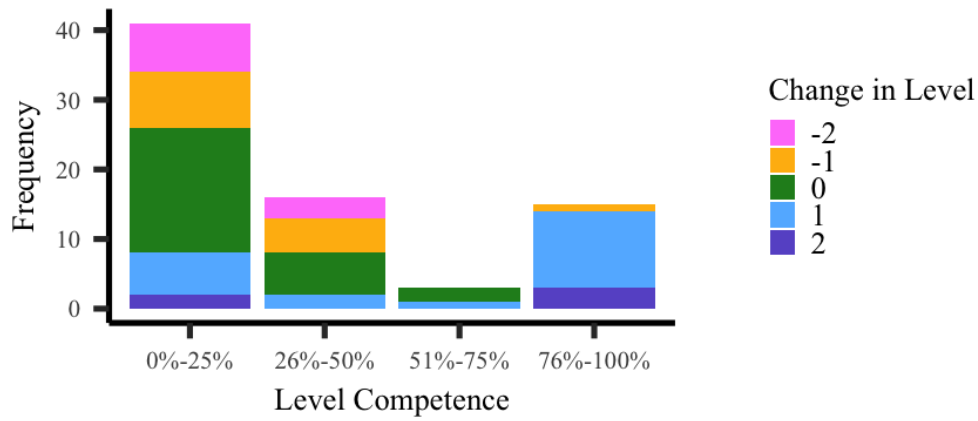


Figure 5: Level adjustments based on children's level competence on the current level. The x-axis measures current level competence as a percentage; the y-axis shows subsequent level adjustment frequency. A level change of 1 implies choosing a game one level harder, while -1 means opting for one level easier. This figure includes participants' selections of levels easier than or equivalent to the goal level. Overall, children tend to remain on the current level when their level competence is less than 75%. However, upon reaching a 76% completion rate, children often transition to more challenging levels. Conversely, when children demonstrate less than 50% level competence, they are more inclined to return to easier levels. Thus, children adapt their learning trajectory based on their performance.

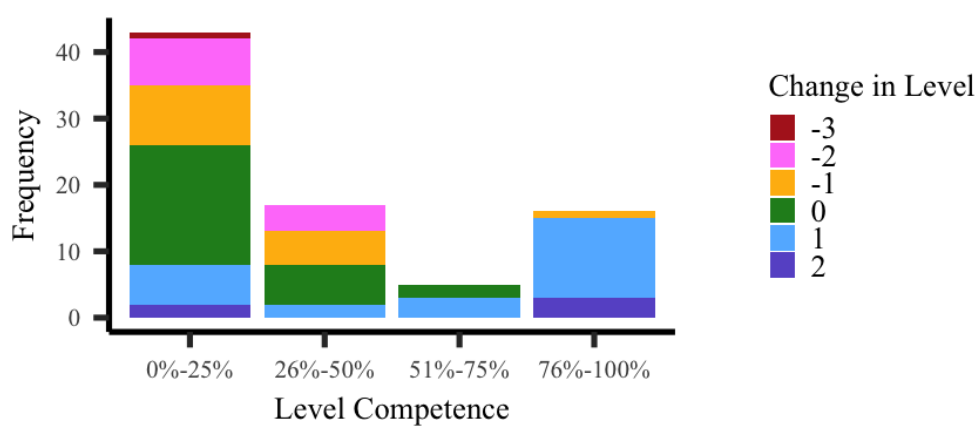
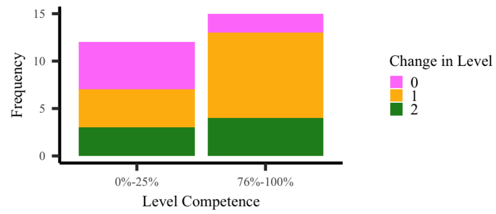
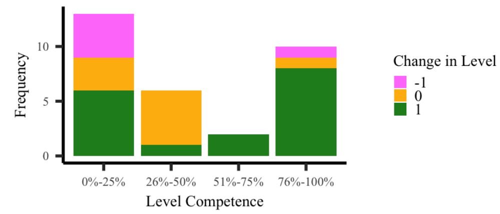


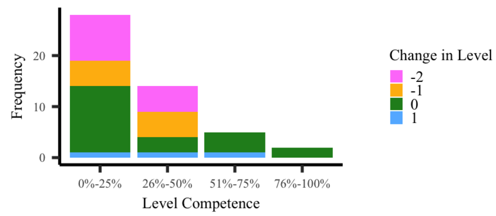
Figure 6: Level adjustments based on children's percentage of competence on the current level with the inclusion of the extra challenging Level 4.



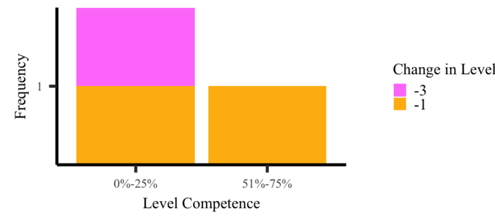
(a) Level 1



(b) Level 2



(c) Level 3 (Target)



(d) Level 4 (Extra Challenging)

Figure 7: Children made level adjustments based on competence within each level. Note that only 2 participants selected Level 4 at any point of the curriculum.

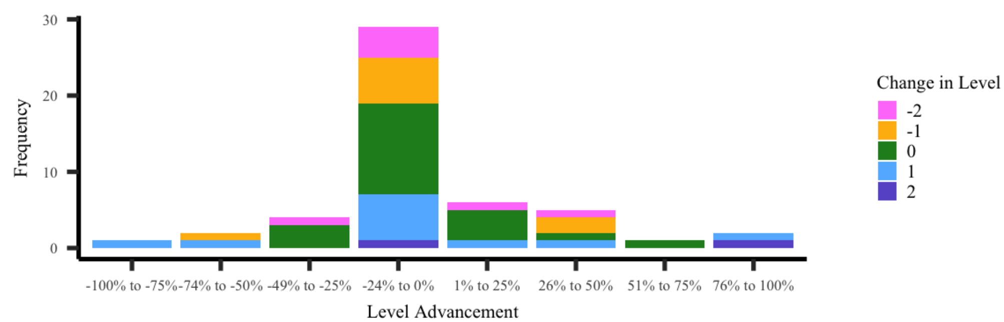


Figure 8: There was no significant correlation between children's percentage of advancement within their current level and children's level adjustment. Most of the time children were making -24% to 0% level advancement, and yet many of them still opted to remain on the same level or select a more challenging level

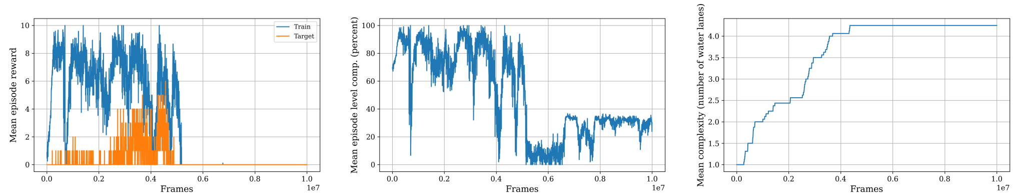
D Hand-designed Curriculum Learning in RL Agents

D.1 Additional Method Details

Motivated by recent work that suggests tasks should be solved in an easiest-to-hardest fashion for better sample difficulty (33), our hand-designed curriculum function ϕ starts at 1 water lane (easiest) and trains the agent until a mean episode reward of 9 (out of possible 10) is achieved. Leaper is a sparse rewards problem, where a reward of 10 is only provided when an episode is solved (and 0 otherwise). Specifically, we run $w = 16$ parallel tasks, initially starting all 16 at 1 water lane. Then, the curriculum function ϕ increments Θ by $1/w$ (e.g., 1/16) to advance to a harder distribution of tasks. After each PPO update, the agent is evaluated against the target task M_t. If the agent successfully solves M_t, training concludes. Otherwise, training continues until a maximum number of frames $f_m = 10 \times 10^6$.

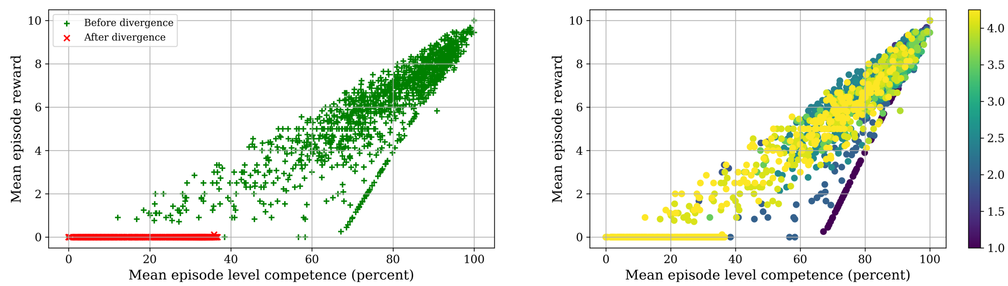
D.2 Results

This section contains more analysis of the experiments conducted in Sec. 3.3 and Sec. 3.4.



(a) (b) (c)

Figure 9: Representative results for baseline curriculum learning with an RL agent. (a) Time history of mean episode reward obtained by the agent in both the training levels and the goal level. Training divergence from catastrophic forgetting results in a regression of reward to zero, which occurs around 5.18 million frames. (b) Time history of mean level competence in the training tasks. (c) Time history of the training task difficulty, as measured by number of water lanes.

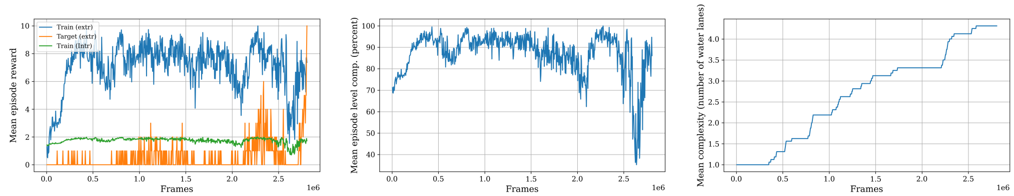


(a) Reward by level competence, colored by divergence. (b) Reward by level competence, colored by difficulty.

Figure 10: Level competence is a proxy for reward. (a) Prior to training divergence, mean episode training reward is proportional to mean episode level competence. After training divergence, this relationship no longer holds: the reward remains at zero regardless of level competence. (b) Before training divergence, the exact relationship of reward and level competence depends on the task difficulty. The easiest task (1 water lane, dark blue) has the greatest slope, since changes in level competence yield relatively greater mean training reward. The slope decreases as difficulty increases because tasks have increasingly more vertical lanes before the goal.

D.3 Random Level Baseline

In this experiment, we evaluate a baseline using a random curriculum: where the level is randomly selected from the possible distribution of levels. We conduct this experiment six times. We consistently saw poor learning performance, and the reward obtained on the goal level was generally zero. This baseline is quite challenging as it is difficult to obtain a consistent learning signal from the extrinsic reward alone.



(a) (b) (c)

Figure 11: Representative results for curriculum learning with an RL agent while training on level competence as an auxiliary reward. (a) Time history of mean episode reward obtained by the agent in both the training tasks and the target task. The intrisic reward used for training is also shown, which is derived from the agent's level competence. The agent begins to generalize to the target task around 2.8 million frames, eventually leading to solving the target task in 2.806 million frames. (b) Time history of mean level competence in the training tasks. (c) Time history of the training task difficulty, as measured by number of water lanes.

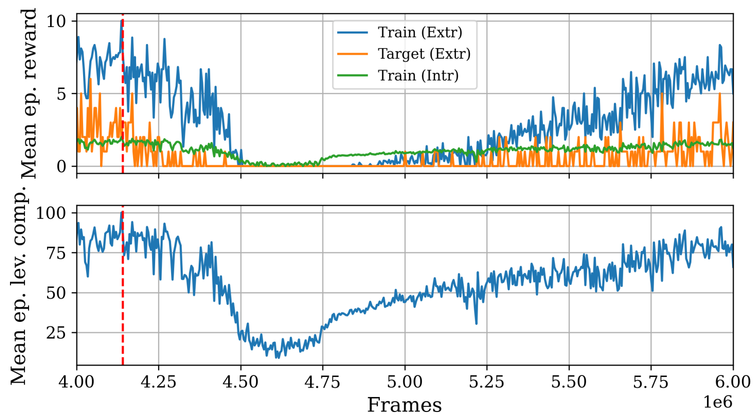


Figure 12: Training using level competence as an intrinsic reward can recover from catastrophic forgetting that would have otherwise led to training divergence. The vertical red line marks the increase in task difficulty from 4 to 4.0625 water lanes, precipitating (recoverable) catastrophic forgetting.