Tiny Paper on Empowerment and Causal Learning

Annya Dahmani
UC Berkeley
adahmani@berkeley.edu

Aly Lidayan
UC Berkeley
dayan@berkeley.edu

Alison Gopnik
UC Berkeley
gopnik@berkeley.edu

We are submitting this tiny paper to showcase our early research ideas on empowerment and to foster discussions and collaborations in the cognitive science and artificial intelligence communities.

In this work, we are interested in bridging causal learning in humans and reinforcement learning (RL) in agents. Earlier work in cognitive science on causal learning has found that both adults and children are strongly motivated to discover causal structure in their environment. Meanwhile, research in RL has focused on learning to maximize rewards without explicitly attempting to discover causal structure. We hypothesize that the concept of "empowerment" in reinforcement learning can provide a bridge between reinforcement learning and causal learning. "Empowerment" is an intrinsic reward that involves maximizing the mutual information between an agent's actions and outcomes in the world, and so maximizing the agent's ability to control the environment, rather than maximizing particular external rewards. This ability to control the environment is also at the heart of "interventionist" accounts of causality and causal learning (e.g. Woodward, 2005; Pearl 2000, 2009). From the machine learning perspective *Empowerment* may thus be an especially promising intrinsic motivation for RL agents to discover causal structure. From the cognitive science perspective we will explore whether human causal learning can be explained by a drive to maximize empowerment, compared to other forms of novelty-seeking drives.

We will build on ongoing work from Gopnik (2024) that has started examining ideas of empowerment as causal learning. Recent work from Rule et al. (2023) have described children's play as optimizing for empowerment gain. Du et al. (2023) have used the Crafter environment to study intrinsic motivation in humans and agents, suggesting that empowerment may be an important motivator for open-world exploration.

This research aims to add another component to this previous work that we believe is important: causal learning. On the cognitive science side, we plan to run empirical studies on children and adults to investigate if causal judgments track empowerment relationships. On the reinforcement learning side, this work seeks to address the gap in causal learning for RL. We plan to investigate if an agent trained with empowerment can discover more causal relationships in an environment compared to agents trained with other intrinsic motivations such as information gain. Moreover, we will ask if such an agent can learn to solve tasks more efficiently than RL systems that learn predictive rather than causal world models (Hafner et al. 2023).

We have already started coming up with rough ideas for human study designs. Some ideas include studies that disentangle empowerment from other information-seeking objectives and use it to predict human decision-making and judgements about causal relationships. We have further ideas about preference and performance when the magnitude of novelty and empowerment are varied. We also plan to explore how causal learning unfolds when participants in an environment either act as an agent engaging in interventions, see another agent engage in interventions, or see outcomes of events happen with no agent involved.

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

Relevant work

[1] Rule, J., Goddu, M., Chu, J., Pinter, V., Reagan, E. R., Bonawitz, E., & Ullman, T. (2023). Fun isn't easy: Children choose more difficult options when "playing for fun" vs."trying to win".

[2] Du, Y., Kosoy, E., Dayan, A., Rufova, M., Abbeel, P., & Gopnik, A. (2023, November). What can AI Learn from Human Exploration? Intrinsically-Motivated Humans and Agents in Open-World Exploration. In NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems.

[3] Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., & Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. Nature Human Behaviour, 7(9), 1481-1489.

[4] Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005, September). Empowerment: A universal agent-centric measure of control. In 2005 ieee congress on evolutionary computation (Vol. 1, pp. 128-135). IEEE.

[5] Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. Developmental psychology, 48(5), 1215.

[6] Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. Cognition, 115(1), 104-117.

[7] Zeng, Y., Cai, R., Sun, F., Huang, L., & Hao, Z. (2023). A Survey on Causal Reinforcement Learning. arXiv preprint arXiv:2302.05209.

[8] Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. Developmental psychology, 43(4), 1045.

[9] Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. Cognition, 120(3), 341-349.

[10] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104.

[11] Woodward, J. (2005). Making things happen: A theory of causal explanation. Oxford university press.

[12] Pearl, J. (2009). Causality. Cambridge university press.

[13] Gopnik, A. (2024). Empowerment as Causal Learning, Causal Learning as Empowerment: A bridge between Bayesian causal hypothesis testing and reinforcement learning.