*A selection of "big" research questions I have:*

## Hierarchical abstraction in continual reinforcement learning, and collective learning

Continual RL agents must learn and adapt in environments that are "much much larger than the agent" where they can't hope to visit every state even once in their lifetime. This creates pressure to learn efficient abstractions from limited data. How can we develop a theory of hierarchical abstraction that allows continual RL agents to efficiently learn and transfer knowledge across different levels of task complexity?

For instance, in real-world scenarios like robotic process automation or autonomous driving, agents encounter new tasks and variations constantly. Can we design architectures that learn reusable skills (like "making coffee" or "navigating an intersection") that remain useful across diverse scenarios? This connects to the idea of option discovery in hierarchical RL, but in a continual learning context where the set of useful options may itself evolve over time. Similarly, concerning the memory constraints and abstractions, how do we make sure that certain options like "being able to close a door" or "walking to the cafe and getting coffee" are retained whilst learning how to "solve math problems"?

Drawing inspiration from how humans form abstractions could be valuable here. Humans have a natural ability to perform Bayesian inference at different levels of abstraction— low-level (day to day decisions aligning with some active reward function) and high-level (communal interactions that impacts multiple individuals). The lack of research and attention towards making agents learn to adapt to tasks in active environments where preferences and states may be defined by hierarchical MDPs (such as MOMDPs and HM-MDPs), pertaining to different levels of non-stationarity and abstraction, is one of my main motivations.

With a similar notion, Abel et al. (2024) provide an informal definition of a continual reinforcement learning agent as one that searches over a possible set of (varying) history-based policies endlessly. For example, considering a neural network to consist of multiple agents (hyperparameters and network parameters) that search over the various possbilities (basis), via SGD for example, that may be capped by physical, problem-specific constraints is a continual learning agent. This intuition builds upon the notion of collective intelligence among various agents that are trying to solve a common goal. A moving target of such a goal, although unstable, would instead lead to endless adaptation.

As an aside, from a continual agent's POV, humans would be the main drivers of active non-stationarity, and it's even harder to assume our action distribution. But, if we can learn to classify actions taken by humans (in a closed system) and humans become a part of the active Markov game (as non-focal agents), could an active equilibrium can be guaranteed (Kim et al., 2022) in the long term?

With this, I would like to investigate different sparse representation learning methods in the context of online MBRL. In particular, I am interested in addressing the following "lower" level questions: how can we work with abstractions efficiently in an online model-based RL setting, particularly for applications with high-dimensional observations? how can these methods be leveraged to mitigate catastrophic forgetting in continual learning scenarios? Particular to MA settings: how do agents decide which abstractions to learn from their continual interaction streams?

Additionally, I would like to build upon existing frameworks for evaluating the quality of learned skills from a mathematical standpoint that focuses on skill acquisition, reacquisition, and extinction be managed over time?

## Multi-agent coordination in non-stationary environments

As pointed by Doina Precup, very little of the non-stationarity that people experience in their lives is from literal changes to the physics of the world but rather from "the changing behaviours of the many other agents in the world." This suggests that multi-agent reinforcement learning (MARL) provides a natural framework for studying continual RL. Indeed, a key motivational question behind this research direction is "How can we rethink exploration strategies in large, non-explorable environments using information-theoretic methods?"

This connects to the idea of active Markov games, where each agent's learning creates non-stationarity from the perspective of other agents. Can we develop algorithms that converge to an "active equilibrium" - a stationary periodic distribution in the joint policy space - while still allowing for continual adaptation? This might involve agents learning to shape the learning of other agents.

In multi-agent settings, how can agents learn to actively shape the non-stationarity of their environment through strategic interaction with other agents? What game-theoretic principles can guide the design of algorithms for mutual adaptation? It's likely that agents need to specialize their abstractions while retaining enough overlap to allow for communication. Can we effectively combine multiple representations (e.g., skill embeddings, latent embeddings, successor representations) to create more robust MARL agents in active environments? How do these different representations complement each other in handling various levels of non-stationary environments? What new composition operators are needed for concurrent execution in temporally extended models?

In this context, how can we design decentralized learning algorithms that balance specialization and synergy across agents while remaining robust to environmental changes? The agents could potentially "teach other agents as they learn so that each learns to better perform tasks that are required of them." What mechanisms would allow for efficient knowledge sharing without creating bottlenecks or vulnerabilities? To this end, how can we design architectures that can plan an agent's POV, while incorporating how other agent's learn in a meta-learning model? (This setting does not include providing the focal agent with the actions other agents take.)

This balance between specialization and commonality in abstractions is crucial for effective multi-agent learning and interaction. Understanding abstraction in multi-agent contexts emerges as a key direction for further research, especially in how it relates to the agents' ability to cooperate, communicate, and adapt in a shared, dynamic environment.

## Convergent realism in CRL; shared model; causal learning

A very optimistic area of research closely related to the notion of "convergent realism" (Hardin and Rosenberg, 1982; Putnam, 1982; Cao and Yamins, 2024) that might seem counter-intuitive, is the hypothesis that a true world model exists. Towards ensuring a stable and efficient continual learning system, the notion of a "platonic world model", i.e. a shared statistical model of the environment, is ought to be established formally. Considering the social learning process that humans go through, one possible intuition to verify the platonic world model hypothesis would be to consider a multi-agent continual setting where each agent's model of the environment (e.g. neural network params) could be stiched together (Bansal et al., 2021) and maintain compatability. Following recent works (Huh et al., 2024; Richens and Everitt, 2024) in the domain of multimodal supervised representation learning which suggest that networks trained on different data with different objectives converge to a shared statistical model of reality, I would like to build upon the definition of continual reiforcement learning, that challenges the view of "platonic world model" in the case of active non-stationarity, while proving it for forming a causal world model in the continual setting. Building a formal intuition for the necessity of a causal model for decision tasks in CRL is still unclear.

## Task-information retrieval from causal structure

One of the key problems pertaining to AI and ethics is alignment. In particular, one of the open questions that arises from the reward hypothesis (Silver et al., 2021) is – how can an agent retrieve the context from a low-dimensional "reward" in an unsupervised manner? One ways to think about reward is the accumulated return, which provides some more information to the agent in terms of dimensionality of the context. To further this concept in the continual setting, average reward accumulation is prefered over exponentially-discounted reward as it enables focus on information retention of the "relevant" knowledge and forgetting the less-catastrophic knowledge. How do we account for the passive non-stationarity in the enviroment, e.g. day-night cycles, fundamental laws of the nature, and even current beliefs based on the insufficiently explored states?

Animals and humans subconsciously categorize skills into habits and reactions or niche and globally irrelevant behaviors. A particular problem occurs when an agent collects contradicting experiences or attains knowledge that contradicts with its prior beliefs. Taking into account the average reward accumulation setting, an agent's expectations after a period contradictory-experiences die out and lead to forgetting of prior knowledge. In an ideal continual agent, every discrepancy between expectations and experiences should lead to a change in the agent's causal model of the world, making sure the causal model does not match a previous state. This can be inefficient in a single-agent setting of the CRL problem where an agent's experiences can be vastly different than another agent from the same initial set of agents.

Highlighting the importance of "explicit sequential reasoning" for intelligence, including the ability to make "forward-looking planning" and "backward-looking causal relations.", in continual RL settings where exhaustive exploration is impossible poses causal reasoning to be crucial for efficient learning. In this context, can we develop causal reasoning capabilities that allow continual RL agents to efficiently explore and build models in open-ended environments? This might involve extending concepts like eligibility traces and source traces to approximate state spaces and control settings. Could these be combined with generative models to allow agents to reason about hypothetical states and actions?

Information-theoretic methods for exploration as potentially more suitable than "optimism in the face of uncertainty" for continual RL. How can causal knowledge guide information-seeking behavior in non-stationary environments? This could connect to ideas around artificial curiosity and intrinsic motivation in RL.

How can humans efficiently translate their task preferences into a reward function for the continual agent? Can we always translate preferences into a Markovian reward function, and what is the complexity of doing so? How do we discover sub-goals naturally in tasks like solving a Rubik's cube without explicit instructions?

As a broad overview of the ideal learning system, an agent that is able to learn the underlying casual relationships between various interactions (active and passive), via cooperation for example, and able to to constantly form reaching goals that lead to endless adaptation and efficient knowledge retention, might address the highlighting problems of continual reinforcement learning and help materialize the goal of creating a "child machine" (McCarthy, 1998).

## References

1. Kumar, S., Marklund, H., Rao, A., Zhu, Y., Jeon, H. J., Liu, Y., Benjamin, V. R. (2023, July 10). *Continual learning as computationally constrained reinforcement learning.* arXiv.org. https://arxiv.org/abs/2307.04345

2. Abel, D., Barreto, A., Benjamin, V. R., Precup, D., Hado, V. H., Singh, S. (2023, July 20). *A definition of continual reinforcement learning.* arXiv.org. https://arxiv.org/abs/2307.11046

3. Kim, D., Riemer, M., Liu, M., Foerster, J. N., Tesauro, G., How, J. P. (2022, October 28). *Game-Theoretical Perspectives on Active Equilibria: A Preferred Solution Concept over Nash Equilibria.* arXiv.org. https://arxiv.org/abs/2210.16175

4. Silver, D., Singh, S., Precup, D., Sutton, R. S. (2021). *Reward is enough.* Artificial Intelligence, 299, 103535. https://doi.org/10.1016/j.artint.2021.103535

5. McCarthy, J. What is artificial intelligence. URL: http://www-formal.stanford.edu/jmc/whatisai.html, 1998.

6. Richens, J., Everitt, T. (2024, February 16). Robust agents learn causal world models. arXiv.org. https://arxiv.org/abs/2402.10877

7. Putnam, H. Three kinds of scientific realism. The Philosophical Quarterly (1950), 32(128):195–200, 1982.

8. Hardin, C. L. and Rosenberg, A. In defense of convergent realism. Philosophy of Science, 49(4):604–615, 1982.

9. Cao, R. and Yamins, D. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. Cognitive Systems Research, 85, 2024.

10. Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. Advances in neural information processing systems, 34:225–236, 2021.

11. Huh, M., Cheung, B., Wang, T., Isola, P. (2024, May 13). The platonic representation hypothesis. arXiv.org. https://arxiv.org/abs/2405.07987