Openmind Research Institute Fellowship Proposal

Personal Introduction

My name is Paolo Di Prodi. I am a software and automation engineer and currently a Senior Researcher at CrowdStrike, an American cybersecurity company. For the past 20 years I have worked on automating human decision processes in high-stakes environments using reinforcement learning, causal inference, and deep learning with a strong engineering mindset.

For instance I have worked in other companies such as Microsoft, Fortinet and R&D university projects to implement cutting edge ML systems with a strong emphasis on security, safety and performance. My research spans both software and hardware (robotics): my PhD focused on multi-agent systems—software and physical agents that learn to act and cooperate based on principles from cybernetics and social psychology.

I have developed both research prototypes and commercial systems using state-of-the-art RL algorithms, and in the past I received several entrepreneurship awards, including a Royal Society fellowship for applying advanced machine learning to IoT sensors for senior citizens and people with disabilities.

I now believe that the next leap in AI requires systems that learn and reason like children—building causal world models, imagining counterfactuals, and planning over options—which aligns directly with the OAK architecture and with the Alberta Plan's focus on model-based, interpretable, human-aligned intelligence and Proto AI.

Hence my proposal below has a special focus on the implementation of causality in the OAK architecture from a developmenental view.

Background: The Ladder of Causality and Proto AI

A key theoretical foundation for my proposal is Judea Pearl's "Ladder of Causality," which hierarchically organizes cognitive capabilities into three distinct rungs:

1. **Association (Seeing):** The ability to detect patterns and correlations in data. This is the domain of statistics and standard supervised deep learning.
2. **Intervention (Doing):** The ability to predict the outcomes of deliberate actions and changes in the environment. This is the domain of standard Reinforcement Learning, where agents learn policies by interacting with the world ($P(y|do(x))$).
3. **Counterfactuals (Imagining):** The ability to reason about hypothetical scenarios, alternative pasts, and "what if" questions. Unlike the first two rungs, this capability remains nebulous in modern RL, lacking standardized algorithms or formal implementation frameworks—a gap this proposal aims to fill.

While many animals exhibit associative learning and some demonstrate interventional capabilities, only humans have fully developed the capacity for **counterfactual discovery and reasoning**. This ability to imagine non-existent worlds and compare them with reality is what enables scientific discovery, moral judgment, and complex planning. I consider this capability essential for the next generation of AI. Integrating counterfactual reasoning into the "Proto AI" is not just an enhancement but a necessity for creating agents that can truly understand their environment, learn efficiently from limited data, and explain their decisions.

Pearl also argues that the human mind internally encodes causal knowledge in symbolic form—essentially as directed acyclic graphs (DAGs) that support intervention and counterfactual queries. He explicitly notes that explaining how such structured causal representations are *acquired and refined developmentally* (from raw interaction and episodic memory to stable abstract causal graphs) is an open research problem that he did not have time to solve (Pearl & Mackenzie, 2018, "The Book of Why"). This proposal treats that unsolved question as central: rather than assuming a hand-crafted DAG, Proto AI within the OAK framework will allow causal structure to *emerge* from world-model rollouts, option-based interventions, and counterfactual (retrospective) credit assignment signals.

While this research may confirm or disprove Pearl's intuition about the developmental origins of symbolic causal representation, this is a secondary outcome. The primary goal is to investigate how agents can learn to act causally and robustly through counterfactual reasoning, regardless of whether explicit symbolic structures or meta-policies emerge in the process.

Developmental Trajectory of Counterfactual Intelligence

The development of counterfactual thinking in humans offers a critical blueprint for artificial intelligence. Unlike associative learning, which is present from birth, counterfactual reasoning emerges in distinct stages, driven by a combination of biological maturation (executive functions) and social scaffolding.

Stage 1: Basic Counterfactuals (Ages 3-4)

- **Capability:** Children begin to answer simple "what if" questions about the immediate past (e.g., "If I hadn't touched the hot stove, I wouldn't have been burned"). They can mentally undo a single premise to simulate an alternative outcome.
- **Mechanism:** This stage correlates with the development of language and basic working memory. It appears to be a cognitive leap driven by the maturation of the prefrontal cortex, allowing the child to inhibit the "reality bias" (what actually happened) to entertain a false premise.
- **Role of Learning:** While the capacity is biological, the *structure* is often learned. Parents play a crucial role by asking guiding questions ("What could you have done differently?"), effectively co-constructing counterfactual scenarios and teaching the child how to run these mental simulations.

Stage 2: Counterfactual Emotions (Ages 5-7)

- **Capability:** The emergence of **regret** and **relief**. Children start to compare reality not just to a neutral alternative, but to a better or worse one, and feel emotions based on that comparison. This is the "dopaminergic" stage where counterfactuals begin to drive reinforcement learning.
- **Mechanism:** Requires advanced executive functions, specifically the ability to hold two conflicting representations (actual vs. counterfactual) in mind simultaneously and perform a value comparison. This mirrors the OFC-striatal circuit maturation described in the neuroimaging section.
- **Role of Learning:** Social feedback is essential here. Parents label these complex emotions ("You're sad because you wanted the other toy") and validate the causal link between choice and outcome, helping children internalize the regret signal as a learning tool rather than just a negative feeling.

Stage 3: Complex Planning and Moral Reasoning (Ages 7+)

- **Capability:** Children use counterfactuals for complex future planning and moral judgment (understanding negligence vs. accident). They can simulate long chains of causality and understand "opportunity cost" in abstract domains.
- **Mechanism:** Full integration of the episodic simulation network (hippocampus, mPFC, parietal cortex).
- **Role of Learning:** Formal education and complex social play refine this skill, teaching children to use counterfactuals for argumentation, scientific reasoning (hypothesis testing), and strategic gaming.

Implication for AI: This trajectory suggests that Proto AI should not be "born" with full counterfactual capabilities. Instead, it should follow a curriculum: first learning to simulate simple physical alternatives (Stage 1), then integrating these simulations into a value/reward system to feel "regret" (Stage 2), and finally using them for abstract planning and communication (Stage 3).

Research Goals

The overarching goal of this project is to identify and implement the algorithms that will allow a Proto AI, instantiated within the OAK architecture, to *develop* causal and counterfactual abilities in a staged, child-like manner rather than having them hard-coded. Concretely, this breaks down into four interlocking goals:

1. **Stage 1 – Associative and Simple Counterfactual World Models**
   Develop learning algorithms that enable Proto AI agents to progress from pure association (pattern recognition) to simple, episodically grounded "what if" simulations about the immediate past. Here the focus is on:
   - Learning a world model that can generate plausible alternative outcomes for single-step actions in physical or grid-world environments.
   - Implementing basic reality-inhibition mechanisms (analogous to early prefrontal maturation) so the agent can maintain both actual and hypothetical states without confusion.
2. **Stage 2 – Counterfactual Emotions and Composite Learning Signals**
   Implement and test composite prediction-error mechanisms (actual + counterfactual) inspired by dopaminergic circuits, allowing Proto AI to experience analogues of *regret* and *relief*.
   - Integrate counterfactual outcomes into the agent's value updates, so unchosen actions are learned about via simulated outcomes.
   - Systematically compare agents with and without these composite signals to quantify gains in sample efficiency, adaptation speed, and robustness under change.
3. **Stage 3 – Abstract Planning, Social Reasoning, and Communication**
   Extend counterfactual reasoning from concrete actions to abstract options and social settings:
   - Design multi-agent tasks where Proto AIs must model what *other* agents would have done under alternative choices (theory of mind via counterfactuals).
   - Investigate how communication can be used to share counterfactual information ("If I had gone left, we would have collided") and how this shapes joint planning.
4. **Developmental Curriculum and Alignment with Human Data**
   Formalize and evaluate a training curriculum that mirrors human developmental stages:
   - Start with simple, single-step physical counterfactuals; progress to counterfactual emotions in decision tasks; and finally to long-horizon planning and moral-like trade-offs.
   - Use behavioral and neural benchmarks from developmental psychology and cognitive neuroscience (e.g., emergence of regret around 5–7 years, performance on child counterfactual tasks) as external targets for Proto AI behavior, grounding the architecture in human-like causal learning trajectories.

Across all stages, the central research question is: **which algorithms and architectural constraints are necessary and sufficient for an agent to climb Pearl's ladder of causality—from association, through intervention, to genuine counterfactual discovery—using only experience, world models, and internally generated simulations?**

Causal learning and counterfactuals

Planning vs. Counterfactual Reasoning: Neural and Computational Distinctions

Both planning ("what will happen if...") and counterfactual reasoning ("what would have happened if...") use overlapping brain systems—primarily the hippocampus (memory), prefrontal cortex (simulation and control), and parietal regions (integration). However, they differ in critical ways that inform how we should build learning agents.

Key Differences

Planning (Forward Thinking) imagines possible futures by recombining past experiences in open-ended ways. It relies on general knowledge and schemas, becoming more automatic for highly probable scenarios.

Counterfactual Reasoning (Retrospective Thinking) imagines alternative pasts by taking a known event and asking "what if I had done differently?" It's tightly constrained by what actually happened, requiring extra mental effort to maintain both the real and imagined scenarios simultaneously.

Brain imaging reveals three critical distinctions:

1. **More conflict monitoring for counterfactuals:** Comparing "what happened" vs. "what could have happened" requires juggling competing representations, engaging additional control regions in the prefrontal cortex.
2. **Opposite memory engagement:** Hippocampal activity *increases* for plausible counterfactuals (because they must fit stored context) but *decreases* for plausible futures (because familiar futures need less memory reconstruction).
3. **Regret and value comparison:** The orbitofrontal cortex (OFC) is essential for counterfactual value judgments. Lesion studies show that patients with OFC damage cannot experience regret or use "what might have been" to improve future decisions—they learn only from direct experience.

How Dopamine Encodes Both Actual and Counterfactual Outcomes

The dopamine system, traditionally understood as signaling reward prediction errors (actual reward minus expected reward), does something more sophisticated: it encodes *composite* error signals that include counterfactual information.

Composite Learning Signal: When you choose an action and receive a reward, dopamine neurons respond not just to whether the reward was better or worse than expected, but also to whether it was better or worse than what you *could have* gotten with a different choice. The same reward triggers different dopamine responses depending on the foregone alternative—this is the neural basis of regret and relief.

Three-Component Circuit:

- **Hippocampus** provides the memory of what actually happened and constraints for simulating plausible alternatives
- **OFC** compares actual vs. counterfactual outcomes, computing how much better or worse the alternative would have been
- **Dopamine/Striatum** combines both signals into a unified teaching signal that updates values for both chosen and unchosen actions

This circuit enables learning from actions you *didn't* take, not just actions you did take—a form of off-policy learning grounded in biology.

Implementation in the OAK Architecture

The OAK framework can naturally incorporate composite dopaminergic signals by extending its learning mechanism to include counterfactual credit assignment:

1. **World Model (Knowledge) generates counterfactuals:** When an option terminates and an outcome is observed, the agent's world model can simulate what would have happened under alternative options from the same state. The hippocampus-based episodic memory retrieves the actual trajectory and context, while mPFC and parietal cortex generate constrained alternatives (as in eCFT).

2. **Composite prediction error:** The agent computes two parallel errors:

    - δ_actual = R_obtained - V(s, o_chosen) [standard temporal-difference error]
    - δ_counterfactual = max(V(s, o_other)) - R_obtained [opportunity cost or regret signal]

    These are combined into a composite error: δ_composite = δ_actual + α·δ_counterfactual, where α controls the weight given to counterfactual information (potentially modulated by OFC and fronto-parietal control signals).

3. **Update option values:** Both chosen and unchosen option values are updated:

    - V(s, o_chosen) ← V(s, o_chosen) + β·δ_composite
    - V(s, o_unchosen) ← V(s, o_unchosen) + γ·δ_counterfactual [off-policy learning from simulated alternatives]

4. **Dopamine as teaching signal:** In biological terms, dopamine release encodes δ_composite and modulates synaptic plasticity throughout striatum and cortex. In OAK, this composite error serves as the teaching signal for both model-free option values and model-based world model updates, enabling the agent to learn simultaneously from what it experienced and what it inferred it missed.

5. **OFC as comparator and gain modulator:** The OFC component in OAK (potentially part of the "Options" evaluation system) computes the explicit comparison between actual and counterfactual values, sets the α parameter based on task demands and uncertainty, and broadcasts this comparative signal back to the learning system. When counterfactual information is reliable and relevant (high constraint from episodic memory), α increases; when counterfactuals are speculative or unreliable, α decreases.

This architecture enables OAK agents to:

- Learn more efficiently by leveraging unchosen alternatives (reducing sample complexity)
- Develop regret-sensitive and relief-sensitive policies that account for opportunity costs
- Integrate retrospective counterfactual reasoning (eCFT) with prospective planning (EFT) in a unified value-learning framework
- Modulate the balance between model-free (cached option values) and model-based (simulated counterfactuals) learning dynamically based on task structure

The convergence of regret circuitry (OFC/vmPFC), hippocampal constraint checking, striatal encoding, and composite dopaminergic signaling provides a biologically grounded blueprint for implementing counterfactual reinforcement learning in artificial agents. This goes beyond standard model-free or model-based RL: it is a *counterfactual RL* mechanism where the agent's learning signal inherently incorporates "what could have been," leveraging the structural constraints of past episodes to update both experienced and unexperienced option values simultaneously.

Neurally-Grounded Counterfactual RL: The Zhang et al. Framework

The computational approach proposed by Zhang et al. (2015) provides a cognitive model that directly operationalizes the neuroimaging findings into a practical learning algorithm suitable for OAK. Their model, developed to explain human performance in change-detection tasks, implements a hybrid learning system that combines standard RL updates with counterfactual credit assignment—a design that maps naturally onto the neural architecture documented above.

Core Algorithm: In Zhang et al.'s framework, after selecting strategy *s* and observing outcome *r*, the agent performs two parallel updates:

1. **Actual RL update (standard TD learning):**

    - U(s_chosen) ← U(s_chosen) + β·[r - U(s_chosen)]

2. **Counterfactual update (hypothetical evaluation):**

    - U(s_unchosen) ← U(s_unchosen) + γ·[r_counterfactual - U(s_unchosen)]

where *r_counterfactual* is the outcome that *would have* occurred under the alternative strategy, estimated via the agent's internal model of the environment.

Neural Implementation Mapping:

This algorithm directly implements the neural circuit described in the neuroimaging literature:

- **Hippocampus (episodic memory + world model):** Stores the actual state and outcome, provides the episodic context needed to simulate "what would have happened" under the unchosen option. The hippocampus's constraint-checking role ensures counterfactual simulations respect stored contextual information.

- **mPFC (simulation and recombination):** Generates the counterfactual trajectory by flexibly recombining episodic elements from the actual experience under a hypothetical action. The stronger mPFC activation seen in counterfactual vs. future thinking reflects the additional computational demand of maintaining both actual and counterfactual representations simultaneously.

- **OFC (comparative valuation):** Computes the explicit comparison between actual and counterfactual outcomes (r_obtained vs. r_counterfactual), generating the regret/relief signal. This comparison determines the sign and magnitude of counterfactual updates and modulates the weight γ given to counterfactual information.

- **Dopamine/Striatum (composite learning signal):** The actual update is driven by standard dopaminergic RPE (r - U), while the counterfactual update reflects the dopamine system's sensitivity to foregone alternatives. Ventral striatum's encoding of comparative value (relief for "better than counterfactual," regret for "worse than counterfactual") directly modulates both learning rates and the relative weight given to actual vs. counterfactual terms.

- **Fronto-parietal control (adaptive weighting):** Modulates β and γ based on task demands, uncertainty, and conflict. When counterfactual information is reliable (high episodic constraint, low model uncertainty), γ increases; when counterfactuals are speculative or the model is poor, γ decreases to prevent learning from unreliable simulations.

Behavioral and Neural Fit:

Zhang et al. demonstrated that humans performing change-detection tasks fit this hybrid model better than pure RL or pure Bayesian models, and adapt faster after environmental changes precisely because they update utilities for unchosen strategies based on counterfactual reasoning. This behavioral pattern mirrors the neural evidence:

1. Humans with intact OFC show regret-based learning and update beliefs about unchosen options; OFC-lesioned patients do not.
2. Dopamine signals in striatum encode composite errors that include counterfactual terms, not just actual outcomes.
3. Hippocampal activity scales differently with likelihood for counterfactuals (increases with plausibility) vs. futures (decreases), consistent with the hippocampus providing constrained, episodically-grounded counterfactual simulations rather than free-form predictions.

Advantages for OAK Over SCM-Based Counterfactual RL:

Most counterfactual RL methods in the machine learning literature (e.g., SCM-based offline RL, causal discovery + counterfactual reasoning) require:

- Explicit structural causal models (DAGs) learned in a separate discovery phase
- Offline batch processing of logged datasets
- Strong identifiability assumptions and full observability of causal variables

In contrast, the Zhang et al. approach—grounded in cognitive neuroscience—requires only:

- A learned forward model (world model / Knowledge component in OAK) that can simulate $s' = f(s, a)$
- Real-time, online learning from continuous interaction
- No preconceived causal graph—the agent simply simulates alternative actions from experienced states and compares outcomes

This makes it ideal for OAK, which emphasizes:

- **Continual learning:** Updates happen after every transition, not in batch
- **Model-based reasoning:** The Knowledge component provides the forward model needed to generate counterfactuals
- **Biological plausibility:** The algorithm mirrors documented neural computations in hippocampus, OFC, and dopamine systems
- **Scalability:** No need for costly causal discovery; counterfactuals emerge naturally from world-model simulations

Implications for the OAK Architecture

In the context of the OAK (Options And Knowledge) framework:

- **Planning (EFT)** corresponds to using the world model (Knowledge) to simulate future trajectories under different options, relying on semantic schemas and forward projection with minimal constraint from specific past episodes.
- **Counterfactual reasoning (eCFT)** involves retrospectively simulating alternative trajectories from known past states, requiring explicit representation of both what happened and what could have happened, and engaging conflict-monitoring systems to maintain consistency with stored episodic constraints.

An OAK agent equipped with both modes can:

1. Use forward planning to select options in novel or open-ended contexts (relying on schematic knowledge and likelihood-reduced hippocampal engagement).
2. Use counterfactual reasoning to refine option values and causal models post-hoc, leveraging stored episodes and dopaminergic learning signals to update beliefs about what actions would have led to better outcomes.

The integration of these two processes—one prospective and schema-driven, the other retrospective and constraint-bound—enables richer learning and more flexible decision-making, mirroring the dual roles of the hippocampus and dopamine system in both memory-based simulation and value-based learning.

Knowing When to Reason: The Role of Habit and Cognitive Efficiency

However, a key aspect of intelligence is not just the ability to perform complex reasoning, but knowing *when* to do so. Many human decisions are not the product of deliberate counterfactual simulation. Instead, we rely on:

- **Habits and Cached Behaviors:** Actions like tying shoelaces, driving a familiar route, or making coffee are executed automatically from cached motor programs. These are computationally cheap and fast, governed by model-free systems that do not require simulating alternatives.
- **Heuristics and Simple Rules:** In many situations, we use simple rules of thumb ("if the light is red, stop") that are learned through association and do not involve imagining what would happen if we broke the rule on this specific occasion.

This reflects a "dual-process" cognitive architecture where a fast, intuitive, and low-energy "System 1" (habits) is balanced against a slow, deliberative, and energy-intensive "System 2" (counterfactual reasoning and planning). The brain arbitrates between these systems constantly. We default to habits until something unexpected happens—a roadblock on a familiar route, a strange taste in our coffee—which triggers the more complex reasoning systems to engage.

Implications for Proto AI: A truly intelligent Proto AI must not be a "pure reasoner" that always engages in costly counterfactual simulations. It must also develop a robust habitual system and, crucially, learn *how to arbitrate* between the two. The OAK architecture should therefore include a mechanism that decides whether to:

1. Execute a cheap, cached option (habitual action).
2. Engage the world model for prospective planning (forward simulation).
3. Engage the world model for retrospective analysis (counterfactual simulation).

This arbitration could be modulated by signals like prediction error, uncertainty, or novelty. When the world is predictable and outcomes match expectations, the habitual system dominates. When surprise or high stakes are detected, the deliberative, counterfactual system is called upon. This allows the agent to be both efficient in familiar contexts and robustly intelligent in novel ones, a hallmark of human cognition.

Related Research

The integration of counterfactual reasoning into reinforcement learning is an active area of research, though approaches vary significantly in their mechanisms and goals. Current literature can be broadly categorized into six main implementation styles:

1. **SCM-based Counterfactual RL:** These methods, often used in offline or risk-sensitive settings, explicitly build Structural Causal Models (SCMs) to generate counterfactual transitions. By using abduction-action-prediction cycles, they augment datasets with "what would have happened" scenarios to improve data efficiency and mitigate distributional shift (e.g., *Sample-Efficient RL via Counterfactual Reasoning*, 2020; *Counterfactually-Guided Causal RL*, 2024).

2. **Counterfactuals for Sequence Models:** Recent work with Decision Transformers uses counterfactuals as sequence augmentation. By altering actions or rewards in logged trajectories and using learned models to infer outcomes, these approaches help transformers generalize better from limited data and "stitch" together suboptimal trajectories (e.g., *CRDT*, 2025).

3. **Cognitive Models of Human Learning:** Psychological models, such as the one by Zhang et al. (2015), integrate counterfactual evaluation directly into RL updates. These models modify temporal-difference learning by adding terms for unchosen actions, better fitting human behavioral data in change-detection tasks. This aligns closely with the neuro-computational approach proposed here.

4. **Causal Discovery for Applications:** In applied domains like persuasion systems or medical treatment planning, researchers combine causal graph discovery with RL. Counterfactual estimates are used to construct enriched training signals or to evaluate policy changes safely without executing them in the real world (e.g., *Zeng et al.*, 2025).

5. **Counterfactual Explanations:** Some works use counterfactuals not to improve the policy itself, but to explain it. Techniques like *Counterfactual Explainer for Deep RL* generate minimal perturbations to states that would have changed the agent's action, providing post-hoc interpretability without altering the learning loop.

6. **General Causal Frameworks:** Foundational theoretical works frame RL within broader causal inference contexts (e.g., Bottou et al., 2013), providing the

mathematical basis for off-policy evaluation and inverse propensity scoring that underpins many practical counterfactual RL methods.

While SCM-based methods dominate the engineering literature, they often rely on heavy offline causal discovery and batch processing. The approach proposed in this fellowship focuses on the *developmental* and *online* emergence of these capabilities—mirroring the cognitive models (Category 3) where counterfactuals arise from real-time interaction and world-model simulation rather than pre-computed causal graphs.

Comparison with Regret-Based RL

This proposal shares the fundamental mathematical objective of standard regret-based algorithms (e.g., UCB, UCRL2, CFR): minimizing the difference between realized rewards and optimal possible rewards. However, the implementation and architectural focus differ significantly in three key areas:

1. Instantaneous vs. Cumulative Regret

Standard algorithms like UCB or UCRL2 focus on minimizing **cumulative regret** over a time horizon to drive exploration (Optimism in the Face of Uncertainty). In contrast, this proposal uses **instantaneous regret** (or "counterfactual error") as a direct **teaching signal**.

- **Standard Approach:** Regret bounds are used to construct confidence intervals that guide future action selection.
- **Proposed Approach:** The "Composite Learning Signal" ($\delta_{counterfactual} = \max(V(s, o_{other})) - R_{obtained}$) immediately updates the value function of unchosen options. This allows the agent to learn from what it *didn't* do, rather than just using uncertainty to explore.

2. Source of Counterfactuals: World Model vs. Statistical Bounds

Standard bandit algorithms generate "counterfactuals" implicitly via statistical confidence intervals or probability distributions (e.g., Thompson Sampling).

- **Standard Approach:** Relies on mathematical bounds derived from information theory.
- **Proposed Approach:** Generates counterfactuals via **episodic simulation** using the OAK "Knowledge" component (World Model). The agent explicitly simulates $s' = f(s, a_{unchosen})$, bridging Model-Based RL and Regret Minimization. This allows for grounded counterfactual reasoning in complex, non-tabular environments where statistical bounds are intractable.

3. Biological Plausibility and Developmental Curriculum

Most regret-minimizing algorithms are "born" with fixed learning rules derived from theoretical guarantees.

- **Standard Approach:** Algorithms are static and mathematically derived.
- **Proposed Approach:** The mechanism is **neuromorphic**, derived from the OFC-Striatal circuits and the Zhang et al. framework. Furthermore, it follows a **developmental trajectory**: regret-based learning is not a static rule but an emergent capability that comes online only after basic world-modeling is established (Stage 2 of the curriculum). This aligns with the "Proto AI" goal of building intelligence that grows and matures.

Proposed contribution

Experimental Validation: Testing Counterfactual RL in OAK

To validate the counterfactual learning mechanism described above, I propose a systematic experimental program with benchmarks of increasing complexity, designed to isolate and measure the specific advantages of composite dopaminergic learning signals.

Benchmark Tasks

1. Multi-Armed Bandit with Partial Feedback (Baseline)

- **Setup:** Agent chooses among N arms; receives reward for chosen arm only (standard bandit), or sees rewards for all arms (full feedback).
- **Complexity:** Stationary rewards initially, then introduce reward drift/changes to test adaptation.
- **Why this task:** Cleanest test of counterfactual learning. Traditional RL learns only from chosen actions; counterfactual RL should update unchosen arm values when full feedback is available, and infer counterfactuals from its world model under partial feedback.
- **Expected outcome:** Counterfactual agent adapts faster to reward changes because it learns from both experienced and unexperienced arms simultaneously.

2. Grid-World Navigation with Regret Scenarios

- **Setup:** Agent navigates a grid to reach goals with varying rewards. Crucially, after reaching a goal, the agent briefly observes what reward it *would have* received at alternative nearby goals.
- **Complexity:** Start with deterministic transitions, then add stochasticity. Introduce foraging scenarios where the agent must choose between a guaranteed small reward and a risky large reward.
- **Why this task:** Tests whether the agent develops regret-sensitive policies. Human OFC-lesion patients fail to adjust choices based on foregone alternatives—can we replicate this computationally?
- **Expected outcome:** Counterfactual agent shows faster convergence to optimal policy in regret-rich scenarios and develops "relief-seeking" or "regret-avoiding" strategies that pure RL agents miss.

3. Atari Games with Sparse Reward + Counterfactual Feedback (Video Game)

- **Setup:** Select Atari games where players make discrete, high-impact choices (e.g., Ms. Pac-Man: which ghost to evade, which power pellet to collect). Modify environment to occasionally show "what would have happened" if the agent had taken a different action at a critical decision point.
- **Complexity:** High-dimensional visual state space, long temporal credit assignment, partial observability.
- **Why this task:** Tests scalability to complex, realistic domains. Traditional RL agents struggle with sparse rewards; counterfactual feedback from the world model should accelerate learning by providing additional gradient signal for unchosen high-value actions.
- **Expected outcome:** Counterfactual agent achieves higher cumulative reward with fewer environment interactions (improved sample efficiency), and shows more robust performance under distribution shift (new game configurations).

4. Multi-Agent Coordination with Communication (Robot Task)

- **Setup:** Two or more robots (physical or simulated) must coordinate to move objects, where each robot observes only local information. After joint actions, robots can "communicate" counterfactual information: "If I had moved left instead of right, we would have succeeded/failed."
- **Complexity:** Partial observability, non-stationary environment (other agents' policies are changing), credit assignment across agents.

- **Why this task:** Tests whether counterfactual reasoning supports joint intentionality and theory of mind. Agents must not only learn their own counterfactuals but also infer what their partners' alternative actions would have led to.
- **Expected outcome:** Counterfactual agents converge to coordinated policies faster, develop better theory of mind (predicting partner actions), and show more flexible adaptation when partner behavior changes.

5. Real-World Robotic Manipulation with Safety Constraints (Robot Task)

- **Setup:** Robot arm must learn to manipulate objects (pick, place, stack) under strict safety constraints (e.g., never drop fragile items, avoid collisions). After each episode, the world model simulates counterfactual trajectories for unchosen grasps or motion plans.
- **Complexity:** Continuous state/action space, real-time constraints, safety-critical (limited real-world exploration budget).
- **Why this task:** Tests whether counterfactual learning enables safe, sample-efficient learning in domains where traditional trial-and-error is too costly. Counterfactual simulations from the world model should allow the agent to learn from "near-miss" scenarios without physical risk.
- **Expected outcome:** Counterfactual agent learns successful manipulation policies with fewer real-world trials, fewer safety violations, and better generalization to novel objects.

Expected Behavioral Signatures vs. Traditional RL

Sample Efficiency:

- Traditional RL agents learn only from experienced (state, action, reward) tuples.
- Counterfactual agents learn from both experienced and simulated unexperienced alternatives, effectively doubling or tripling the information extracted from each real interaction.
- **Quantitative metric:** Number of episodes to reach 90% of optimal policy performance. Expect a significant reduction for counterfactual agents in tasks with high decision-branch complexity.

Adaptation Speed After Environmental Changes:

- Traditional RL requires extensive re-exploration when reward structure or dynamics change.
- Counterfactual agents, by maintaining updated values for unchosen options, can pivot more quickly when previously suboptimal actions become optimal.
- **Quantitative metric:** Episodes required to recover to 90% optimal performance after an environmental shift. Expect counterfactual agents to adapt faster, matching Zhang et al.'s human behavioral data.

Regret-Sensitive Value Functions:

- Traditional RL agents evaluate states/actions purely by expected return.
- Counterfactual agents should show value functions modulated by opportunity cost: the same objective reward receives different learned value depending on what alternatives were available.
- **Qualitative signature:** In repeated choice scenarios, counterfactual agents should show "disappointment" (reduced value update) for objectively good rewards that are worse than foregone alternatives, and "relief" (enhanced value update) for objectively mediocre rewards that are better than foregone alternatives. This maps directly to ventral striatum BOLD patterns in human neuroimaging.

World Model Quality and Constraint Checking:

- Traditional model-based RL uses world models purely for forward planning (future trajectories).
- Counterfactual agents must use world models for retrospective, *constrained* simulation: generating plausible alternatives that respect what actually happened up to the decision point.
- **Quantitative metric:** Measure world model prediction accuracy specifically for counterfactual queries (predicting outcomes under actions not taken in a specific past context) vs. forward planning queries. Counterfactual agents should develop better-calibrated models for off-distribution actions because they train on these queries.

Ablation Studies:

- Implement agents with $\alpha=0$ (no counterfactual weight) vs. $\alpha>0$ (composite learning signal) to isolate the contribution of counterfactual updates.
- Test with perfect world models (oracle counterfactual outcomes) vs. learned world models (simulated counterfactual outcomes with error) to measure robustness to model uncertainty.
- Compare fixed α vs. adaptive α (modulated by epistemic uncertainty, as in the OFC gain control hypothesis) to validate the fronto-parietal control mechanism.

Implementation and Open Source Contributions

I am a great supporter of open source and will contribute implementations of the counterfactual RL algorithms described above to the institute's codebase, including:

- Modular components for composite prediction error computation that can be integrated into existing OAK agents
- Benchmark task environments with built-in counterfactual feedback mechanisms
- Evaluation tools for measuring sample efficiency, adaptation speed, and regret-sensitivity
- Documentation and tutorials connecting the neuroscience foundations to the algorithmic implementation

I will also be very interested in assessing how the algorithms scale—as described in the scaling principle—from simple robots with small resources to larger robots (such as those recent humanoid robots) with more sensors and computer resources to also infinite scale on the cloud. The counterfactual learning mechanism should provide consistent benefits across scales, as the core computational principle (learning from simulated alternatives) is scale-invariant.

Funding Requirements

I am requesting minimal funding limited strictly to travel and attendance at meetings and conferences directly related to advancing and coordinating work on the OAK architecture and Proto AI (e.g., OAK internal workshops, reinforcement learning and causal reasoning venues, and key neuroscience/AI crossover events). I am not requesting salary support, compute credits, or hardware purchases.

Estimated annual travel budget:

- Flights + lodging for 2–3 international trips (OAK meetings / major conferences)
- Conference registration and workshop/tutorial fees
- Local transportation and incidental costs

Cost range: GBP £6,000–£8,000 total (approx. CAD $10,200–$13,600 assuming £1 ≈ $1.70 CAD). Should remote participation substitute for any planned travel, unused funds would be released or reallocated at the institute’s discretion. I will provide a transparent post-event summary to justify each trip ex-post.

Disclaimer

The work described in this proposal will be undertaken in my personal capacity and conducted entirely outside the scope of my employment duties and hours. No resources, equipment, or proprietary information of my current employer will be used in the execution of this project. Consequently, all intellectual property, including any inventions, discoveries, and creative works generated as a result of this research, will be attributed to the Openmind Research Institute which follows a policy of free and open knowledge.

References

Talks and Presentations

1. **Oak Architecture**: YouTube Link
2. **Toward Deep Learning**: YouTube Link
3. **AGI Conference**: YouTube Link
4. **Planning and Action Selection in Option based agents**: YouTube Link

Key Papers

5. Pearl, J., & Mackenzie, D. (2018). **"The Book of Why: The New Science of Cause and Effect"**. *Basic Books*. [Link]
6. Sutton, R. S., Precup, D., & Singh, S. (1999). **"Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning"**. *Artificial Intelligence*. [Link]
7. Zhang, S., Maddox, W. T., & Glass, B. D. (2015). **"Reinforcement Learning and Counterfactual Reasoning"**. *Topics in Cognitive Science*. [Link]
8. Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). **"Counterfactual thinking: An fMRI study on changing the past for a better future"**. *Social Cognitive and Affective Neuroscience*. [Link]
9. Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). **"Cognitive neuroscience of human counterfactual reasoning"**. *Frontiers in Human Neuroscience*. [Link]
10. Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). **"Regret and its avoidance: A neuroimaging study of choice behavior"**. *Nature Neuroscience*. [Link]
11. Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., White, J. P., Ellis, T. L., Phillips, P. E. M., & Montague, P. R. (2016). **"Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward"**. *Proceedings of the National Academy of Sciences*. [Link]
12. Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). **"Children's thinking about counterfactuals and future hypotheticals as possibilities"**. *Child Development*. [Link]
13. O'Connor, E., McCormack, T., & Feeney, A. (2014). **"The development of regret and relief about the outcomes of risky decisions"**. *Journal of Experimental Child Psychology*. [Link]
14. Guajardo, N. R., Parker, J. L., & Turley-Ames, K. J. (2009). **"The co-construction of counterfactual worlds in parent-child conversations"**. *Merrill-Palmer Quarterly*. [Link]
15. Jaksch, T., Ortner, R., & Auer, P. (2010). **"Near-optimal Regret Bounds for Reinforcement Learning"**. *Journal of Machine Learning Research*. [Link]
16. Zinkevich, M., Johanson, M., Bowling, M., & Piccione, C. (2007). **"Regret Minimization in Games with Incomplete Information"**. *Advances in Neural Information Processing Systems (NIPS)*. [Link]
17. Agrawal, S., & Goyal, N. (2012). **"Analysis of Thompson Sampling for the Multi-armed Bandit Problem"**. *Conference on Learning Theory (COLT)*. [Link]

Personal Links

Old Research

Linkedin

Twitter