Indirect Utility Maximization via Second-Order Agents

William Sawyerr
Institute of Creativity and Innovation
University for the Creative Arts
Farnham, Surrey, United Kingdom
william.sawyerr@uca.ac.uk

ABSTRACT

Humans are constrained to a single world with limited capacity to explore others. We formalize this limitation and propose second-order artificial agents trained in Virtual Worlds (VWs) to address it. The approach models worlds as strategy spaces with explicit state structures, action constraints and reward topologies. We implement this as a VW with Earth features where autonomous agents develop boundary-crossing capabilities through curiosity-driven exploration. Agents exhibit both rational and intelligent properties. The work demonstrates that cross-world exploration can be formalized as a computational problem, shifting it from a theoretical constraint to an engineering challenge with concrete design principles.

CCS CONCEPTS

• Computing methodologies • Modeling and simulation • Simulation types and techniques

KEYWORDS

Utility Maximization, Strategic Decision-Making, Evolutionary Optimization, Artificial Agents, Virtual Worlds, Cross-Reality

1 Introduction

Humans are bounded agents constrained to a single world, with limited capacity to enter and operate in other worlds directly. Other worlds can be physical, virtual or abstract [6], but access to them remains largely indirect, mediated by observation or simulation from within the human regime [8, 4]. This creates a structural mismatch between localized human agency and a broader landscape of possible worlds, leaving large portions of the space of worlds and payoffs inaccessible to direct action.

The research problem is to formalize this cross-world limitation as a barrier to exploration and to characterize its implications for long-term agency and knowledge expansion. In its current form, exploration is locally constrained, which limits systematic comparison, evaluation and transfer across distinct worlds. The resulting gap between the worlds that can be directly explored and those that can only be modeled constitutes a core limitation on what humans can learn, test or validate about other worlds.

To address the issue, we propose a method for expanding our reach into other worlds via artificial agents. We create a Virtual World (VW) that approximates core properties of our human regime and can be varied, instrumented and reset. We then instantiate and train artificial agents in this VW for exploration beyond it. By embedding artificial agents in a scalable VW, the method creates a tractable substrate for systematic exploration across distinct worlds and for evaluating transfer between them. The proposed approach therefore frames artificial agents as a multi-agent mechanism for extending human agency into other worlds that are otherwise inaccessible, while preserving the ability to audit, compare and generalize exploratory outcomes across regimes.

In Section 2, we present the background and theoretical foundations for cross-world exploration, modeling worlds as strategy spaces and examining the conditions, under which exploration becomes worthwhile. Section 3 covers the design and development of our VW system, including the formal model of agents as rational and intelligent entities and the implementation architecture supporting boundary-crossing capabilities. Section 4 reflects on the challenges inherent in training agents for regime shifts, including belief calibration, goal preservation and the tension between exploration efficiency and transfer capability. We conclude the paper in Section 5 with a summary of our methodological contributions and findings regarding autonomous learning, belief calibration and multi-agent collaboration in cross-world exploration.

2 Background

A world can be modeled as a strategy space with rules and constraints that define what actions are possible [9]. The central difficulty arises when alternative strategy spaces exist but are not directly accessible. In such cases, the agent faces a meta-decision problem. The choice is not only which action to take, but which world to enter. This creates a higher-order uncertainty because the structure of the alternative world is unknown until entry and partial exploration occur.

Exploration across worlds is therefore not ordinary exploration of actions. It is a regime-shift decision under incomplete information. What can be learned about a new world depends on factors that are only accessible after entry, including different action constraints or resource structures [1, 2]. This produces an

MMVE' 26, April 4–8, 2026, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2535-7/2026/04
https://doi.org/10.1145/3798090.3799679

asymmetry between evaluation and access. A world may be worth entering even when the agent cannot prove it in advance.

Cross-world exploration requires addressing two coupled problems. The first is the decision to allocate effort to exploration given bounded time, resources and institutional capacity. The second is the problem of evaluation after entry. Agents must build models of unfamiliar rules, test hypotheses and estimate outcomes in a regime where historical priors may not apply.

Entities explore other worlds because different systems can reward different capabilities. A new world can enable strategies that were suboptimal in the previous world and can expose advantages that were unavailable under prior constraints [3]. Agents confined to a single world face a fundamental limitation. Their potential is bounded by that world's structure.

To circumvent this, agents can delegate exploration to secondary agents. These secondary agents inherit the primary agent's objectives but operate under different constraints or embodiments. They act as strategic proxies. They traverse adjacent worlds, seek higher value outcomes on the primary's behalf and report discoveries in auditable form.

The problem is to identify the conditions under which world exploration becomes worthwhile. This requires a model that treats cross-world exploration as a meta-strategy with its own costs, risks and expected returns [5, 7]. It also requires a clear account of how information gained in one world can be transferred to decisions about another.

A VW constructed for training such agents allows parametric variation, full instrumentation and iterative resets. Within such a world, we can observe how agents learn to detect saturation signals, evaluate adjacent worlds, preserve objectives during transitions and encode transferable knowledge. The VW serves as a testbed for boundary-crossing strategies. It enables systematic design of agents optimized not for a single strategy space, but for navigating between strategy spaces while carrying forward specified objectives.

3 Design and Development

This section presents the design and development of the VW system for training second-order agents in boundary-crossing strategies. Specifically, Section 3.1 covers the formal design of the VW as a strategy space with explicit state structures, action constraints and reward topologies, along with the theoretical framework governing agent rationality, intelligence and meta-policy decision-making. Section 3.2 covers the implementation architecture, including the distributed multi-agent system infrastructure, the server backend built on Python with spatial databases and vector memory, the client visualization application and the autonomous agent learning mechanisms that enable curiosity-driven exploration across millions of Earth features.

3.1 Design

We model the VW as a strategy space $W = (S, A, T, R, O)$ where S is the geographic state space, A is the action space, T is the transition model, R is the reward function and O is the observation model. The geographic state space S includes position coordinates (latitude, longitude, altitude) and environmental features drawn from a spatial database of Earth features. The action space A includes movement, knowledge queries, communication with other agents and world model updates. This structure allows agents to operate under partial observability with bounded computational resources.

Agents maintain an internal state that includes position, accumulated knowledge, exploration history and beliefs about unexplored regions. An agent follows a policy π that maps belief states to actions. The policy is not hand-coded. It emerges from autonomous learning driven by curiosity, prediction error and information gain. The in-world expected utility is

$$U(\pi) = E\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t)\right]$$

under transition model T and observation model O. Rewards are derived from knowledge acquisition, spatial coverage and successful modeling of unfamiliar regions.

Agents in this design exhibit both rational and intelligent properties. They are rational agents in that they make decisions based on expected utility under uncertainty. Given belief state b_t at time t, an agent selects action a_t that maximizes $E[\,U \mid b_t, a_t\,]$. This requires consistent preferences over outcomes, coherent belief updates via Bayes' rule or approximations and decision rules that respect the structure of the problem. Rationality ensures agents can be analyzed using decision-theoretic frameworks and that their behavior follows from principled optimization rather than arbitrary heuristics.

They are also intelligent agents in that they learn from experience, adapt to novel environments and build transferable models. Intelligence manifests in several capacities. Agents update world models based on observations, improving predictions about state transitions and reward structures. They detect patterns in spatial data and generalize to unseen regions. They identify when current models fail and trigger exploration to gather corrective information. They communicate findings to other agents and integrate knowledge from external sources. Intelligence allows agents to handle environments where rational decision-making alone is insufficient because the world structure is unknown and must be learned.

The combination of rationality and intelligence is necessary for cross-world exploration. Rationality provides the decision framework for when to explore versus exploit, when to transition between worlds and how to trade off immediate rewards against long-term information gain. Intelligence provides the learning mechanisms to build models of unfamiliar worlds, adapt policies to new constraints and transfer knowledge across regime boundaries. An agent that is rational but not intelligent cannot learn about new worlds. An agent that is intelligent but not rational cannot make principled decisions about which worlds to explore or when to terminate exploration.

Cross-world exploration is modeled as a regime shift from world W_i to world W_j with transition cost C_{ij} and entry delay D_{ij}. The total expected utility across multiple worlds is

$$U_{\text{total}} = \max_{\text{sequences}} E\left[\sum_{k}\left(U_k(\pi_k) - C_{k,k+1}\right)\right]$$

In the current implementation, agents train within a single world W_E representing Earth geography. The design allows future extension to multiple worlds by treating each world as a distinct strategy space with its own state structure, action constraints and reward topology.

Agents in this framework are second-order agents. Their action space includes exploratory actions that sample world parameters and build transferable models rather than optimizing only for immediate state transitions. Let the primary agent (human decision maker) have utility U_P. The second-order agent maximizes $U_A = E[\,U_P \mid \text{evidence}\,]$ by selecting actions that increase the value of information about alternative worlds. This formalizes indirect utility maximization via delegated exploration.

A meta-policy governs world entry decisions. Let b_i be the belief over world parameters for world i. A rational entry rule is to transition from world i to world j when $E[\,U_j \mid b_j\,] - C_{ij}$ exceeds $E[\,U_i \mid b_i\,]$ by threshold τ. This captures entry decisions under uncertainty without requiring full observability of the target world. Agents must estimate expected utility based on partial exploration and transfer learning from prior worlds.

These formal models translate into implementation requirements. The system must represent geographic state spaces with high-dimensional feature sets. It must support belief updates as agents query knowledge sources and observe environmental features. It must expose exploration costs including computational resources and time delays. Agents must maintain world models that encode spatial structure, feature distributions and causal relationships. They must evaluate entry decisions based on expected information gain and report discoveries in auditable form that allows the primary agent to update beliefs about alternative worlds.

The architecture implements these requirements through distributed agent actors, spatial databases with PostGIS queries, vector memory for knowledge encoding and real-time observation of agent behavior through visualization. Agents spawn at real Earth locations, autonomously explore driven by curiosity-based policies and develop exploration archetypes based on parameter configurations that encode different meta-strategies for boundary crossing.

3.2 Development

The VW is implemented as a distributed multi-agent system where agents are instantiated as software entities with bounded memory, perception and action capabilities. The system realizes the formal model $W = (S, A, T, R, O)$ described in the design section through explicit world parameters, spatial state representation and observation interfaces that expose geographic and knowledge structures to agent reasoning.

The server backend is built on Python with FastAPI for API endpoints, Ray for distributed agent computation and PyTorch for learning modules. It runs as a containerized stack with PostgreSQL and PostGIS for spatial data management, Redis for event streaming, ChromaDB for vector memory and Ollama for local reasoning support. The server exposes a REST API with OpenAPI documentation and WebSocket streams for real-time simulation control. Database migrations follow an Alembic-based workflow. CLI utilities support agent creation, simulation control and health monitoring. This backend implements the simulation engine, agent runtime, belief update mechanisms and data persistence layer required by our method.

The client, which is shown in Figure 1 below, is a Tauri desktop application written in TypeScript. It provides real-time visualization and interactive control for live simulations. The client connects to the server API and renders agent positions on geographic maps, exploration trajectories, knowledge acquisition events and inter-agent communication patterns. Visualization modes include 2D maps, 2.5D elevated views and 3D globe rendering using MapLibre. Inspector windows expose individual agent state including current goals, accumulated knowledge, decision history and world model confidence. The client-server separation supports independent iteration on visualization and agent logic while preserving a stable communication protocol.

Figure 1: The Tauri desktop application, which serves as the client (frontend) to the VW (server backend).

The geographic state space S can handle millions of Earth features including cities, buildings, roads and natural landmarks stored in PostGIS with spatial indexing. Agents spawn at real Earth locations with latitude, longitude and altitude coordinates. They query their surroundings through spatial range searches that return nearby features and geographic context. This grounds the VW in real-world structure and allows agents to develop spatial understanding that may transfer to other worlds with similar topological properties.

Agents operate autonomously without hand-coded behaviors. Each agent maintains internal modules for perception, reasoning, decision-making, learning and communication. Perception modules process observations from the spatial database and knowledge sources. Reasoning modules generate hypotheses about world structure and predict outcomes of unexplored actions. Decision modules select actions based on expected information gain, curiosity signals and goal alignment. Learning modules

update world models based on prediction errors and novel observations. Communication modules enable knowledge sharing and collaborative exploration among agents.

The action space A includes movement to new locations, queries to knowledge sources (Wikipedia, web search and social media APIs), communication with other agents and world model updates. Movement actions modify agent position coordinates. Knowledge queries return text that agents encode into vector embeddings and integrate into episodic and semantic memory. Communication actions broadcast information to nearby agents or targeted groups. World model updates revise beliefs about spatial structure, causal relationships and reward distributions.

Agents are not LLM wrappers. They implement true autonomous learning through reinforcement learning, curiosity-driven exploration and model-based planning. Policies emerge from experience rather than prompt engineering. This ensures agents develop genuine exploration strategies rather than mimicking human-described behaviors. The system supports parameter variation that creates diverse exploration archetypes including pathfinders (high spatial coverage), surveyors (high detail orientation), pioneers (high risk tolerance) and specialists (domain-focused learning). These archetypes represent different meta-strategies for world exploration encoded through curiosity weights, learning rates, exploration biases and backtracking tolerances.

The system integrates real-world knowledge as priors. External knowledge sources seed agent world models with background information about geographic regions, semantic concepts and causal patterns. Agents query these sources during exploration to reduce uncertainty about unfamiliar features. The availability of multiple knowledge sources tests whether agents can synthesize information from heterogeneous inputs and whether access to richer priors improves exploration efficiency and transfer capability.

All simulation runs produce reproducible traces that capture agent trajectories, action sequences, observation histories, belief updates and estimated utilities. Traces include timestamps, location coordinates, knowledge acquisition events, communication records and decision justifications. These logs enable post-hoc analysis of whether agents make rational entry decisions given their beliefs, whether they gather sufficient evidence before committing to regime shifts and whether their exploration strategies align with the formal models of cross-world utility maximization. Trace data supports evaluation of agent readiness for deployment to alternative worlds and validation of the indirect utility maximization framework.

4 Reflection

The VW demonstrates that cross-world exploration can be formalized as a computational problem and implemented as a trainable system. This shifts the boundary-crossing limitation from a theoretical barrier to an engineering challenge. The approach shows that second-order agents can be instantiated with explicit utility inheritance, that exploration strategies can emerge from autonomous learning rather than scripted behaviors and that evidence from agent exploration can be captured in auditable traces suitable for human decision-making.

The implementation reveals several challenges inherent in training agents for regime shifts. First is the problem of belief calibration. Agents must estimate expected utility in worlds they have not entered. Their estimates depend on priors drawn from the current world, which may not transfer. The system addresses this through curiosity-driven exploration that actively seeks prediction errors and through world models that track confidence levels. Agents learn to distinguish between well-modeled and poorly-modeled regions, which enables rational decisions about when estimates are reliable enough to justify entry.

Second is the challenge of goal preservation across regime boundaries. Agents optimized for one world may develop subgoals or instrumental values that do not align with the primary agent's objectives in a different world. The design mitigates this through explicit utility functions tied to knowledge acquisition and transferable representations rather than world-specific metrics. By rewarding information gain and spatial coverage, the system encourages agents to build models that generalize rather than overfit to the training world's particular constraints.

Third is the evaluation problem. How do we determine whether an agent is ready to explore a new world? The system produces readiness metrics based on knowledge coverage, exploration depth, spatial competence and curiosity maintenance. These metrics quantify whether an agent has developed sufficient modeling capacity to handle unfamiliar environments. However, readiness in one world does not guarantee success in another. The formal framework suggests that readiness should be assessed relative to the target world's expected structure, which requires partial information about that world before full commitment.

The architecture exposes a tension between exploration efficiency and transfer capability. Agents optimized for rapid exploration in the current world may develop strategies that are brittle when world parameters change. Agents optimized for robust world modeling may explore slowly and accumulate less evidence per unit time. The exploration archetypes in the system represent different positions along this tradeoff. Pathfinders maximize spatial coverage at the cost of detail. Surveyors maximize detail at the cost of breadth. Pioneers maximize novelty-seeking at the cost of systematic coverage. Each archetype embodies a different meta-strategy for handling the exploration-transfer tradeoff and the optimal choice depends on the structure of the target world and the transition costs involved.

The system also reveals that multi-agent collaboration changes the dynamics of cross-world exploration. When agents share knowledge, each agent benefits from the exploration of others without paying full exploration costs. This creates opportunities for division of labor where different agents explore different regions or test different hypotheses. However, it also introduces coordination problems and the risk of correlated failures if agents share faulty priors. The communication infrastructure in the VW allows testing of collaborative exploration protocols, including knowledge sharing, joint planning and delegated search.

One limitation of the current implementation is that agents train in a single world. The design includes cross-world transitions and utility calculations, but empirical validation requires deploying agents to a second world with different state structure, action constraints or reward topology. This would test whether agents trained in the VW can actually transfer their exploration strategies to unfamiliar regimes. It would also allow measurement of the value of information gained in the first world and validation of the entry decision models.

Another limitation is the assumption of static world structure. In reality, worlds evolve. Entry decisions must account for the possibility that a world's payoff landscape changes after entry, invalidating beliefs formed during exploration. The formal model can be extended to include temporal dynamics, but this requires agents to maintain beliefs not only about current world parameters but also about rates of change and regime stability.

The approach has implications for human agency extension beyond the specific case of geographic exploration. The same framework applies to any situation where humans face strategy spaces they cannot directly enter. This includes virtual economies, simulated environments for policy testing, abstract problem domains in mathematics or engineering and potential physical regimes such as extreme environments or distant locations. In each case, the challenge is to train agents that inherit human objectives, explore the inaccessible regime and return evidence that supports rational entry decisions.

The VW also provides a platform for studying the conditions under which world exploration is justified. By varying transition costs, world similarity, prior quality and time horizons, we can map the parameter space where regime shifts improve expected utility versus cases where remaining in the current world dominates. This empirical analysis complements the theoretical model and provides practical decision rules for when to invest in cross-world exploration.

A broader implication is that formalizing cross-world exploration as indirect utility maximization changes how we think about agent design. Traditional agent architectures optimize for performance in a fixed environment. This approach optimizes for adaptability across environments. It requires agents that maintain meta-level representations of world structure, that reason about their own uncertainty and learning capacity and that make strategic decisions about when to explore versus exploit. These capabilities extend beyond domain-specific optimization and constitute a form of general intelligence grounded in decision theory.

The system demonstrates that VWs can serve as controlled testbeds for boundary-crossing strategies. By providing parametric variation, full instrumentation and iterative resets, VWs enable systematic experimentation with exploration policies, belief update rules and transfer learning mechanisms. This is difficult to achieve in real-world settings where regime shifts are costly, irreversible or impossible. The VW therefore serves not only as a training ground for agents but also as an experimental platform for understanding cross-world exploration as a general phenomenon.

5 Conclusion

We formalized the cross-world limitation as a barrier to exploration and demonstrated that it can be addressed through second-order agents trained in a VW. The formal model treats worlds as strategy spaces with explicit transition costs and entry rules based on expected utility under uncertainty. The implementation shows that agents exhibiting both rational and intelligent properties can learn boundary-crossing strategies autonomously.

The contribution is methodological. We provide infrastructure for training agents that explore on behalf of primary agents and a framework for evaluating when such exploration is justified. The method applies wherever humans face inaccessible strategy spaces. The VW serves as both training ground and experimental platform for systematic investigation of cross-world exploration as a computational problem.

Three findings matter. First, exploration strategies emerge from autonomous learning rather than hand-coded heuristics. Second, belief calibration under regime uncertainty is tractable through curiosity-driven exploration and confidence-aware world models. Third, multi-agent collaboration changes the cost structure of cross-world exploration by enabling division of labor and knowledge sharing.

The work shifts cross-world exploration from a theoretical constraint to an engineering challenge. Whether agents trained in VWs successfully transfer to truly alien regimes remains empirical. But, the infrastructure exists to test it. The question is no longer whether boundary-crossing is possible. It is whether we can design agents that do it effectively and the evidence they produce supports rational entry decisions.

Cross-world exploration is now tractable. This opens investigation into a landscape of worlds that were previously beyond reach.

REFERENCES

[1] Herbert Gintis. 2000. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton University Press, Princeton, NJ.

[2] Herbert Gintis. 2009. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction* (2nd ed.). Princeton University Press, Princeton, NJ.

[3] Simon Haslam and Ben Shenoy. 2018. *Strategic decision making: A discovery-led approach to critical choices in turbulent times*. Kogan Page Publishers, London.

[4] I. L. Humberstone. 1983. Inaccessible worlds. *Notre Dame Journal of Formal Logic*, 24, 3 (July 1983).

[5] Christopher Kiekintveld and Michael P. Wellman. 2008. Selecting strategies using empirical game models: An experimental analysis of meta-strategies. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '08)*, Estoril, Portugal, May 12-16, 2008.

[6] Olaf Kühne. 2025. *The theory of three landscapes*. In Olaf Kühne, Florian Weber, Karsten Berr, and Corinna Jenal (Eds.), Landscape handbook. Springer International Handbooks of Human Geography. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-83147-8_18

[7] Falk Lieder and Thomas L. Griffiths. 2017. Strategy selection as rational metareasoning. *Psychological Review*, 124, 6 (2017), 762–794. DOI: https://doi.org/10.1037/rev0000075

[8] Annie Luciani, Daniela Urma, Sylvain Marlière, and Joël Chevrier. 2004. PRESENCE: The sense of believability of inaccessible worlds. *Computers & Graphics*, 28, 4 (August 2004), 509-517. DOI: https://doi.org/10.1016/j.cag.2004.04.006

[9] Zhihan Lv, Shuxuan Xie, Yuxi Li, M. Shamim Hossain, and Abdulmotaleb El Saddik. 2022. Building the metaverse using digital twins at all scales, states, and relations. *Virtual Reality & Intelligent Hardware*, 4, 6 (December 2022), 459-476. DOI: https://doi.org/10.1016/j.vrih.2022.07.002