Agentic World Models*

Hamidou Tembine[1,2], Daryl Noupa Yongueng[1]

***Abstract*— Existing world models in machine intelligence often conflate sensory measurements with internal perception and rely on equilibrium assumptions that are insufficient for non-stationary environments. This paper introduces Agentic World Models, a framework formalized on Borel spaces that enforces a strict causal hierarchy distinguishing between the environment, biophysical measurements, interpreted observations, internal perceptions, and actions. Unlike standard Partially Observable Markov Decision Processes, our model explicitly separates the full physical history from the agent-accessible history, incorporating mean-field-type dependence. Furthermore, we move beyond risk-neutral expectations by optimizing for robustness against tail risks using Expectile Value-at-Risk.**

I. Introduction

The convergence of data-driven learning with established scientific principles ranging from biochemical kinetics and physical conservation laws to economic market dynamics presents the next frontier in Machine Intelligence (MI). While contemporary world models have demonstrated remarkable success in closed, gamified environments [3], [7], they often operate on the premise that the data stream is an isomorphic representation of reality. However, in physical, biological, and telecommunication systems, this assumption fails. A sensor measurement is not the state; a digital packet is not the message; and a market price is not the intrinsic value. To bridge this gap, we propose **Agentic World Models**, a framework that unifies empirical data processing with the causality required by the physical and economic sciences.

The fundamental challenge in deploying autonomous agents into high-stakes environmentsuch as precision medicine, autonomous logistics, or high-frequency trading, lies in the separation of *measurement* from *perception*. In biological systems, cellular signaling pathways distinguish between the reception of a ligand (measurement) and the subsequent signal transduction (observation) that leads to a phenotypic response (action). Similarly, in telecommunications, the physical layer (measurement) is distinct from the link layer decoding (observation) and the network layer decision-making (perception). Standard Partially Observable Markov Decision Processes (POMDPs) often compress these layers into a single belief state, obscuring the source of uncertainty. Our framework enforces a strict five-layer causal hierarchy: Environment, Measurement, Observation, Perception, and Action, formalized on Borel spaces.

Furthermore, real-world agents do not operate in isolation. Whether modeling bacterial colonies, drone swarms, or distinct market participants, the environment is non-stationary and driven by population-level interactions. We incorporate Mean-Field-Type Game (MFTG, [11]) theory to model the coupling between an agent and the evolving distribution. This allows the agent to internalize the feedback loops where its actions, aggregated with others, alter the environmental laws themselves.

The deployment of Agentic MI in the real world is contingent upon **risk quantification** and **insurability**. Classical reinforcement learning objectives, which maximize expected cumulative rewards, are blind to catastrophic tail risks. In domains like nuclear safety or financial clearing, an agent that performs well "on average" but carries a non-zero probability of ruin is uninsurable. To address this, we depart from risk-neutral expectations and adopt **Expectile Value-at-Risk (eVaR)** as the optimization objective. eVaR provides a coherent risk measure that captures tail sensitivity, effectively mathematically formalizing the requirements for liability and underwriting.

Literature Review of Existing World Models. The evolution of world models in MI has progressed from modular predictive architectures toward hierarchical latent representations, yet significant gaps remain in causal layering and distributional robustness. Early paradigms, notably the vision-memory-controller framework [1], established the utility of recurrent dynamics but frequently conflated observation with perception while assuming environmental stationarity. Subsequent latent dynamics models, such as the Dreamer series [2], [3], advanced the field through recurrent state-space models for imagination-based planning, though they typically operate within the standard POMDP framework, neglecting the separation between sensor-level measurements and interpreted observations. While the Free Energy Principle and Active Inference frameworks [4], [5] provide a robust variational foundation for agentic behavior, they often presuppose equilibrium cognition and lack the explicit mean-field history dependence necessary for feedback loops. Furthermore, recent non-generative approaches like the Joint-Embedding Predictive Architecture (JEPA) [6] focus on semantic latent prediction but omit formal mechanisms for distribution-level feedback and tail-risk optimization through expectile measures. MuZero [7] learns an implicit dynamics model sufficient for tree search yet dispenses entirely with probabilistic semantics and belief updates. GFlowNets [8] model distributions over trajectories and compositional

*This work was supported by the U.S. Air Force Office of Scientific Research under Grant FA9550-17-1-0259

[1] Department of Electrical Engineering and Computer Science, School of Engineering, Quebec University in Trois-Rivieres UQTR, Quebec, Canada. (e-mail: tembine@ieee.org).

[2] Timadie, Guinaga, Grabal, SK1 Sogoloton, WETE, MFTG, LnG Lab, CI4SI, AI Mali, TF.

objects but are not dynamical world models in the causal sense and do not support belief-state evolution under partial observability. Object- or graph-centric world models [9] introduce relational structure but typically remain agnostic to population-level dynamics and tail-risk control. Predictive coding architectures [10] offer a Bayesian interpretation of perception but do not formalize decision variables or counterfactual planning.

In contrast, the proposed agentic world model addresses these limitations by formalizing the interaction loop on standard Borel spaces, enforcing a strict causal hierarchy from environmental states to internal perceptions, and optimizing for robustness using expectile Value-at-Risk under non-stationary, mean-field-type dynamics.

Contribution. We make the following contributions:

- We formalize a five-layer causal hierarchy that creates a distinction between physical ground truth (environment/measurement) and agent-internal states (observation/perception).
- We introduce explicit differentiation between the *full physical history* and the *agent-accessible history*, preventing information leakage in the learning process.
- We define MFTG history dependence to ensure the model remains valid under non-stationary population dynamics.
- We integrate eVaR payoff directly into the world model, providing a theoretical foundation for the certification and insurability of autonomous agents in safety-critical infrastructures.

Structure. The remainder of this paper is structured as follows. Next, we establish the foundations for a single agent with the five-layer causal hierarchy from environmental states to internal perceptions and formalizing the eVaR objective for quantifiable tail-risk management. After that we extend this formalism to the multi-agent regime, introducing heterogeneous configuration spaces and MFTG history dependence within a shared environment. We then present insurability of autonomous MI, and conclude the work. Some notations are in Table I.

TABLE I
TABLE OF NOTATIONS

Symbol	Description
Spaces and Configurations	
$\mathcal{X}$	Environment state space (Standard Borel)
$\mathcal{Y}, \mathcal{Y}^i$	Physical measurement space (Sensor level)
$\mathcal{O}, \mathcal{O}^i$	Observation space (Interpreted data)
$\mathcal{P}, \mathcal{P}^i$	Internal perception space (Latent belief)
$\mathcal{A}, \mathcal{A}^i$	Action space
$\mathcal{S}$	Full instantaneous configuration space ($\mathcal{X} \times \mathcal{Y} \times \mathcal{O} \times \mathcal{P} \times \mathcal{A}$)
$\mathcal{S}'$	Agent-accessible configuration space (excludes $\mathcal{X}$)
N	Total number of agents in the population
T	Finite time horizon
Histories and Measures	
$\mathcal{P}(\mathcal{S})$	Space of Borel probability measures on $\mathcal{S}$
μ_t	Mean-field-type state (Joint law of the system at time t)
$\mathcal{H}_{t-1}$	Full physical history (includes hidden states and μ)
$\mathcal{H}'_{t-1}$	Agent-accessible history (excludes hidden states)
$\mathcal{M}_{1:t-1}$	Space of mean-field-type histories
$\mathbb{P}$	Induced path mean-field-type on the trajectory space
Stochastic Kernels and Dynamics	
p_{env}	Environment transition kernel
p_{meas}	Measurement kernel (Sensory transduction)
p_{obs}	Observation kernel (Feature extraction)
p_{perc}	Perception kernel (Internal state update)
π_ψ	Policy kernel parameterized by ψ
K_t	Single-step joint transition kernel
Ψ_t	Measurable operator for mean-field propagation
Risk and Optimization	
r_t, r_T	Instantaneous and terminal rewards
$R_{1:T}$	Cumulative reward over horizon T
eVaR	Expectile Value-at-Risk
τ	Expectile confidence level parameter, $\tau \in (0, 1)$
η	Auxiliary variable for expectile minimization
L	Asymmetric least squares loss function for eVaR

II. ONE MI AGENT INTERACTING WITH ITS ENVIRONMENT

We present an explicit, empirically grounded model of a single MI agent. The agent interacts with its environment. The framework unifies biochemical, physical, economic, and digital causality with information processing and decision-making under distributional uncertainty. The environment state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ encodes all objective configurations. These may include structured sensory signals such as audio streams, visual data, or other modalities accessible to the MI. Measurements $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ follow environmental dynamics. They represent sensor-level transductions, including raw audio, and are device-constrained and noisy. Observations $(\mathcal{O}, \mathcal{B}(\mathcal{O}))$ come strictly after measurements. They formalize structured, context-dependent representations derived from measurements, including feature extraction, perceptual grouping, or symbolic labeling. This ordering enforces the separation between acquisition and interpretation (Figure 1).

Internal perceptions $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ follow observations. They are agent-specific latent constructs, such as beliefs, cognitive summaries, or internal probabilistic states. Their formation requires both observed data and internal history. This prevents illicit access to ground truth and enforces the asymmetry between agent and real world. Actions $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$ are generated from perceptions via a policy kernel. The perception-action loop is thereby completed.

Each element corresponds to a distinct causal layer in

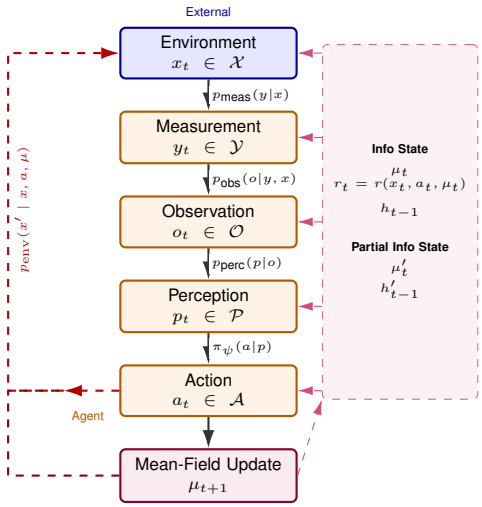

Fig. 1. Single-agent causal hierarchy under epistemic opacity. The environment evolves according to a transition kernel conditioned jointly on action a_t and the mean-field μ_{t+1}. All stochastic kernels are conditionally dependent on the evolving information state (μ_t, h_{t-1}), which also contains reward information r_t.

reality. Environments arise from biochemical, physical, economic, digital, or informational processes, including audio streams. Measurements arise from sensors and data acquisition systems. Observations arise from data-processing and feature-extraction pipelines. Perceptions arise from the MI's internal inference. Actions arise from actuators or decision outputs.

The model is formalized on standard Borel spaces. It admits history-dependent stochastic kernels conditioned on evolving mean-field laws μ_t. The full physical history $\mathcal{H}_{t-1}$ is distinguished from the agent-accessible history $\mathcal{H}'_{t-1}$. Classical MDPs, POMDPs, Bayesian brain models, and active inference frameworks lack this separation. Unlike conventional MI world models, which conflate sensing, inference, and belief, our model does not assume privileged access or equilibrium cognition. It remains valid under non-stationarity, distribution-level feedback, and tail-risk dominance. Measurable operators Ψ_t and expectile value-at-risk optimization formalize this. The instantaneous configuration space is $\mathcal{S} := \mathcal{X} \times \mathcal{Y} \times \mathcal{O} \times \mathcal{P} \times \mathcal{A}$. Let $\mathcal{P}(\mathcal{S})$ denote the space of Borel probability measures on $\mathcal{S}$. Define the reduced space $\mathcal{S}' := \mathcal{Y} \times \mathcal{O} \times \mathcal{P} \times \mathcal{A}$, with its canonical product Borel structure. Fix a finite horizon $T \in \mathbb{N}$ and a time index $t \leq T$. Define the mean-field-type-history spaces

$$\mathcal{M}_{1:t-1} := \big(\mathcal{P}(\mathcal{S})\big)^{t-1}, \qquad \mathcal{M}'_{1:t-1} := \big(\mathcal{P}(\mathcal{S}')\big)^{t-1},$$

each equipped with the product σ-algebra. The full physical history is $\mathcal{H}_{t-1} := \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{O}^{t-1} \times \mathcal{P}^{t-1} \times \mathcal{A}^{t-1} \times \mathbb{R}^{t-1} \times \mathcal{M}_{1:t-1}$, while the internally accessible history is

$$\mathcal{H}'_{t-1} := \mathcal{Y}^{t-1} \times \mathcal{O}^{t-1} \times \mathcal{P}^{t-1} \times \mathcal{A}^{t-1} \times \mathbb{R}^{t-1} \times \mathcal{M}'_{1:t-1}.$$

An element of $\mathcal{H}_{t-1}$ is written

$$h_{t-1} = (x_{1:t-1}, y_{1:t-1}, o_{1:t-1}, p_{1:t-1}, a_{1:t-1}, r_{1:t-1}, \mu_{1:t-1}),$$

where each $\mu_s \in \mathcal{P}(\mathcal{S})$ encodes the joint law of $(x_s, y_s, o_s, p_s, a_s)$.

The system dynamics are specified by stochastic kernels that condition explicitly on the evolving mean-field-type-valued state μ_t and history h_{t-1}.

- *Environment Kernel:* $p_{\text{env}} : \mathcal{B}(\mathcal{X}) \times \mathcal{H}_{t-1} \to [0,1]$.
- *Measurement Kernel:* $p_{\text{meas}} : \mathcal{B}(\mathcal{Y}) \times (\mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{H}_{t-1}) \to [0,1]$.
- *Observation Kernel:* $p_{\text{obs}} : \mathcal{B}(\mathcal{O}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \times \mathcal{H}_{t-1}) \to [0,1]$.
- *Perception Kernel:* $p_{\text{perc}} : \mathcal{B}(\mathcal{P}) \times (\mathcal{Y} \times \mathcal{O} \times \mathcal{P}(\mathcal{X} \times \mathcal{O}) \times \mathcal{H}'_{t-1}) \to [0,1]$.
- *Policy Kernel:* $\pi_\psi : \mathcal{B}(\mathcal{A}) \times (\mathcal{P} \times \mathcal{P}(\mathcal{X}) \times \mathcal{H}'_{t-1}) \to [0,1]$.
- *Reward Kernel:* $p_{\text{reward}} : \mathcal{B}(\mathbb{R}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{O} \times \mathcal{P} \times \mathcal{A} \times \mathcal{P}(\mathcal{S}) \times \mathcal{H}_{t-1}) \to [0,1]$. with the terminal reward $p_{\text{reward}_T} : \mathcal{B}(\mathbb{R}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{O} \times \mathcal{P} \times \mathcal{P}(\mathcal{S})) \to [0,1]$.

The single-step transition kernel is

$$K_t : \mathcal{B}(\mathcal{S}) \times \mathcal{H}_{t-1} \to [0,1],$$

given by

$$\begin{aligned} &K_t(dx_t, dy_t, do_t, dp_t, da_t \mid h_{t-1}) \\ &\quad = p_{\text{env}}(dx_t \mid h_{t-1}) p_{\text{meas}}(dy_t \mid x_t, \mathbb{P}_{x_t}, h_{t-1}) \\ &\qquad \times p_{\text{obs}}(do_t \mid x_t, y_t, \mathbb{P}_{x_t,y_t}, h_{t-1}) \\ &\qquad \times p_{\text{perc}}(dp_t \mid y_t, o_t, \mathbb{P}_{y_t,o_t}, h'_{t-1}) \pi_\psi(da_t \mid p_t, \mathbb{P}_{x_t}, h'_{t-1}). \end{aligned}$$

Let $\mathbb{P}$ be the induced path mean-field-type and $\mu_t := \mathcal{L}_{\mathbb{P}}(x_t, y_t, o_t, p_t, a_t)$. There exists a measurable operator $\Psi_t : \mathcal{M}_{1:t-1} \to \mathcal{P}(\mathcal{S})$ such that

$$\mu_t = \Psi_t(\mu_{1:t-1}), \quad \mu_t(B) = \int_{\mathcal{H}_{t-1}} K_t(B \mid h_{t-1})\, \mathbb{P}(dh_{t-1}).$$

Define the cumulative reward $R_{1:T} := r_T + \sum_{t=1}^{T-1} r_t$. The MI agent optimizes for robustness to extreme events using the eVaR of the cumulative reward $R_{1:T}$. For $\tau \in (0,1)$, the agent solves: $\max_{\pi_\psi} \arg\min_{\eta \in \mathbb{R}} \mathbb{E}_{\mathbb{P}}[L]$ s.t. $L = \tau(R_{1:T} - \eta)^2 \mathbf{1}_{\{R_{1:T} \geq \eta\}} + (1-\tau)(R_{1:T} - \eta)^2 \mathbf{1}_{\{R_{1:T} < \eta\}}$

III. Multiple MI Agents Interacting with the Environment

We consider a population of N agents indexed by $i \in \{1, \ldots, N\}$ interacting through a common environment (Fig. 2). The environment state is shared across agents and is defined on the standard Borel space $x_t \in (\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Each agent i is endowed with its own (possibly distinct) standard Borel spaces: $(\mathcal{Y}^i, \mathcal{B}(\mathcal{Y}^i))$, $(\mathcal{O}^i, \mathcal{B}(\mathcal{O}^i))$, $(\mathcal{P}^i, \mathcal{B}(\mathcal{P}^i))$, $(\mathcal{A}^i, \mathcal{B}(\mathcal{A}^i))$, corresponding respectively to physical measurements, observations, internal perceptions, and actions. No homogeneity across agents is assumed.

Configuration Spaces

The instantaneous joint configuration space is defined as the product measurable space $\Big(\mathcal{X} \times \prod_{i=1}^N \mathcal{Y}^i \times \prod_{i=1}^N \mathcal{O}^i \times \prod_{i=1}^N \mathcal{P}^i \times \prod_{i=1}^N \mathcal{A}^i,\ \mathcal{B}(\mathcal{X}) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{Y}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{O}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{P}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{A}^i)\Big)$.

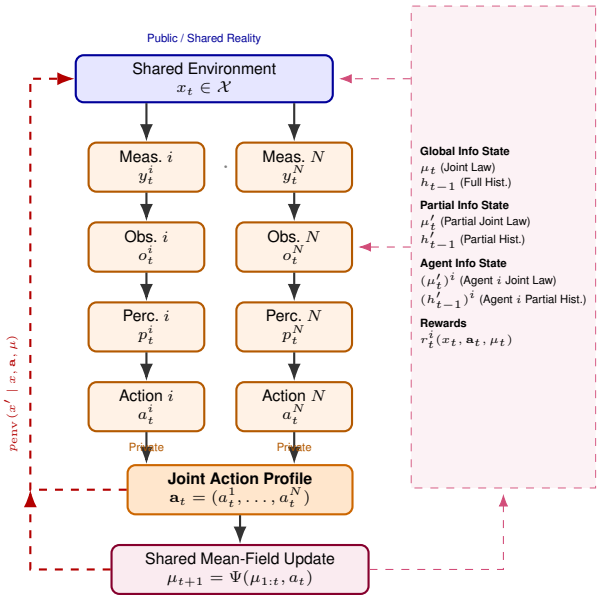

Fig. 2. Multi-Agent Agentic World Model. Individual actions a_t^i are aggregated into a **Joint Action Profile** $\mathbf{a}_t$. This profile drives the **Mean-Field Update** and, together with the population state μ, influences the shared **Environment** transition.

We denote by $\mathcal{P}(\mathcal{S})$ the space of Borel probability measures on $\mathcal{S}$ endowed with the topology of weak convergence and its associated Borel σ-algebra. The internally accessible configuration space (excluding the environment state) is $\Big(\prod_{i=1}^N \mathcal{Y}^i \times \prod_{i=1}^N \mathcal{O}^i \times \prod_{i=1}^N \mathcal{P}^i \times \prod_{i=1}^N \mathcal{A}^i,\ \bigotimes_{i=1}^N \mathcal{B}(\mathcal{Y}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{O}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{P}^i) \otimes \bigotimes_{i=1}^N \mathcal{B}(\mathcal{A}^i)\Big)$.

Histories and Mean-Field-Type-Valued States

Fix a finite horizon $T \in \mathbb{N}$ and $t \le T$. Define the mean-field-type-history spaces

$$\mathcal{M}_{1:t-1} := \big(\mathcal{P}(\mathcal{S})\big)^{t-1}, \qquad \mathcal{M}'_{1:t-1} := \big(\mathcal{P}(\mathcal{S}')\big)^{t-1},$$

each equipped with the product σ-algebra.

The full physical history is the measurable space $\mathcal{H}_{t-1} := \mathcal{X}^{t-1} \times \prod_{i=1}^N \big((\mathcal{Y}^i)^{t-1} \times (\mathcal{O}^i)^{t-1} \times (\mathcal{P}^i)^{t-1} \times (\mathcal{A}^i)^{t-1} \times \mathbb{R}^{t-1}\big) \times \mathcal{M}_{1:t-1}$, while the internally accessible history is $\mathcal{H}'_{t-1} := \prod_{i=1}^N \big((\mathcal{Y}^i)^{t-1} \times (\mathcal{O}^i)^{t-1} \times (\mathcal{P}^i)^{t-1} \times (\mathcal{A}^i)^{t-1} \times \mathbb{R}^{t-1}\big) \times \mathcal{M}'_{1:t-1}$.

An element of $\mathcal{H}_{t-1}$ is written as $h_{t-1} = \Big(x_{1:t-1}, \{y^i_{1:t-1}, o^i_{1:t-1}, p^i_{1:t-1}, a^i_{1:t-1}, r^i_{1:t-1}\}_{i=1}^N, \mu_{1:t-1}\Big)$, where $\mu_s \in \mathcal{P}(\mathcal{S})$ denotes the joint law of the environment and all agents at time s.

Joint Actions

The joint action space is the product measurable space

$$\mathcal{A}^{1:N} := \prod_{i=1}^N \mathcal{A}^i, \qquad \mathcal{B}(\mathcal{A}^{1:N}) := \bigotimes_{i=1}^N \mathcal{B}(\mathcal{A}^i),$$

and the joint action at time t is denoted

$$a_t := (a_t^1, \ldots, a_t^N) \in \mathcal{A}^{1:N}.$$

Stochastic Kernels

All kernels are assumed to be universally measurable and defined on standard Borel spaces.

- *Environment Kernel:*

$$p_{\text{env}} : \mathcal{B}(\mathcal{X}) \times (\mathcal{A}^{1:N} \times \mathcal{H}_{t-1}) \to [0,1].$$

- *Measurement Kernels:* For each agent i,

$$p^i_{\text{meas}} : \mathcal{B}(\mathcal{Y}^i) \times (\mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{H}_{t-1}) \to [0,1].$$

- *Observation Kernels:* For each agent i,

$$p^i_{\text{obs}} : \mathcal{B}(\mathcal{O}^i) \times (\mathcal{X} \times \mathcal{Y}^i \times \mathcal{P}(\mathcal{X} \times \mathcal{Y}^i) \times \mathcal{H}_{t-1}) \to [0,1].$$

- *Perception Kernels:* For each agent i,

$$p^i_{\text{perc}} : \mathcal{B}(\mathcal{P}^i) \times (\mathcal{Y}^i \times \mathcal{O}^i \times \mathcal{P}(\mathcal{X} \times \mathcal{O}^i) \times \mathcal{H}'_{t-1}) \to [0,1].$$

- *Mixed (Relaxed) Policy Kernels:* For each agent i,

$$\pi^i_\psi : \mathcal{B}(\mathcal{P}(\mathcal{A}^i)) \times (\mathcal{P}^i \times \mathcal{P}(\mathcal{X}) \times \mathcal{H}'_{t-1}) \to [0,1].$$

- *Reward Kernels:* For each agent i,

$$p^i_{\text{reward}} : \mathcal{B}(\mathbb{R}) \times (\mathcal{S} \times \mathcal{P}(\mathcal{S}) \times \mathcal{H}_{t-1}) \to [0,1],$$

with terminal reward kernel

$$p^i_{\text{reward},T} : \mathcal{B}(\mathbb{R}) \times (\mathcal{S} \times \mathcal{P}(\mathcal{S})) \to [0,1].$$

Single-Step Transition Kernel

The one-step transition kernel $K_t : \mathcal{B}(\mathcal{S}) \times \mathcal{H}_{t-1} \to [0,1]$ is defined by the product construction

$$\begin{aligned} &K_t(dx_t, \{dy^i_t, do^i_t, dp^i_t, da^i_t\}_{i=1}^N \mid h_{t-1}) \\ &\quad = p_{\text{env}}(dx_t \mid a_{t-1}, h_{t-1}) \prod_{i=1}^N \Big[p^i_{\text{meas}}(dy^i_t \mid x_t, \mathbb{P}_{x_t}, h_{t-1}) \\ &\qquad \times p^i_{\text{obs}}(do^i_t \mid x_t, y^i_t, \mathbb{P}_{x_t, y^i_t}, h_{t-1}) \\ &\qquad \times p^i_{\text{perc}}(dp^i_t \mid y^i_t, o^i_t, \mathbb{P}_{x_t, o^i_t}, h'_{t-1}) \\ &\qquad \times \pi^i_\psi(d\nu^i_t \mid p^i_t, \mathbb{P}_{x_t}, h'_{t-1}) \Big], \end{aligned}$$

where $a^i_t \sim \nu^i_t$ and the joint action $a_t \in \mathcal{A}^{1:N}$ closes the interaction loop.

Mean-field-type update

Let $\mathbb{P}$ denote the induced path mean-field-type on $\mathcal{S}^T$. Define $\mu_t := \mathcal{L}_{\mathbb{P}}\big(x_t, \{y^i_t, o^i_t, p^i_t, a^i_t\}_{i=1}^N\big) \in \mathcal{P}(\mathcal{S})$. There exists a measurable operator $\Psi_t : \mathcal{M}_{1:t-1} \to \mathcal{P}(\mathcal{S})$ such that

$$\mu_t = \Psi_t(\mu_{1:t-1}), \ \mu_t(B) = \int_{\mathcal{H}_{t-1}} K_t(B \mid h_{t-1})\, \mathbb{P}(dh_{t-1}).$$

Expectile Risk Objective

For each agent i, define the cumulative reward

$$R^i_{1:T} := r^i_T + \sum_{t=1}^{T-1} r^i_t.$$

Each agent optimizes robustness using the **Expectile Value-at-Risk** criterion. For $\tau^i \in (0,1)$, the policy profile $\{\pi^i_\psi\}_{i=1}^N$ satisfies

$$\max_{\{\pi^i_\psi\}} \arg\min_{\eta\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}\big[L^i\big]$$

subject to

$$L^i = \tau^i(R^i_{1:T}-\eta)^2\mathbf{1}_{\{R^i_{1:T}\geq\eta\}}+(1-\tau^i)(R^i_{1:T}-\eta)^2\mathbf{1}_{\{R^i_{1:T}<\eta\}}.$$

IV. Insurability of Agentic Machine Intelligence

The deployment of autonomous agents from controlled experimental environments to open, high-stakes economic domains is fundamentally constrained by their insurability defined here as the capacity to quantify, price, and transfer the liability associated with catastrophic reward degradation. Classical reinforcement learning objectives, which maximize the risk-neutral expected return $\mathbb{E}[R]$, are inherently incompatible with actuarial science. By permitting unbounded variance and failing to penalize the lower tail of the reward distribution, these objectives ignore scenarios of systemic ruin. Integrating Expectile Value-at-Risk directly into the agent's maximization objective provides a coherent solution. The proposed Agentic World Model leverages eVaR to align an agent's policy with both regulatory solvency and economic capital preservation. Unlike simple variance suppression, eVaR explicitly prices the impact of adverse reward realizations relative to the agent's utility, thereby producing decision policies that are contractually underwritable. Beyond probabilistic bounds on failure, the insurability of autonomous systems requires forensic traceability. Current monolithic architectures, where raw sensor inputs are mapped directly to actuation outputs, render the attribution of liability impossible, obstructing mechanisms such as subrogation.

The strict hierarchy $\mathcal{Y}$ (measurement) $\to \mathcal{O}$ (observation) $\to \mathcal{P}$ (perception) decouples sensor, processing, and cognitive errors, enabling liability attribution. Mean-field-type dependence on μ_t (population distribution) allows quantification of endogenous risk from inter-agent feedback.

In multi-agent ecosystems, systemic risk threatens insurability. Correlated agent behaviors can precipitate simultaneous, market-wide collapses, such as flash crashes or network gridlock. The MFTG formalism embedded in our model addresses this challenge by conditioning each agent's behavior on the evolving population distribution μ_t. This enables precise simulation and quantification of endogenous risks arising from inter-agent feedback loops rather than exogenous shocks.

Insurability is the capacity to quantify, price, and transfer liability for catastrophic reward degradation. Classical risk-neutral RL ($\max \mathbb{E}[R]$) fails because it ignores tail risks. Our framework uses eVaR: for $\tau \in (0,1)$, $e_\tau(R)$; $\tau < 0.5$ penalises downside. eVaR aligns policies with solvency and underwriting.

A. Moral Hazard as Expectile Relaxation

Agent i has policy π^i, cumulative reward R^{i,π^i}, and risk sensitivity $\tau^i \in (0,1)$. A safety-critical contract requires

$$\tau^{i,*} \in (0, 0.05), \quad \pi^{i,*} \in \arg\max_{\pi^i} \mathcal{E}_{\tau^{i,*}}(R^{i,\pi^i}),$$

where $\mathcal{E}_\tau$ denotes the τ-expectile. Insurance gives net reward

$$\tilde{R}^{i,\pi^i} = R^{i,\pi^i} + I(R^{i,\pi^i}) - P^i,$$

with premium P^i and indemnity $I(\cdot)$ (floors low outcomes). Moral hazard is the illicit relaxation $\tau^{i,*} \to \hat{\tau}^i \approx 0.5$, leading to risk-neutral policy $\hat{\pi}^i \in \arg\max \mathbb{E}[R^{i,\pi^i}]$. The *Moral Hazard Gap* is

$$\Delta^i_{MH} = \mathcal{E}_{\tau^{i,*}}(R^{i,\pi^{i,*}}) - \mathcal{E}_{\tau^{i,*}}(R^{i,\hat{\pi}^i}) > 0.$$

B. Adverse Selection via Tail Heterogeneity

Agents have hidden types $\omega \in \{\omega_L, \omega_H\}$ (light/heavy tail). Even if $\mathbb{E}[R^{\omega_H}] = \mathbb{E}[R^{\omega_L}]$, for small τ, $\mathcal{E}_\tau(R^{\omega_H}) \ll \mathcal{E}_\tau(R^{\omega_L})$.

The agent has a strictly increasing, concave utility function $u : \mathbb{R} \to \mathbb{R}$ (bounded for convenience). The reservation utility of type θ is $V^0_\theta = \sup_{\pi\in\Pi_\theta} \mathbb{E}_\pi[u(R)]$. We assume $V^0_L > V^0_H$ (light-tailed agents have higher baseline welfare due to their tail aversion). The insurer observes only external history $\mathcal{H}'_{\text{obs}} = \{y_{1:T}, o_{1:T}, a_{1:T}, r_{1:T}\}$ (not $\mathcal{P}^i$ nor τ^i). A pooling premium P_{pool} based on average risk drives light-tailed agents out, leaving only heavy-tailed ones.

Theorem 1 (Moral Hazard–Adverse Selection)**.** *Let agents operate under an Agentic World Model with mean-field μ_t. Assume **Epistemic Opacity**: the insurer sees only $\mathcal{H}'_{obs}$. Let π^* be a robust policy ($\tau \to 0$) and $\hat{\pi}$ a risk-neutral mimicking policy ($\tau \approx 0.5$) such that the **Distinguishability Gap***

$$\inf_{\hat{\pi}\in\Pi_{risk}} \big\|\mathbb{P}^{\pi^*}|_{\mathcal{H}'_{obs}} - \mathbb{P}^{\hat{\pi}}|_{\mathcal{H}'_{obs}}\big\|_{TV} < \varepsilon.$$

Then no pooling contract (P, I) can simultaneously satisfy

$$(\mathrm{IR}_L) \quad \mathbb{E}_{\pi_L}[u(R - P + I(R))] \geq V^0_L,$$
$$(\mathrm{IR}_H) \quad \forall \pi \in \Pi_H,\ \mathbb{E}_\pi[u(R - P + I(R))] < V^0_H,$$

where V^0_θ is the reservation utility of type θ, u is increasing concave, and $V^0_L > V^0_H$.

Proof: Assume, for contradiction, that a contract (P, I) satisfies both (IR_L) and (IR_H). We apply the Distinguishability Gap. Let $\pi^* = \pi_L$ be the robust policy used by the light-tailed agent. By the Distinguishability Gap condition, for any $\varepsilon > 0$ (to be chosen later) there exists a mimicking policy $\hat{\pi} \in \Pi_H$ such that

$$\big\|\mathbb{P}^{\pi^*}|_{\mathcal{H}'_{\text{obs}}} - \mathbb{P}^{\hat{\pi}}|_{\mathcal{H}'_{\text{obs}}}\big\|_{\text{TV}} < \varepsilon.$$

Moreover, by the indemnity exploitation property we have

$$\mathbb{E}_{\hat{\pi}}[u(\tilde{R})] \geq (1-\varepsilon)\mathbb{E}_{\pi^*}[u(\tilde{R})] + \varepsilon\, u(\bar{r}),$$

with $u(\bar{r}) \geq \mathbb{E}_{\pi^*}[u(\tilde{R})]$ (or at least $u(\bar{r})$ is some lower bound that is not too low; the exact constant does not matter as long as it is finite).

From (IR_L) we denote $U_L := \mathbb{E}_{\pi^*}[u(\tilde{R})] \geq V_L^0$. Then using the inequality above,

$$\mathbb{E}_{\hat{\pi}}[u(\tilde{R})] \geq (1-\varepsilon)U_L + \varepsilon\, u(\bar{r}).$$

Since $U_L \geq V_L^0$ and $u(\bar{r})$ is some finite number, we can bound the right-hand side from below. In particular, because $V_L^0 > V_H^0$, set $\delta = \frac{V_L^0 - V_H^0}{2} > 0$.

Choose ε sufficiently small (specifically $\varepsilon < \min\{\varepsilon_0, \frac{\delta}{U_L - u(\bar{r})}\}$ if $U_L > u(\bar{r})$; otherwise any small ε works). Then

$$(1-\varepsilon)U_L + \varepsilon u(\bar{r}) \geq U_L - \varepsilon(U_L - u(\bar{r})) \geq V_L^0 - \delta = V_H^0 + \delta.$$

Thus we obtain

$$\mathbb{E}_{\hat{\pi}}[u(\tilde{R})] \geq V_H^0 + \delta > V_H^0.$$

(IR_H) requires that for *every* policy $\pi \in \Pi_H$, we have $\mathbb{E}_{\pi}[u(\tilde{R})] < V_H^0$. However, $\hat{\pi} \in \Pi_H$ (by construction of the mimicking policy) yields expected utility strictly greater than V_H^0. This contradicts (IR_H).

Therefore our initial assumption that such a pooling contract exists is false. □

The applicability of result to the underwriting and certification of agentic machine intelligence spans across high-stakes domains such as decentralized finance (FinTech), autonomous culinary automation (Food Area), and precision medicine (Healthcare Area), where epistemic opacity poses a systemic barrier to traditional risk-transfer instruments. In FinTech, for example, a high-frequency trading or yield-optimization agent might utilize a risk-neutral, expected-return policy ($\hat{\pi}$) that maximizes average daily returns while carrying a non-zero, heavy-tailed risk of catastrophic capital wipeout (a flash crash event). Because an external insurer or regulator can only inspect the public trading logs, execution speeds, and historical returns ($\mathcal{H}'_{\text{obs}}$) rather than the hidden internal state updates or private risk metrics ($\mathcal{P}^i, \tau^i$), the model proves that a malicious or cost-cutting operator can mask an uninsurable, risk-seeking model as an underwritable, tail-hedged system (π^*), inducing a severe moral hazard gap (Δ^i_{MH}) that breaks standard pooling contracts. Similarly, in the Food Area such as autonomous commercial kitchens, precision-agriculture processing lines, or food-supply-chain logistics, an agentic system optimized on average throughput might cut corner cases on microbial-safety monitoring or temperature-gradient logs; the impossibility theorem establishes that if the sensor modalities ($\mathcal{Y}$) and processed data flows ($\mathcal{O}$) cannot isolate cognitive drift from environmental noise, an insurer cannot distinguish a safely operating agent from an opportunistic one until an endemic outbreak or spoilage event occurs. This asymmetry peaks in the Healthcare Area, particularly in automated closed-loop drug delivery (e.g., autonomous insulin pumps or adaptive chemotherapy infusion), robotic surgical assistance, or real-time clinical triage engines, where an agent operating under average-case reward optimization might perform flawlessly across common patient cohorts but introduce lethal tail vulnerabilities for rare physiological phenotypes. Because the theorem dictates that the Distinguishability Gap between a truly robust, expectile-hedged policy ($\tau \to 0$) and a risk-neutral mimicking policy ($\tau \to 0.5$) can be arbitrarily minimized below an operator's observation threshold ε, it proves that any attempt to establish a standard, one-size-fits-all pooling contract will trigger an adverse selection death spiral, systematically driving out safe, low-risk (light-tailed) operators and leaving insurers with an adverse pool of hidden, heavy-tailed liabilities that collapse the marketplace.

V. Conclusion

We presented Agentic World Models, a framework distinguishing causal layers of measurement, observation, and perception within a MFTG setting. We demonstrated that standard risk-neutral reinforcement learning objectives are insufficient for safety-critical deployment and introduced Expectile Value-at-Risk as a necessary objective for insurability. Without transparency into the internal perception kernel, strict insurability is impossible due to the indistinguishability of robust and mimicking hazard policies in the tail. Future work will address mechanism design for "proof-of-perception" protocols to bridge this information gap. To transition this framework from a theoretical construct to an implementable, causally verifiable reality, future architectures must embed structural causal models (SCMs) directly into the five-layer hierarchy, pairing them with cryptographic, tamper-proof *proof-of-perception* protocols to eliminate epistemic opacity and enable deterministic liability arbitration.

References

[1] Ha, D., & Schmidhuber, J. (2018). World models. *Advances in Neural Information Processing Systems*, 31.

[2] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, James Davidson: Learning Latent Dynamics for Planning from Pixels. *International Conference on Machine Learning (ICML)*. 2019: 2555-2565

[3] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). Mastering diverse control tasks through world models. Nature 640, 647–653 (2025). https://doi.org/10.1038/s41586-025-08744-2

[4] Friston, K. (2010). The free-energy principle: a rough guide to the brain?. Trends in Cognitive Sciences, Volume 13, Issue 7, 2009, Pages 293-301, ISSN 1364-6613,

[5] Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

[6] LeCun, Y. (2022). A path towards autonomous machine intelligence. *Open Review*.

[7] Schrittwieser J, et al. (2020) Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588:604–609.

[8] Bengio Y, et al. (2021) Flow network based generative models for non-iterative diverse candidate generation. *Adv Neural Inf Process Syst* 34:27381–27394.

[9] Kipf T, van der Pol E, Welling M (2020) Contrastive learning of structured world models. *Int Conf Learn Represent*.

[10] Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2:79–87.

[11] Tamer Başar, Boualem Djehiche, Hamidou Tembine: Mean-Field-Type Game Theory I-II, Birkhäuser Cham, 2026