

# On Decision-Making Agents and Higher-Order Causal Processes

Matt Wilson\*

*Université Paris-Saclay, CNRS, ENS Paris-Saclay,  
Inria, CentraleSupélec, Laboratoire Méthodes Formelles*

(Dated: December 12, 2025)

We establish a precise correspondence between decision-making agents in partially observable Markov decision processes (POMDPs) and one-input process functions, the classical limit of higher-order quantum operations. In this identification an agent’s policy and memory update combine into a process function  $w$  that interacts with a POMDP environment via the link product. This suggests a dual interpretation: in the physics view, the process function acts as the environment into which local operations (agent interventions) are inserted, whereas in the AI view it encodes the agent and the inserted functions represent environments. We extend this perspective to multi-agent systems by identifying observation-independent decentralized POMDPs as natural domains for multi-input process functions.

## INTRODUCTION

Agency, the capability of an entity to act upon and receive information from its surroundings, is a fundamental notion in both artificial intelligence and the foundations of physics. In AI, agents interact with partially observable environments to maximise cumulative reward, forming the basis of reinforcement learning and multi-agent systems [1, 2]. In the informational foundations of physics, “agents in laboratories” are modelled as local operations inserted into a spacetime environment, formalised by higher-order quantum operations [3–6]. Both frameworks hinge on the interplay between an agent and its environment, yet they have developed independently and no direct mathematical correspondence between them has been established.

Bridging these viewpoints could bring new ideas to both fields. From a physical perspective, higher-order processes allow causal (even indefinite-causal) structure to be treated as a resource [7–9]. A mapping to decision-making agents opens the door to consideration of multi-agent tasks in which optimal causal and indefinite causal strategies could be learned. From an AI perspective, interpreting agents as higher-order processes suggests a principled route to quantum generalizations of reward-seeking agency [10–12].

On the more formal and logic perspective, the compositional and logical tools developed for higher-order quantum maps [13–19] and their generalisation to arbitrary monoidal categories [20–22] might be utilised in the context of such an identification to provide tools for reasoning about composite multi-agent systems. Conversely, already existing compositional and logic tools for modelling aspects of reward-seeking agents and environments in categorical cybernetics [23–25] and open game theory [26] could be lifted to the quantum domain to bring new techniques and ways of thinking about quantum reinforcement learning [27] and quantum game theory [28].

*Contributions.* In this work we build a precise correspondence between agents and environments in AI and

the foundations of physics. Our main result is a precise correspondence between agent-state deterministic POMDP agents [1, 29] and one-input process functions, the classical limit of higher-order quantum operations [30, 31]. We show that:

1. Every deterministic agent with memory space  $M$ , action set  $A$  and observation set  $\Omega$  in the sense of [1, 29] uniquely determines a one-input process function  $w : (A \rightarrow \Omega) \rightarrow (M \rightarrow M)$  by combining its policy and memory update into a classical comb. Conversely, any one-input process function admits a unique decomposition into a policy and update, yielding a deterministic agent. Two agents are behaviourally equivalent precisely when they induce the same process function.
2. As a result, the same higher-order object admits two dual interpretations: in the physics view,  $w$  represents a spacetime environment into which local operations (agent interventions) are inserted; in the AI view,  $w$  encodes the agent, and the inserted maps represent 1-step evaluations on POMDP environments.
3. Observation-independent decentralised POMDPs provide a natural host for multi-input process functions and thus for multi-agent strategies that may not respect a definite causal order.

These results identify the agents of reinforcement learning with the classical limit of higher-order processes, establishing a direct bridge between agency in AI and agency in physics, with the aim of discovering how causal structure might be used as a resource in AI, whether indefinite causal structures can be efficiently learned, how best to quantize decision-making agents, and how to compose classical and/or quantum multi-agent systems.

## PRELIMINARIES

### Decision-making agents

We will work with deterministic partially observable Markov decision processes (POMDPs) [32]. A deterministic POMDP is defined by the tuple  $\langle S, A, \Omega, T, O, R \rangle$  where  $S$  is a set of states,  $A$  a set of available actions,  $\Omega$  a set of observations and  $T : S \times A \rightarrow S$ ,  $O : S \times A \rightarrow \Omega$  and  $R : S \times A \rightarrow \mathbb{R}$  are deterministic functions specifying the state update, the observation, and the reward associated to performing an action in a given state.

In a fully observable Markov decision process the observation set  $\Omega$  coincides with  $S$ , and an optimal agent can choose an action based solely on the current state. In contrast, in a POMDP the observation  $o \in \Omega$  at a given time may not uniquely determine the underlying state, and so the agent must retain a memory of the history of its past interactions in order to act optimally. In this article we will work with agent-states which generalise to allow for any agent to retain an abstract memory, representing a suitable compression of the history of past observations [1, 29, 33, 34]. Whilst in these works, agent policies and updates are permitted to be stochastic, we will work entirely within the deterministic picture.

Accordingly, a deterministic agent is specified by two maps:

1. A *policy*  $\pi : M \rightarrow A$  that selects an action based on the current memory state.
2. A *memory update*  $\mathcal{U} : M \times A \times \Omega \rightarrow M$  that updates the agent's memory based on the previous memory, the chosen action and the observed outcome.

The pair  $\mathcal{A} = (\pi, \mathcal{U})$  will be called a decision-making agent. Finite-memory agents generalise both reactive policies  $\pi : \Omega \rightarrow A$  and belief-state updates, and they do not require the agent to know the environment model in advance.

Whilst the body of this article focusses on the 1-input case. We will later identify a class of POMDPs on which indefinite causal order strategoes return well-formed logically consistent outcomes, the observation-independent decentralised environments [35].

**Definition 1** (*n*-party deterministic factored dec-POMDP). Fix  $n \geq 2$ . For each party  $i \in \{1, \dots, n\}$  let  $S_i$  be a local state space,  $A_i$  a local action space, and  $\Omega_i$  a local observation space. Define the products

$$S := \prod_{i=1}^n S_i, \quad A := \prod_{i=1}^n A_i, \quad \Omega := \prod_{i=1}^n \Omega_i.$$

A *deterministic factored n-party dec-POMDP* is, simply a POMDP on  $S, A, \Omega$ .

In factored and decentralised POMDPs, we have the opportunity to state constraints on the dependency between actions and observations.

**Definition 2** (Observation independence). The dec-POMDP  $\mathcal{P}$  is *observation independent* if for each  $i$  the component  $O_i^{\mathcal{P}}(a, s)$  is independent of  $a_j$  for all  $j \neq i$ . Equivalently, there exist functions

$$O_i : S \times A_i \longrightarrow \Omega_i \quad (i = 1, \dots, n)$$

such that for all  $a = (a_1, \dots, a_n) \in A$  and  $s \in S$ ,

$$O^{\mathcal{P}}(s, a) = (O_1(s, a_1), \dots, O_n(s, a_n)). \quad (1)$$

*Remark.* Equation (1) is the deterministic expression of the no-signalling constraint  $A_i \not\rightarrow \Omega_k$  for  $i \neq k$  within one environment step.

### Process functions

The framework of higher-order quantum operations allows one to treat quantum channels as resources and to compose them in ways that may not respect a fixed causal order [3, 6, 36]. In the classical deterministic limit these higher-order operations reduce to *process functions*. Such functions are of particular interest because they model deterministic closed timelike curves that remain paradox-free by ensuring the existence of a unique fixed point for any choice of inserted functions [30]. Here we use the precise formulation given in [31].

**Definition 3** (One-input process function). Let  $P, O, F$  and  $I$  be sets. A *one-input process function* of type  $(I \rightarrow O) \rightarrow (P \rightarrow F)$  is a function  $w : P \times O \rightarrow F \times I$  satisfying the following *unique fixed-point condition*. For every function  $f : I \rightarrow O$  and every  $p \in P$ , consider the equation

$$o = f(w_I(p, o)),$$

where  $w_I : P \times O \rightarrow I$  denotes the second component of  $w$ . The map  $w$  is a process function if and only if, for each  $f$  and  $p$ , this equation has a unique solution  $o \in O$ .

Process functions can be generalised to multiple-inputs, in which case they allow for non-causal concatenations. In this article, we will focus on the 1-input case.

**Definition 4** (*n*-input process functions). An *n-input process function* is a function

$$w : P \times \Omega_1 \times \dots \times \Omega_n \longrightarrow F \times A_1 \times \dots \times A_n$$

satisfying the following unique fixed-point condition: for every choice of deterministic functions

$$f_i : A_i \rightarrow \Omega_i \quad (i = 1, \dots, n)$$

and every  $p \in P$ , the system of equations

$$o_i = f_i(w_{A_i}(p, \vec{o})) \quad (i = 1, \dots, n) \quad (2)$$

has a unique solution  $\vec{o} = (o_1, \dots, o_n) \in \Omega$ . Here  $w_{A_i}$  denotes the  $A_i$ -component of  $w$  and we write

$$w(m, o_1, \dots, o_n) = (w_P(p, \vec{o}), w_{A_1}(p, \vec{o}), \dots, w_{A_n}(p, \vec{o})).$$

### CORRESPONDENCE BETWEEN AGENTS AND PROCESS FUNCTIONS

To relate agents to process functions we begin by formalising a single round of interaction. Let  $\mathcal{A} = (\pi, \mathcal{U})$  be a deterministic agent with memory space  $M$  and let  $\mathcal{P} = \langle S, A, \Omega, T, O, R \rangle$  be a deterministic POMDP. Given an initial memory state  $m \in M$  and environment state  $s \in S$ , the one-step interaction proceeds as follows:

1. The agent chooses an action  $a = \pi(m)$ .
2. The environment updates its state to  $s' = T(s, a)$ , outputs an observation  $o = O(s, a)$  and produces a reward  $r = R(s, a)$ .
3. The agent updates its memory according to  $m' = \mathcal{U}(m, a, o)$ .

We write  $\mathcal{A} \bullet \mathcal{P} : M \times S \rightarrow M \times S \times \mathbb{R}$  for the function mapping  $(m, s)$  to  $(m', s', r)$  defined by this procedure. Two agents should be regarded as behaviourally equivalent if they induce the same memory-state update, environment transition, and reward for every possible environment.

**Definition 5.** Let  $\mathcal{A} = (\pi_{\mathcal{A}}, \mathcal{U}_{\mathcal{A}})$  and  $\mathcal{B} = (\pi_{\mathcal{B}}, \mathcal{U}_{\mathcal{B}})$  be two deterministic agents with the same memory space. We say that  $\mathcal{A}$  and  $\mathcal{B}$  are *equivalent*, written  $\mathcal{A} \cong \mathcal{B}$ , if for every deterministic POMDP  $\mathcal{P}$  and every memory-state/environment-state pair  $(m, s)$  the equality

$$\mathcal{A} \bullet \mathcal{P}(m, s) = \mathcal{B} \bullet \mathcal{P}(m, s)$$

holds.

Concretely, for all POMDPs  $\mathcal{P} = \langle S, A, \Omega, T, O, R \rangle$  and all  $m \in M$ ,  $s \in S$  one has

$$\begin{aligned} T(s, \pi_{\mathcal{A}}(m)) &= T(s, \pi_{\mathcal{B}}(m)), \\ R(s, \pi_{\mathcal{A}}(m)) &= R(s, \pi_{\mathcal{B}}(m)), \end{aligned}$$

$$\mathcal{U}_{\mathcal{A}}(m, \pi_{\mathcal{A}}(m), O(s, \pi_{\mathcal{A}}(m))) = \mathcal{U}_{\mathcal{B}}(m, \pi_{\mathcal{B}}(m), O(s, \pi_{\mathcal{B}}(m))).$$

This definition captures the idea that agents can use different updating mechanisms and still induce the same behaviour.

We now prove that equivalence classes of such agents are in bijection with one-input process functions. To do so, let us first establish a basic lemma regarding 1-input process functions, that they exhibit a comb-decomposition generalising that of quantum and stochastic higher order processes [7, 15, 36].

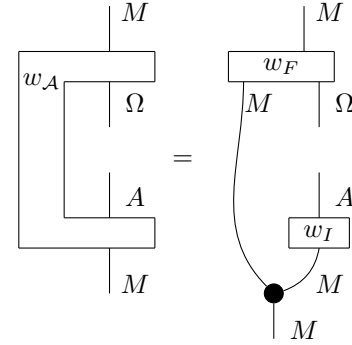
**Lemma 1** (Decomposition of process functions). *Let  $w : P \times O \rightarrow F \times I$  be a one-input process function. Then there exist functions  $w_F : P \times O \rightarrow F$  and  $w_I : P \rightarrow I$  such that*

$$w(p, o) = (w_F(p, o), w_I(p)),$$

for all  $p \in P$  and  $o \in O$ . In particular,  $w_I(p)$  is independent of  $o$ .

*Proof.* Write  $w(p, o) = (w_F(p, o), w_I(p, o))$ . Suppose for contradiction that there exist  $o, o' \in O$  with  $w_I(p, o) \neq w_I(p, o')$  for some fixed  $p \in P$ . Define a function  $f : I \rightarrow O$  by setting  $f(w_I(p, o)) = o$  and  $f(w_I(p, o')) = o'$ . The fixed-point equation  $\bar{o} = f(w_I(p, \bar{o}))$  then has two solutions  $\bar{o} = o$  and  $\bar{o} = o'$ , contradicting the unique fixed-point condition. Therefore  $w_I(p, o)$  must be constant as a function of  $o$ , and we may write  $w(p, o) = (w_F(p, o), w_I(p))$ .  $\square$

In diagrammatic representation, this decomposition reads



where we have used the black dot to represent the copy map in the category of sets and functions. This particular form of comb (which is exhaustive in any cartesian monoidal category), is referred to in the categorical literature as a lens [37, 38]. From this lemma we can now connect process functions to decision-making agents.

**Theorem 1.** *There is a one-to-one correspondence between equivalence classes of deterministic agents and one-input process functions of type  $(A \rightarrow \Omega) \rightarrow (M \rightarrow M)$ .*

*Proof. (Agent  $\implies$  process function).* Fix a deterministic agent  $\mathcal{A} = (\pi, \mathcal{U})$  with memory space  $M$ , action set  $A$  and observation set  $\Omega$ . Define a map

$$\begin{aligned} w_{\mathcal{A}} : M \times \Omega &\longrightarrow M \times A, \\ w_{\mathcal{A}}(m, o) &:= (\mathcal{U}(m, \pi(m), o), \pi(m)). \end{aligned}$$

We claim that  $w_{\mathcal{A}}$  is a one-input process function. Indeed, for any function  $f : A \rightarrow \Omega$  and any  $m \in M$  the fixed-point equation  $o = f(w_{\mathcal{A}}^I(m, o))$  becomes  $o = f(\pi(m))$ , which has a unique solution since the right-hand side is independent of  $o$ .

Now suppose  $\mathcal{A} \cong \mathcal{B}$ . We show that  $w_{\mathcal{A}} = w_{\mathcal{B}}$ . Fix  $m \in M$  and  $o \in \Omega$ . Consider the POMDP  $\mathcal{P}$  with state space  $S = \Omega \times A$ , transition function  $T((o, a'), a) = (o', a)$  (for some fixed state  $o'$ ), observation function  $O((o, a'), a) = o$  and reward function trivial  $\mathcal{R}((o, a), a) = 1$ . For the single step of interaction with starting state  $(m, (o, a'))$ , since  $\mathcal{A} \cong \mathcal{B}$  we have

$$\mathcal{A} \bullet \mathcal{P}(m, (o, a')) = \mathcal{B} \bullet \mathcal{P}(m, (o, a')),$$

and therefore  $\mathcal{U}_{\mathcal{A}}(m, \pi_{\mathcal{A}}(m), o) = \mathcal{U}_{\mathcal{B}}(m, \pi_{\mathcal{B}}(m), o)$  and  $\pi_{\mathcal{A}}(m) = \pi_{\mathcal{B}}(m)$ . It follows that  $w_{\mathcal{A}}(m, o) = w_{\mathcal{B}}(m, o)$  for all  $m, o$ , and hence equivalent agents induce the same process function.

(*Process function*  $\implies$  *agent*). Conversely, let  $w : M \times \Omega \rightarrow M \times A$  be a one-input process function. By the decomposition lemma there are functions  $w_F : M \times \Omega \rightarrow M$  and  $w_I : M \rightarrow A$  such that  $w(m, o) = (w_F(m, o), w_I(m))$ . We define a deterministic agent  $\mathcal{A}_w$  with policy  $\pi_w := w_I$  and memory update  $\mathcal{U}_w(m, a, o) := w_F(m, o)$ .

We now show that composing these two constructions yields the identity up to equivalence. Starting with a process function  $w$ , constructing the agent  $\mathcal{A}_w = (\pi_w, \mathcal{U}_w)$  and then forming the associated process function  $w_{\mathcal{A}_w}$  gives

$$\begin{aligned} w_{\mathcal{A}_w}(m, o) &= (\mathcal{U}_w(m, \pi_w(m), o), \pi_w(m)) \\ &= (w_F(m, o), w_I(m)) = w(m, o), \end{aligned}$$

so  $w_{\mathcal{A}_w} = w$ . Conversely, start with an agent  $\mathcal{A} = (\pi, \mathcal{U})$ , form  $w_{\mathcal{A}}$  and then construct the agent  $\mathcal{A}_{w_{\mathcal{A}}} = (\pi_{w_{\mathcal{A}}}, \mathcal{U}_{w_{\mathcal{A}}})$ . One easily checks that  $\pi_{w_{\mathcal{A}}} = \pi$  and then we have that

$$\begin{aligned} &\mathcal{U}_{w_{\mathcal{A}}}(m, \pi_{w_{\mathcal{A}}}(m), O(\pi_{w_{\mathcal{A}}}(m), s)) \\ &= w_{\mathcal{A}}(m, O(\pi(m), s)) \\ &= \mathcal{U}_{\mathcal{A}}(m, \pi(m), O(\pi(m), s)), \end{aligned}$$

and similarly for rewards, so that  $\mathcal{A}_{w_{\mathcal{A}}} \cong \mathcal{A}$ . This shows that the constructions are mutually inverse up-to equivalence and so establishes a bijection between equivalence classes of agents and one-input process functions.  $\square$

### Observation-independent dec-POMDPs support indefinite order strategies

We now generalise multi-agent systems in artificial intelligence, allowing for the possibility that they might implement strategies in an indefinite causal order.

**Proposition 1** (Multi-agent link product under observation independence). *Let  $\mathcal{P}$  be a deterministic factored  $n$ -party dec-POMDP as in Definition 1. Assume  $\mathcal{P}$  is observation independent in the sense of Definition 2. Let*

*$w$  be an  $n$ -input process function as in Definition 4. Then there is a well-defined deterministic map*

$$w \star \mathcal{P} : M \times S \longrightarrow M \times S \times \mathbb{R}$$

*computing one step of interaction between  $w$  and the environment  $\mathcal{P}$ .*

*Proof.* To define  $w \star \mathcal{P}$  we must show how to apply it to an arbitrary  $(m, s) \in M \times S$ . By Definition 2, there exist functions  $O_i : S \times A_i \rightarrow \Omega_i$  such that (1) holds. For the fixed state  $s$ , we then define

$$f_i^s : A_i \rightarrow \Omega_i, \quad f_i^s(a_i) := O_i(s, a_i). \quad (3)$$

Now, consider the fixed point equation

$$\vec{o}_i = f_i^s(w_{A_i}(m, \vec{o})) \quad (i = 1, \dots, n). \quad (4)$$

Since  $w$  is an  $n$ -input process function, Definition 4 applies to the choice of functions  $(f_1^s, \dots, f_n^s)$  and the fixed point equation has a unique solution

$$\vec{o}^*(m, s) = (o_1^*, \dots, o_n^*).$$

We now evaluate  $w$  on this unique fixed point:

$$(m', a_1^*, \dots, a_n^*) := w(m, \vec{o}^*(m, s)). \quad (5)$$

Let  $\vec{a}^* := (a_1^*, \dots, a_n^*) \in A$ . Using the decomposition of  $\mathcal{P}$  from Definition 1, define

$$(s', r) := (T^{\mathcal{P}}(s, \vec{a}^*), R^{\mathcal{P}}(s, \vec{a}^*)). \quad (6)$$

Finally, we can set

$$(w \star \mathcal{P})(m, s) := (m', s', r). \quad (7)$$

This completes the construction.  $\square$

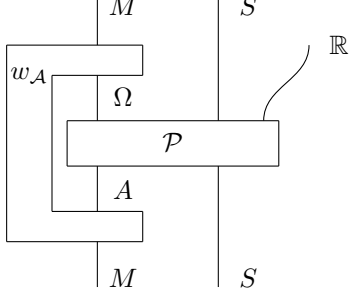
### Interpretation in Terms of Higher-Order Maps

Quantum supermaps or equivalently process matrices are higher-order transformations that send quantum channels to quantum channels, without assuming a definite causal order [36, 39]. In the deterministic classical limit the same notion of higher-order process is instead captured by process functions [7]. Consequently, we may think of process functions as transformations from functions to functions, and this viewpoint gives a clean way to understand the action of any process function on a POMDP.

Indeed, encoding a deterministic POMDP as a single function is straightforward. Define

$$\begin{aligned} \mathcal{P} : A \times S &\rightarrow \Omega \times S \times \mathbb{R} \\ \mathcal{P}(a, s) &:= (O(s, a), T(s, a), R(s, a)). \end{aligned}$$

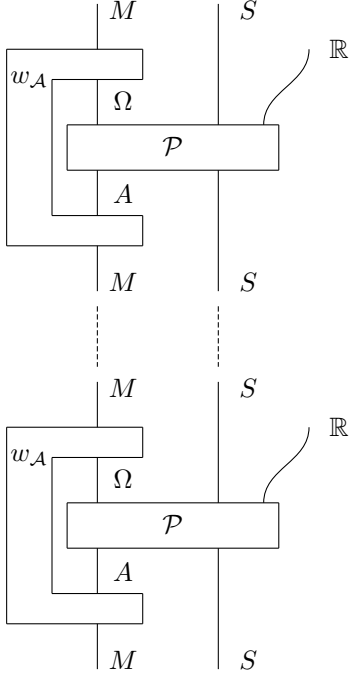
If  $w^A$  is the process function associated to an agent  $A$  then as seen in the previous section the link product  $w^A \star \mathcal{P} : M \times S \rightarrow M \times S \times \mathbb{R}$  computes exactly one step of interaction between the agent and the environment. In the usual graphical notation used to represent higher-order maps in the foundations of physics, this application of a process function to a POMDP via a link product can be expressed by



Furthermore, multi-step interactions correspond to iterated link products. For a finite-horizon task of  $h$  steps one composes the one-step functions  $w^A \star \mathcal{P}$  in sequence:

$$(w^A \star \mathcal{P})^h := (w^A \star \mathcal{P}) \circ \dots \circ (w^A \star \mathcal{P}) : M \times S \longrightarrow M \times S \times \mathbb{R}^h,$$

which yields both the final memory and state as well as the sequence of rewards. Again, graphically, this  $n$ -step procedure reads simply as



In this way the process-function formalism points towards a compact graphical framing for the composition of agents, environments, and their cumulative rewards close in spirit to the graphical representations for protocols in categorical cybernetics in terms of coend-optics [23, 25].

For a deterministic observation-independent **dec**-POMDP  $\mathcal{P}$  and  $n$ -party process function strategy  $w$  with

memory space  $M$ , the result of a  $h$ -step interaction is component-wise

$$(\bigcirc_{t=1}^h (w \star \mathcal{P}))(m_0, s_0) = (m_h, s_h, r_1, \dots, r_h),$$

where  $r_t$  denotes the reward obtained at time-step  $t$ . As a result of this identification, we can define the discounted reward for an arbitrary process-function strategy.

**Definition 6** (Discounted Reward). Let  $\mathcal{P}$  be an observation-independent  $n$ -party **dec**-POMDP and  $w$  be an  $n$ -party process function strategy with memory  $M$ . Given a discount factor  $\gamma \in [0, 1)$  the discounted reward for  $w$  on  $\mathcal{P}$  with discount  $\gamma$  is the map

$$D^w : M \times S \rightarrow \mathbb{R}$$

defined by

$$\sum_{t=1}^{\infty} \gamma^{t-1} r_t.$$

**Definition 7** (Performance of a Process Function Policy). Given a distribution  $\mu$  on  $S$  and memory state  $m_0 \in M$ , define

$$E^w(m_0, \mu) := \sum_{s_0 \in S} \mu(s_0) D^w(m_0, s_0).$$

By the representation for 1-input process functions in terms of a standard agent-state policies, the performance of a 1-input process function policy reduces to the performance of its associated agent-state policy as defined in [33]. Nonetheless, the definition can be applied to arbitrary process functions, leaving open the possibility that there might exist observation independent decentralized partially observable Markov decision processes, in which there exist process functions which outperform in the above sense; any policy which operates using a predefined causal order.

## CONCLUSION

The correspondence established in this article shows that deterministic agents in artificial intelligence and one-input process functions in the foundations of physics are mathematically equivalent. Our main result proves a bijection between equivalence classes of finite-memory POMDP agents and one-input process functions: combining an agent's policy and memory update yields a 1-input process function, whilst every such process function decomposes uniquely back into a policy and update. Two agents are behaviorally indistinguishable precisely when they induce the same process function. This correspondence comes with a duality of interpretations: in physics  $w$  is viewed as the spacetime environment into which local operations are inserted, whereas in artificial intelligence

$w$  encodes the agent’s decision-making procedure and the inserted map corresponds to the environment. Furthermore, we discovered that multi-input process functions yield legitimate assignments to observation-independent decision problems.

Regarding future work, with the connection made, precise methods developed for higher-order causality could now be transported to decision-making problems. First, is the question of whether there exist practical already known observation independent decentralised-POMDPs [2], in which general process function strategies outperform traditional definite ordered ones, or if it might be possible to build toy examples with indefinite order advantage from well-known causal games such as *guess your neighbors input* [7]. Furthermore, it is unclear at this stage how efficiently indefinite causal order strategies can be constructed, via a suitable generalization from policy iteration to process-function iteration.

The identification also motivates a particular fully quantum generalization of POMDPs. More precisely, the upgrading of POMDPs from functions to quantum channels of type  $\mathcal{P} : A \times S \rightarrow \Omega \times S \times R$  with each  $A, S, \Omega, R$  a Hilbert space (with the greatest complication coming from defining a suitable Hilbert space  $R$  for coherently collecting rewards). A quantum decision-making agent in this perspective then corresponds to a (possibly multi-input) quantum super-channel (process matrix). Exactly how this viewpoint compares with the quantum partially observable Markov decision processes of [11], the quantum Markov decision processes of [12], as well as quantum games [28], quantum agents for algorithmic discovery [40], and potential advantages of quantum resources in multi-agent protocols [41], is left for future work.

---

\* matthew.wilson@centralesupelec.fr

- [1] X. Lu, B. V. Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen, Reinforcement learning, bit by bit (2023), arXiv:2103.04047 [cs.LG].
- [2] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, 1st ed., SpringerBriefs in Intelligent Systems (Springer, Cham, 2016) p. 134.
- [3] P. Taranto, S. Milz, M. Murao, M. T. Quintino, and K. Modi, Higher-Order Quantum Operations (2025), arXiv:2503.09693.
- [4] J. Barrett, R. Lorenz, and O. Oreshkov, Quantum causal models (2020), arXiv:1906.10726 [quant-ph].
- [5] O. Oreshkov, F. Costa, and Č. Brukner, Quantum correlations with no causal order, *Nature Communications* **3**, 10.1038/ncomms2076 (2012).
- [6] G. Chiribella, G. M. D’Ariano, P. Perinotti, and B. Valiron, Quantum computations without definite causal structure, *Phys. Rev. A* **88**, 022318 (2013).
- [7] Å. Baumeler and S. Wolf, The space of logically consistent classical processes without causal order, *New Journal of Physics* **18**, 013036 (2016).
- [8] D. Ebler, S. Salek, and G. Chiribella, Enhanced communication with the assistance of indefinite causal order, *Physical Review Letters* **120**, 10.1103/physrevlett.120.120502 (2018).
- [9] P. A. Guérin, A. Feix, M. Araújo, and Č. Brukner, Exponential communication complexity advantage from quantum superposition of the direction of communication, *Physical Review Letters* **117**, 10.1103/physrevlett.117.100502 (2016).
- [10] G. Acampora, A. Ambainis, N. Ares, L. Banchi, P. Bhardwaj, D. Binosi, G. A. D. Briggs, T. Calarco, V. Dunjko, J. Eisert, O. Ezratty, P. Erker, F. Fedele, E. Gil-Fuster, M. Gärttner, M. Granath, M. Heyl, I. Kerenidis, M. Klusch, A. F. Kockum, R. Kueng, M. Krenn, J. Lässl, A. Macaluso, S. Maniscalco, F. Marquardt, K. Michielsen, G. Muñoz-Gil, D. Müssig, H. P. Nautrup, S. A. Neubauer, E. van Nieuwenburg, R. Orus, J. Schmiedmayer, M. Schmitt, P. Slusallek, F. Vicentini, C. Weitenberg, and F. K. Wilhelm, Quantum computing and artificial intelligence: status and perspectives (2025), arXiv:2505.23860 [quant-ph].
- [11] J. Barry, D. T. Barry, and S. Aaronson, Quantum partially observable markov decision processes, *Phys. Rev. A* **90**, 032311 (2014).
- [12] N. Saldi, S. Sanjari, and S. Yuksel, Quantum markov decision processes: General theory, approximations, and classes of policies (2024), arXiv:2402.14649 [quant-ph].
- [13] A. Bisio and P. Perinotti, Theoretical framework for higher-order quantum theory, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **475**, 20180706 (2019).
- [14] L. Apadula, A. Bisio, and P. Perinotti, No-signalling constrains quantum computation with indefinite causal structure, *Quantum* **8**, 1241 (2024).
- [15] A. Kissinger and S. Uijlen, A categorical semantics for causal structure, *Logical Methods in Computer Science* **15**, 10.23638/LMCS-15(3:15)2019 (2019).
- [16] W. Simmons and A. Kissinger, A complete logic for causal consistency (2024), arXiv:2403.09297 [cs.LO].
- [17] W. Simmons and A. Kissinger, Higher-order causal theories are models of bv-logic (2022), arXiv:2205.11219 [cs.LO].
- [18] T. Hoffreumon and O. Oreshkov, Projective characterization of higher-order quantum transformations (2024), arXiv:2206.06206 [quant-ph].
- [19] A. Jenčová, On the structure of higher order quantum maps (2024), arXiv:2411.09256 [quant-ph].
- [20] J. Hefford and M. Wilson, A Profunctorial Semantics for Quantum Supermaps, in *Proceedings of the 39th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS ’24* (Association for Computing Machinery, New York, NY, USA, 2024).
- [21] J. Hefford and M. Wilson, A bv-category of spacetime interventions (2025), arXiv:2502.19022 [quant-ph].
- [22] M. Wilson, G. Chiribella, and A. Kissinger, Quantum supermaps are characterized by locality (2025), arXiv:2205.09844 [quant-ph].
- [23] M. Capucci, B. Gavranović, J. Hedges, and E. F. Rischel, Towards foundations of categorical cybernetics, *Electronic Proceedings in Theoretical Computer Science* **372**, 235 (2022).
- [24] J. Hedges and R. Rodríguez Sakamoto, Value iteration is optic composition, *Electronic Proceedings in Theoretical*

- Computer Science **380**, 417 (2023).
- [25] J. Hedges and R. Rodríguez Sakamoto, Reinforcement learning in categorical cybernetics, *Electronic Proceedings in Theoretical Computer Science* **429**, 270 (2025).
  - [26] N. Ghani, J. Hedges, V. Winschel, and P. Zahn, Compositional game theory (2018), arXiv:1603.04641 [cs.GT].
  - [27] S. Y.-C. Chen, An introduction to quantum reinforcement learning (qrl) (2024), arXiv:2409.05846 [quant-ph].
  - [28] G. Gutoski and J. Watrous, Toward a general theory of quantum games, in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, STOC07 (ACM, 2007) pp. 565–574.
  - [29] S. Dong, B. V. Roy, and Z. Zhou, Simple agent, complex environment: Efficient reinforcement learning with agent states (2021), arXiv:2102.05261 [cs.LG].
  - [30] Ä. Baumeler and E. Tselentis, Equivalence of grandfather and information antinomy under intervention, *Electronic Proceedings in Theoretical Computer Science* **340**, 1 (2021).
  - [31] A. A. Abbott, M. Mhalla, and P. Pocreau, Quantum query complexity of boolean functions under indefinite causal order, *Phys. Rev. Res.* **6**, L032020 (2024).
  - [32] B. Bonet, Deterministic pomdps revisited (2012), arXiv:1205.2659 [cs.AI].
  - [33] A. Sinha and A. Mahajan, Agent-state based policies in pomdps: Beyond belief-state mdps (2024), arXiv:2409.15703 [eess.SY].
  - [34] A. Sinha, M. Geist, and A. Mahajan, Periodic agent-state based q-learning for pomdps (2024), arXiv:2407.06121 [cs.LG].
  - [35] M. Allen and S. Zilberstein, Complexity of decentralized control: Special cases, in *Advances in Neural Information Processing Systems*, Vol. 22, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Curran Associates, Inc., 2009).
  - [36] G. Chiribella, G. M. D’Ariano, and P. Perinotti, Transforming quantum operations: Quantum supermaps, *EPL (Europhysics Letters)* **83**, 30004 (2008).
  - [37] M. Riley, Categories of optics (2018), arXiv:1809.00738 [math.CT].
  - [38] J. N. Foster, M. B. Greenwald, J. T. Moore, B. C. Pierce, and A. Schmitt, Combinators for bidirectional tree transformations: A linguistic approach to the view-update problem, *ACM Trans. Program. Lang. Syst.* **29**, 17 (2007).
  - [39] G. Chiribella, G. M. D’Ariano, and P. Perinotti, Theoretical framework for quantum networks, *Phys. Rev. A* **80**, 022339 (2009).
  - [40] I. Kerenidis and E.-A. Cherratt, Quantum agents for algorithmic discovery (2025), arXiv:2510.08159 [quant-ph].
  - [41] V. Anantharam, Quantum advantage in decentralized control of pomdps: A control-theoretic view of the mermin-peres square (2025), arXiv:2501.16690 [math.OC].