Agent policies from higher-order causal functions

Matt Wilson
Université Paris-Saclay, CNRS, ENS Paris-Saclay, Inria, CentraleSupélec, Laboratoire Méthodes Formelles
matthew.wilson@centralesupelec.fr

Based on the preprint of the same title availliable at *arxiv.org/abs/*2512.10937.

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 for planning and learning in single- and multi-agent systems [26, 28]. In the informational foundations of physics, "agents in laboratories" are modelled as local operations inserted into a spacetime environment, formalised by higher-order quantum [32, 6, 29, 13] and classical [9] processes. Whilst the agent–environment interaction is central to both fields, their formalisations of the concept have developed independently, and no direct mathematical correspondence between them has been established.

A bridge between these two models of the agent–environment interaction has the potential to bring new ideas to both fields. First, regarding causality and spacetime, higher-order processes allow causal (even indefinite-causal) structure to be treated as a resource for communication and computation [9, 14, 16]. A mapping to planning and learning agents opens the door to considering multi-agent tasks in which optimal causal and indefinite-causal strategies might exist and might further be discovered and learned. Second, from a formal and logical perspective, the compositional and logical tools developed for higher-order quantum maps [10, 3, 25, 31, 30, 23, 24] and their generalisation to arbitrary monoidal categories [21, 22, 33] might be utilised to provide tools for reasoning about composite multi-agent systems. Conversely, existing compositional and logical tools for modelling aspects of reward-seeking agents and environments in categorical cybernetics [11, 19, 20] and open game theory [15] could be lifted to the quantum domain, bringing new techniques and ways of thinking about quantum reinforcement learning [12] and quantum game theory [18]. Finally, and most broadly, such an identification might open up new perspectives on the quantisation of agency and artificial intelligence, based on the theory of higher-order quantum operations.

The agent–environment interaction in artificial intelligence: Given S (a set of states), A (a set of available actions), and Ω (a set of observations), a deterministic POMDP is a tuple of functions $\langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle$, where $\mathcal{T} : A \times S \to S$, $\mathcal{O} : A \times S \to \Omega$, and $\mathcal{R} : A \times S \to \mathbb{R}_{\geq 0}$ specify the state update, the observation, and the reward associated with performing an action in a given state.

In artificial intelligence, the POMDP represents a fundamental model of an environment with which an agent (which seeks rewards) can interact. To obtain the corresponding environment model for multi-agent systems, one simply factorises the action and observation spaces. Accordingly, a *deterministic n-party* `dec`*-POMDP* is a POMDP with state set S, action set $\prod_{i=1}^{n} A_i$, and observation set $\prod_{i=1}^{n} \Omega_i$. Observation independence is the deterministic expression of the no-signalling constraint $A_i \not\to \Omega_k$ for $i \neq k$ within one environment step. In general, this kind of signalling constraint occurs quite naturally within the `dec`-POMDP framework; that is, it is common to study environments in which the action of each agent does not influence, at that time step, the observations of the other agents [27, 2].

A deterministic agent is specified by declaring a deterministic agent-state policy: a pair of functions:

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

A deterministic agent is decentralised if the joint policy and update can be factorised into individual policies and updates assigned to each agent, $\pi = \prod_i \pi_i$ and $\mathcal{U} = \prod_i \mathcal{U}_i$.

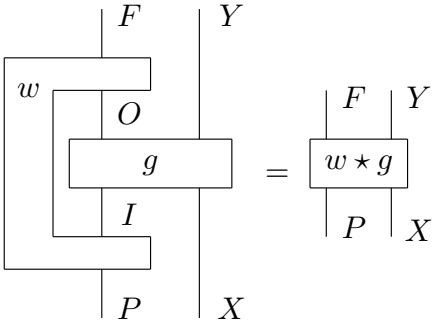
A one-step interaction between an agent and a POMDP can be constructed, mapping the agent and environment states (m_t, s_t) at time t to new states (in addition to collecting a reward) $(m_{t+1}, s_{t+1}, r_{t+1})$ at time $t+1$. The finite-horizon performance of an agent-state policy, given an initial state distribution μ, is then

$$J_{\mu}^{(\pi,\mathcal{U})} = \sum_{s_0} \mu(s_0) \sum_{t=1}^{T} \gamma^{t-1} r_t(m_0, s_0).$$

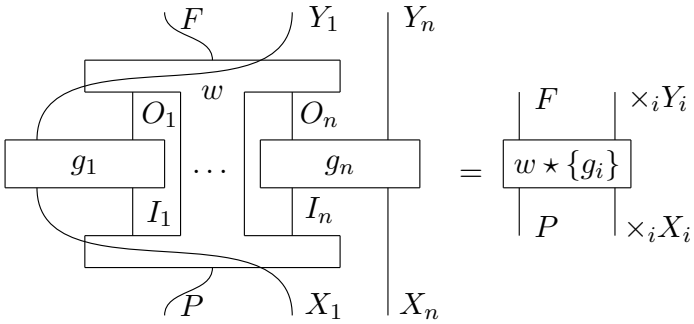
Two agents might not use exactly the same update and policy mechanics and yet be behaviourally equivalent, in the sense that for all possible POMDP environments, the results of their interactions are identical. We write $(\pi, \mathcal{U}) \bullet \langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle : M \times S \to M \times S \times \mathbb{R}_{\geq 0}$ for the function mapping (m_t, s_t) to $(m_{t+1}, s_{t+1}, r_{t+1})$, and we organise agents into equivalence classes; two agent-state policies are equivalent if, for every POMDP,

$$(\pi_1, \mathcal{U}_1) \bullet \langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle = (\pi_2, \mathcal{U}_2) \bullet \langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle.$$

Process function representation for multi-agent systems: In this paper, we show that equivalence classes of agent-state policies correspond to higher-order classical deterministic functions, referred to in the literature as process functions [8, 9, 7, 1]. Defined through unique fixed-point criteria, process functions are the kinds of functions that can be contracted over standard functions. Their generalisation to multi-input process functions allows for the study of indefinite causal structure in deterministic classical theories.



(a) A process function w can be evaluated on an arbitrary function g, using the unique fixed-point criterion.



(b) A multi-input process function can be evaluated on an arbitrary family of functions using the unique fixed-point criterion.

Formally, there is a one-to-one correspondence between equivalence classes of deterministic agent-state policies and one-input process functions of type $(A \to \Omega) \to (M \to M)$, such that for each equivalence class $[\pi, \mathcal{U}]$ the associated process function $w_{[\pi,\mathcal{U}]}$ satisfies

$$[\pi, \mathcal{U}] \bullet \langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle = w_{[\pi,\mathcal{U}]} \star \mathcal{P}_{\langle \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle}.$$

This theorem extends to a decentralised setting: one can define what it means abstractly for any process function (single- or multi-input) to be decentralised, and show that decentralised one-input process functions correspond to decentralised agent-state policies.

A linear logic for multi-agent systems and process functions: We then show that process functions can be defined as passing through arbitrary types $\mathscr{P} \subseteq \mathbf{Set}(A, A')$. We show that process functions form a category **PF**. More precisely, given process functions $w : \mathscr{R} \to \mathscr{Q}$ and $u : \mathscr{Q} \to \mathscr{P}$, we can define their contraction-composition law $u \star w : \mathscr{R} \to \mathscr{P}$, which is associative $(u \star w) \star v = u \star (w \star v)$ as well as unital. We further show that process functions form a model of linear logic: more precisely, a $*$-autonomous category [5]. Every type $\mathscr{P}$ for process functions has an associated dual type $\mathscr{P}^*$, and there exists a tensor product given by $\mathscr{P} \otimes \mathscr{Q} = \mathscr{P} \times \mathscr{Q}$, with De Morgan dual denoted $\mathscr{P} \parr \mathscr{Q} = (\mathscr{P}^* \times \mathscr{Q}^*)^*$ and internal hom, denoted $\mathscr{P} \Rightarrow \mathscr{Q} \cong \mathscr{P}^* \parr \mathscr{Q}$. In this language of $*$-autonomous categories, the following identifications can be made for POMDPs and process functions within **PF**:

- $(A \Rightarrow \Omega) \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0})) \cong$ POMDPs,
- $(\parr_i (A_i \Rightarrow \Omega_i)) \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0})) \cong$ `dec`-POMDPs,
- $(\otimes_i (A_i \Rightarrow \Omega_i)) \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0})) \cong$ observation-independent `dec`-POMDPs,
- $(A \Rightarrow \Omega) \Rightarrow (M \Rightarrow M') \cong$ process functions,
- $\otimes_i ((A_i \Rightarrow A_i') \Rightarrow (M_i \Rightarrow M_i')) \cong$ decentralised process functions,
- $(\otimes_i (A_i \Rightarrow \Omega_i')) \Rightarrow (M \Rightarrow M') \cong$ multi-input process functions.

The logic of $*$-autonomous categories further allows the evaluation of a process function of the form $w : \mathscr{Q} \to (M \Rightarrow M)$ on a POMDP of the form $\mathcal{P} \in \mathscr{Q} \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0}))$ by cutting along $\mathscr{Q}$, that is,

$$(w \star \mathcal{P}) := (w \parr 1_{S \Rightarrow (S \times \mathbb{R}_{\geq 0})}) \circ \mathcal{P}.$$

The result of this cut has type $(M \Rightarrow M) \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0})) \cong (M \times S) \Rightarrow (M \times S \times \mathbb{R}_{\geq 0})$, meaning that the resulting reward of a t-step interaction can then be expressed in terms of $(w \star \mathcal{P})$ (by repeated composition along $M \times S$), allowing for the definition of finite-horizon reward for general process-function policies.

More precisely, for each $\mathcal{P} \in \mathcal{Q} \parr (S \Rightarrow (S \times \mathbb{R}_{\geq 0}))$ and $w : \mathcal{Q} \to (M \Rightarrow M)$, along with discount factor $\gamma \in [0, 1)$, distribution μ on S, time horizon T, and initial memory state m_0, we define the process-function performance

$$J_\mu^w := \sum_{s_0 \in S} \mu(s_0) \sum_{t=1}^{T} \gamma^{t-1} \, r_t(m_0, s_0),$$

where r_t is computed from the last component of $(\bigcirc_{k=1}^{t}((w \star \mathcal{P}) \otimes 1_{\otimes_{i=1}^{k} \mathbb{R}_{\geq 0}})(m_0, s_0)$, i.e., $r_t = (\bigcirc_{k=1}^{t}((w \star \mathcal{P}) \otimes 1_{\otimes_{i=1}^{k} \mathbb{R}_{\geq 0}})^{t+2}(m_0, s_0)$.

Indefinite causal order advantage in multi-agent systems: There is a common sub-expression $(\otimes_i (A_i \Rightarrow \Omega_i))$ which appears both within observation independence and within indefinite causal orders (multi-input process functions). This commonality allows for the construction, via the cut, of a consistent notion of cumulative reward for any indefinite-order strategy over any observation-independent `dec`-POMDP. To see that indefinite causal order can give an advantage over definite-order policies in accumulating finite-horizon reward in multi-agent systems, we adapt known results on the advantages of indefinite causal order within causal games—more precisely, the majority-vote guess-your-neighbour's-input (GYNI) game [9].

In particular, we construct an observation-independent `dec`-POMDP which for the first n rounds returns reward 0, and which plays the same game between rounds $n+1, \ldots, T$. Importantly, a definite-order strategy without decentralisation constraints could win this task perfectly, since it can simply record the inputs to the game in earlier time steps and then use that record to win the GYNI game once it begins.

With decentralisation constraints, however, one can show that for any process function acting on $\mathcal{P}_{\text{GYNI}}^{(n)}$, for each agent i and each round t there exists a function $\varphi_{i,t} : \{0, 1\} \to M_i$ such that $m_i^t = \varphi_{i,t}(x_i)$. In other words, the variable m_i^t is independent of x_j for $j \neq i$, and so no useful information about the game state for parties $j \neq i$ can be recorded in the memory of agent i.

We show that, consequently, for any decentralised multi-input process function with a definite causal order, its return satisfies

$$J \;\leq\; \frac{3}{4} \sum_{t=n+1}^{T} \gamma^{t-1}.$$

By contrast, there exists a between-round decentralised multi-input process function with indefinite causal order (namely the Lugano process [9, 4], which is decentralised since it is memoryless) such that in every reward round $t \geq n + 1$, the realised output y^t satisfies $\mathcal{Q}(x, y^t) = 1$. Consequently,

$$J \;=\; \sum_{t=n+1}^{T} \gamma^{t-1}.$$

Therefore, for any $T \geq n + 1$ and any $\gamma \in (0, 1]$, we bootstrap results on causal games to obtain a strict gap between the process-function performances in multi-agent AI achievable with and without the assumption of a definite background causal structure.

Conclusion and Impact: Many possibilities follow from establishing a formal connection between agency in AI and the foundations of physics. Building on the results of this paper, it may be possible to identify practical, already-known observation-independent decentralised POMDPs [28], and to generalise reinforcement learning to learn indefinite-order policies. There is also the possibility of endowing multi-agent systems in AI with a spatio-temporal logic such as BV-logic [17] (noting that stochastic and quantum higher-order processes form BV-categories [30]), possibly allowing for a purely logical/compositional representation of communication (and constrained communication) within decentralised multi-agent systems [28]. Finally, the results of this paper motivate a particular fully quantum generalisation of POMDPs. More precisely, we consider upgrading POMDPs from functions to quantum channels of type $\mathcal{P} : A \times S \longrightarrow \Omega \times S \times R$, with A, S, Ω, R Hilbert spaces. In this perspective, a quantum agent corresponds to a (possibly multi-input) quantum super-channel (process matrix). Exactly how this viewpoint compares with alternative approaches to the quantisation of POMDPs and agents is left to future work—with the possibility that higher-order quantum operations might provide a natural unifying framework.

References

[1] Alastair A. Abbott, Mehdi Mhalla, and Pierre Pocreau. Quantum query complexity of boolean functions under indefinite causal order. *Phys. Rev. Res.*, 6:L032020, Jul 2024. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.6.L032020, doi:10.1103/PhysRevResearch.6.L032020.

[2] Christopher Amato, Girish Chowdhary, Alborz Geramifard, N. Kemal Üre, and Mykel J. Kochenderfer. Decentralized control of partially observable markov decision processes. In *52nd IEEE Conference on Decision and Control*, pages 2398–2405, 2013. doi:10.1109/CDC.2013.6760239.

[3] Luca Apadula, Alessandro Bisio, and Paolo Perinotti. No-signalling constrains quantum computation with indefinite causal structure. *Quantum*, 8:1241, February 2024. URL: http://dx.doi.org/10.22331/q-2024-02-05-1241, doi:10.22331/q-2024-02-05-1241.

[4] M. Arauújo and A. Feix. Private communication. The process was communicated to Baumeler before it was found by inspecting the extremal points of the non-causal polytope characterized in [9], 2014.

[5] Michael Barr. **-Autonomous Categories*, volume 752 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 1979. doi:10.1007/BFb0064579.

[6] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov. Quantum causal models, 2020. URL: https://arxiv.org/abs/1906.10726, arXiv:1906.10726.

[7] Ämin Baumeler and Eleftherios Tselentis. Equivalence of grandfather and information antinomy under intervention. *Electronic Proceedings in Theoretical Computer Science*, 340:1–12, September 2021. URL: http://dx.doi.org/10.4204/EPTCS.340.1, doi:10.4204/eptcs.340.1.

[8] Ämin Baumeler and Stefan Wolf. Device-independent test of causal order and relations to fixed-points. *New Journal of Physics*, 18(3):035014, April 2016. URL: http://dx.doi.org/10.1088/1367-2630/18/3/035014, doi:10.1088/1367-2630/18/3/035014.

[9] Ämin Baumeler and Stefan Wolf. The space of logically consistent classical processes without causal order. *New Journal of Physics*, 18(1):013036, jan 2016. doi:10.1088/1367-2630/18/1/013036.

[10] Alessandro Bisio and Paolo Perinotti. Theoretical framework for higher-order quantum theory. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2225):20180706, May 2019. URL: http://dx.doi.org/10.1098/rspa.2018.0706, doi:10.1098/rspa.2018.0706.

[11] Matteo Capucci, Bruno Gavranović, Jules Hedges, and Eigil Fjeldgren Rischel. Towards foundations of categorical cybernetics. *Electronic Proceedings in Theoretical Computer Science*, 372:235–248, November 2022. URL: http://dx.doi.org/10.4204/EPTCS.372.17, doi:10.4204/eptcs.372.17.

[12] Samuel Yen-Chi Chen. An introduction to quantum reinforcement learning (qrl), 2024. URL: https://arxiv.org/abs/2409.05846, arXiv:2409.05846.

[13] Giulio Chiribella, Giacomo Mauro D'Ariano, Paolo Perinotti, and Benoît Valiron. Quantum computations without definite causal structure. *Phys. Rev. A*, 88:022318, 2013. doi:10.1103/PhysRevA.88.022318.

[14] Daniel Ebler, Sina Salek, and Giulio Chiribella. Enhanced communication with the assistance of indefinite causal order. *Physical Review Letters*, 120(12), March 2018. URL: http://dx.doi.org/10.1103/PhysRevLett.120.120502, doi:10.1103/physrevlett.120.120502.

[15] Neil Ghani, Jules Hedges, Viktor Winschel, and Philipp Zahn. Compositional game theory, 2018. URL: https://arxiv.org/abs/1603.04641, arXiv:1603.04641.

[16] Philippe Allard Guérin, Adrien Feix, Mateus Araújo, and Časlav Brukner. Exponential communication complexity advantage from quantum superposition of the direction of communication. *Physical Review Letters*, 117(10), September 2016. URL: http://dx.doi.org/10.1103/PhysRevLett.117.100502, doi:10.1103/physrevlett.117.100502.

[17] Alessio Guglielmi. A system of interaction and structure. *ACM Transactions on Computational Logic*, 8(1):1, January 2007. URL: http://dx.doi.org/10.1145/1182613.1182614, doi:10.1145/1182613.1182614.

[18] Gus Gutoski and John Watrous. Toward a general theory of quantum games. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, STOC07, pages 565–574. ACM, June 2007. URL: http://dx.doi.org/10.1145/1250790.1250873, doi:10.1145/1250790.1250873.

[19] Jules Hedges and Riu Rodríguez Sakamoto. Value iteration is optic composition. *Electronic Proceedings in Theoretical Computer Science*, 380:417–432, August 2023. URL: http://dx.doi.org/10.4204/EPTCS.380.24, doi:10.4204/eptcs.380.24.

[20] Jules Hedges and Riu Rodríguez Sakamoto. Reinforcement learning in categorical cybernetics. *Electronic Proceedings in Theoretical Computer Science*, 429:270–286, September 2025. URL: http://dx.doi.org/10.4204/EPTCS.429.15, doi:10.4204/eptcs.429.15.

[21] James Hefford and Matt Wilson. A Profunctorial Semantics for Quantum Supermaps. In *Proceedings of the 39th Annual ACM/IEEE Symposium on Logic in Computer Science*, LICS '24, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3661814.3662123.

[22] James Hefford and Matt Wilson. A bv-category of spacetime interventions, 2025. URL: https://arxiv.org/abs/2502.19022, arXiv:2502.19022.

[23] Timothée Hoffreumon and Ognyan Oreshkov. Projective characterization of higher-order quantum transformations, 2024. URL: https://arxiv.org/abs/2206.06206, arXiv:2206.06206.

[24] Anna Jenčová. On the structure of higher order quantum maps, 2024. URL: https://arxiv.org/abs/2411.09256, arXiv:2411.09256.

[25] Aleks Kissinger and Sander Uijlen. A categorical semantics for causal structure. *Logical Methods in Computer Science*, 15, 2019. doi:10.23638/LMCS-15(3:15)2019.

[26] Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit, 2023. URL: https://arxiv.org/abs/2103.04047, arXiv:2103.04047.

[27] Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Networked distributed pomdps: a synthesis of distributed constraint optimization and pomdps. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 1*, AAAI'05, pages 133–139. AAAI Press, 2005.

[28] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer, Cham, 1 edition, 2016. doi:10.1007/978-3-319-28929-8.

[29] Ognyan Oreshkov, Fabio Costa, and Časlav Brukner. Quantum correlations with no causal order. *Nature Communications*, 3(1092), 2012. doi:10.1038/ncomms2076.

[30] Will Simmons and Aleks Kissinger. Higher-order causal theories are models of bv-logic, 2022. URL: https://arxiv.org/abs/2205.11219, arXiv:2205.11219.

[31] Will Simmons and Aleks Kissinger. A complete logic for causal consistency, 2024. URL: https://arxiv.org/abs/2403.09297, arXiv:2403.09297.

[32] Philip Taranto, Simon Milz, Mio Murao, Marco Túlio Quintino, and Kavan Modi. Higher-Order Quantum Operations, 2025. URL: https://arxiv.org/abs/2503.09693, arXiv:2503.09693.

[33] Matt Wilson, Giulio Chiribella, and Aleks Kissinger. Quantum supermaps are characterized by locality, 2025. URL: https://arxiv.org/abs/2205.09844, arXiv:2205.09844.