View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Hertfordshire Research Archive

Causal blankets: Theory and algorithmic framework

Fernando E. Rosas[1,2,3], Pedro A.M. Mediano[4], Martin Biehl[5], Shamil Chandaria[1,6], and Daniel Polani[7] ⋆

[1] Centre for Psychedelic Research, Imperial College London, London SW7 2DD, UK
[2] Data Science Institute, Imperial College London, London SW7 2AZ, UK
[3] Centre for Complexity Science, Imperial College London, London SW7 2AZ, UK
[4] Department of Psychology, University of Cambridge, Cambridge CB2 3EB, UK
[5] Araya Inc., Tokyo 107-6024, Japan
[6] Institute of Philosophy, School of Advanced Study, University of London, UK
[7] Dept. of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK
f.rosas@imperial.ac.uk pam83@cam.ac.uk martin@araya.org
shamil.chandaria@gmail.com d.polani@herts.ac.uk

Abstract. We introduce a novel framework to identify perception-action loops (PALOs) directly from data based on the principles of computational mechanics. Our approach is based on the notion of *causal blanket*, which captures sensory and active variables as dynamical sufficient statistics — i.e. as the "differences that make a difference." Moreover, our theory provides a broadly applicable procedure to construct PALOs that requires neither a steady-state nor Markovian dynamics. Using our theory, we show that every bipartite stochastic process has a causal blanket, but the extent to which this leads to an effective PALO formulation varies depending on the integrated information of the bipartition.

Keywords: Perception-action loops · Computational Mechanics · Integrated Information · Stochastic processes

1 Introduction

The perception-action loop (PALO) is one of the most important constructs of cognitive science, and plays a fundamental role in many other disciplines including reinforcement learning and computational neuroscience. Despite its importance and pervasiveness, fundamental questions about what kind of systems can be properly described by a PALO are still to a large extent unanswered. The aim of this paper is to introduce a novel framework that allows us to identify PALOs directly from data, which complements existent approaches and serves to deepen our understanding of the essential elements that make a PALO.

⋆ F.R. was supported by the Ad Astra Chandaria foundation. P.M. was funded by the Wellcome Trust (grant no. 210920/Z/18/Z). M.B. was supported by a grant from Templeton World Charity Foundation, Inc. (TWCF). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of TWCF.

1.1 Markov blankets

One of the most encompassing accounts of PALOs can be found in the Free Energy Principle (FEP) literature, which formalises them via *Markov blankets* (MBs) [14]. An interesting contribution of this literature is to characterise "sensory" (S) and "active" (A) variables as having two defining properties: (i) they mediate the interactions between internal variables of the agent (M) and external variables of its environment (E), and (ii) they impose a specific causal structure on these interactions — e.g. sensory variables may affect internal variables, but are not (directly) affected by them [14].

Formally, MBs were originally introduced by Pearl [21] for Markov and Bayesian networks. Within the FEP literature, MBs are usually employed in multivariate stochastic processes with ergodic Markovian dynamics, with a steady-state distribution p^* that is required to satisfy [20]

$$p^*(e_t, m_t|s_t, a_t) = p^*(e_t|s_t, a_t)p^*(m_t|s_t, a_t) \ . \tag{1}$$

However, Eq. (1) does not suffice to guarantee a PALO structure, as noted in Ref. [7]. In effect, the MB condition is insufficient to establish requirement (ii): its symmetry with respect to internal and external variables make it impossible to infer the direction of the loop; additionally, the fact that the condition holds across variables synchronously makes it unsuitable to guarantee a causal relationship [22]. Recent reports [11] acknowledge that this synchronous condition needs to be complemented with additional diachronic restrictions on the system's dynamics, which can be written, for instance, as a set of coupled stochastic differential equations of the form

$$\begin{aligned} \dot{m}_t &= f_{\text{in}}(m_t, a_t, s_t) + \omega_t^{\text{in}} \ , \qquad \dot{a}_t = f_{\text{a}}(m_t, a_t, s_t) + \omega_t^{\text{a}} \ , \\ \dot{e}_t &= f_{\text{ex}}(e_t, a_t, s_t) + \omega_t^{\text{ex}} \ , \qquad \dot{s}_t = f_{\text{s}}(e_t, a_t, s_t) + \omega_t^{\text{s}} \ . \end{aligned} \tag{2}$$

Above, the functions $f_{\text{in}}, f_{\text{a}}, f_{\text{ex}}, f_{\text{s}}$ determine the flow, and $\omega_t^{\text{in}}, \omega_t^{\text{a}}, \omega_t^{\text{ex}}, \omega_t^{\text{s}}$ denote additive Gaussian noise. Interestingly, it has been shown that Eq. (2) implies Eq. (1) under additional assumptions: either block diagonality conditions over the solenoidal flow [11], or strong dissipation [12, Appendix].[8] Hence, PALOs could be interpreted as coupled stochastic dynamical systems of the form in Eq. (2), as long as the flow satisfies any of the two mentioned conditions.

Despite its elegance, this formalisation of PALOs has important limitations. First, this formulation relies strongly on Langevin dynamics, making it difficult to extend it to PALOs appearing in discrete systems. Secondly, this approach depends on a set of assumptions — for one, the aforementioned conditions over the flow and the restriction to systems in their steady-state — that might be too restrictive for some scenarios of interest. Finally, and perhaps most importantly, Eq. (1) forces all interactions between M_t and E_t to be accountable by (S_t, A_t), which imposes — due to the data processing inequality [9] — an information

[8] However, in the general case neither Eqs. (1) or (2) imply each other [7] — hence they need to be taken as complementary conditions.

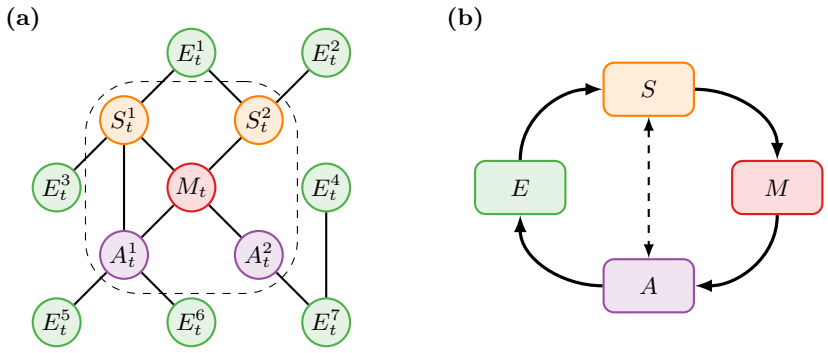

Fig. 1. Two visualisations of PALOs in the FEP literature, either based on **(a)** Markov blankets according to Eq. (1) or **(b)** Langevin dynamics following Eq. (2).

bottleneck of the form $I(M_t; E_t) \leq I(M_t; A_t, S_t)$. Therefore, the MB formalism forbids interdependencies induced by past events that are kept in memory, but may not directly influence the present state of the blankets.[9] This information kept in memory arguably plays an important role in many PALOs, and includes uncontroversial features of cognition (such as old memories that an agent retains but is neither caused by a sensation nor causing an action at the current moment), yet are forbidden by MBs.

1.2 Computational mechanics, causal states, and epsilon-machines

Computational mechanics is a method for studying patterns and statistical regularities observed in stochastic processes by uncovering their hidden causal structure [24,25]. A key insight is that an optimal, minimimal representation of a process can be revealed by grouping past trajectories according to their forecasting abilities into so-called *causal states*. More precisely, the causal states of a (possibly non-Markovian) time series $\{Z_t\}_{t\in\mathbb{Z}}$ are the equivalent classes of trajectories $\overleftarrow{\boldsymbol{z}}_t := (\ldots, z_{t-1}, z_t)$ given by the relationship

$$\overleftarrow{\boldsymbol{z}}_t \equiv_\epsilon \overleftarrow{\boldsymbol{z}}'_t \qquad \text{iff} \qquad p(z_{t+1}|\overleftarrow{\boldsymbol{z}}_t) = p(z_{t+1}|\overleftarrow{\boldsymbol{z}}'_t) \quad \forall z_{t+1} \ .$$

It can be shown that the causal states are the coarsest coarse-graining of past trajectories $\overleftarrow{\boldsymbol{x}}_t$ that retains full predictive power over future variables [10,13]. Moreover, the corresponding process over causal states always has Markovian dynamics, providing the simplest yet encompassing representation of the system's information dynamics on a latent space — known as the *epsilon-machine*.

Please note that the causal states of a system are guaranteed to provide counterfactual relationships [22] only if the system at hand is fully observed. In the case of partially observed scenarios, causal states ought to be understood in the Granger sense, i.e. as states of maximal non-mediated predictive ability [8].

[9] We thank Nathaniel Virgo for first noting this issue.

1.3 Contribution

In this paper we introduce an operationalisation of PALOs based on *causal blankets* (CB), a construction based on a novel definition of dynamical statistical sufficiency. CB capture properties (i) and (ii) in a single mathematical construction by applying informational constructs directly to dynamical conditions. Moreover, CBs can be constructed with great generality for any bipartite system without imposing further conditions, and hence can be applied to non-ergodic, non-Markovian stochastic processes. This generality allows us to explore novel connections between PALOs and integrated information. In the rest of the manuscript, we:

1) Provide a rigorous definition of CBs (Definition 2); and
2) Show every agent-environment partition has a CB, and thus can be described as a PALO (Proposition 1); although
3) Not all systems are equally well described as a PALO, and this can be quantified via information geometry and integrated information (Sec. 3) — providing a principled measure to distinguish preferable candidates for PALO.[10]

2 Causal blankets as informational boundaries

We consider the perspective of a scientist who repeatedly measures a system composed of two interacting parts X_t and Y_t. We assume that, from these observations, a reliable statistical model of the corresponding discrete-time stochastic process can be built — of which all the resulting marginal and conditional distributions are well-defined. Random variables are denoted by capital letters (e.g. X, Y) and their realisations by lower case letters (e.g. x, y); stochastic processes at discrete times (i.e. time series) are represented as bold letters without subscript $\boldsymbol{X} = \{X_t\}_{t\in\mathbb{Z}}$, and $\overleftarrow{\boldsymbol{X}}_t \coloneqq (\ldots, X_{t-1}, X_t)$ denotes the infinite past of $\boldsymbol{X}$ until and including t.

Given two random variables X and Y, a statistic $U = f(X)$ is said to be *Bayesian sufficient for X w.r.t. Y* if $X \perp\!\!\!\perp Y \mid U$, which implies that all the common variability between X and Y is accounted for by U [9]. The first step in our construction is to introduce a dynamical version of statistical sufficiency.

Definition 1 (D-BaSS). *Given two stochastic processes $\boldsymbol{X}, \boldsymbol{Y}$, a process $\boldsymbol{U}$ is a dynamical Bayesian sufficient statistic (D-BaSS) of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$ if, for all $t \in \mathbb{Z}$, the following conditions hold:*

i. Precedence: there exists a function $F(\cdot)$ such that $U_t = F(\overleftarrow{\boldsymbol{X}}_t)$ for all $t \in \mathbb{Z}$.
ii. Sufficiency: $Y_{t+1} \perp\!\!\!\perp \overleftarrow{\boldsymbol{X}}_t \mid (U_t, \overleftarrow{\boldsymbol{Y}}_t)$.

Moreover, a stochastic process $\boldsymbol{M}$ is a minimal D-BaSS of $\boldsymbol{X}$ with respect to $\boldsymbol{Y}$ if it is itself a D-BaSS and for any D-BaSS $\boldsymbol{U}$ there exists a function $f(\cdot)$ such that $f(U_t) = M_t, \forall t \in \mathbb{Z}$.

[10] The proofs of our results can be found in the Appendix.

The first condition above states that $\boldsymbol{U}$ is no more than a simpler, coarse-grained representation of $\boldsymbol{X}$, and the second implies that the influence of $\overleftarrow{\boldsymbol{X}}_t$ on Y_{t+1} given $\overleftarrow{\boldsymbol{Y}}_t$ is fully mediated by U_t. This has interesting consequences for transfer entropy, as seen in the next lemma.

Lemma 1. *If $\boldsymbol{U}$ is a D-BaSS for $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$, then*

$$\mathrm{TE}(\boldsymbol{X} \to \boldsymbol{Y})_t \coloneqq I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) = I(U_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) \ . \tag{3}$$

There are many such D-BaSS; e.g. $U_t = \overleftarrow{\boldsymbol{X}}_t$ would be one valid D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$. However, Theorem 1 shows that minimal D-BaSS's are unique (up to bijective transformations).

Theorem 1 (Existence and uniqueness of the minimal D-BaSS). *Given stochastic processes $\boldsymbol{X}, \boldsymbol{Y}$, the minimal D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$ corresponds to the partition of past-trajectories $\overleftarrow{\boldsymbol{x}}_t$ induced by the following equivalence relationship:*

$$\overleftarrow{\boldsymbol{x}}_t \equiv_p \overleftarrow{\boldsymbol{x}}'_t \qquad \textit{iff} \qquad \forall \overleftarrow{\boldsymbol{y}}_t, y_{t+1} \quad p(y_{t+1} | \overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) = p(y_{t+1} | \overleftarrow{\boldsymbol{x}}'_t, \overleftarrow{\boldsymbol{y}}_t) \ .$$

Therefore, the minimal D-BaSS is always well-defined, and is unique up to an isomorphism.

This result shows that D-BaSSs can be build irrespective of any other possibly latent influences on $\boldsymbol{X}$ and $\boldsymbol{Y}$, as it is defined purely on the joint statistics of these two processes. Moreover, Theorem 1 provides a recipe to build a D-BaSS: group together all the past trajectories that lead to the same predictions, which is a key principle of computational mechanics [10,13,24,25]. Therefore, a minimal D-BaSS distinguishes only "differences that make a difference" for the future dynamics, generalising the construction presented in Ref. [6, Definition 1] for Markovian dynamical systems, and being closely related to the notion of sensory equivalence presented in Ref. [3]. With these ideas at hand, we can formulate our definition of causal blanket.

Definition 2 (Causal blanket). *Given two stochastic processes $\boldsymbol{X}, \boldsymbol{Y}$, a* reciprocal D-BaSS *(ReD-BaSS) is a stochastic process $\boldsymbol{R}$ which satisfies:*

i. Joint precedence: $R_t = F(\overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t)$ for some function $F(\cdot)$.
ii. Reciprocal sufficiency: $\boldsymbol{R}$ is a D-BaSS for $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$, and also is a D-BaSS for $\boldsymbol{Y}$ w.r.t. $\boldsymbol{X}$.

A causal blanket *(CB) is a minimal ReD-BaSS: a time series $\boldsymbol{M}$, itself a ReD-BaSS, such that for all ReD-BaSSs $\boldsymbol{R}$ there exists a function $f(\cdot)$ such that $M_t = f(R_t), \forall t \in \mathbb{Z}$.*

This definition satisfies the two key desiderata discussed in Section 1.1: (i) a CB mediates the interactions that take place between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and (ii) it assesses causality by focusing on statistical relationships between past and future. From this perspective, CBs are the "informational layer" that causally decouples the agent's and environment's temporal evolution from each other (see Proposition 2). Additionally, our next result guarantees that CBs always exist, and are unique to each bipartite system.

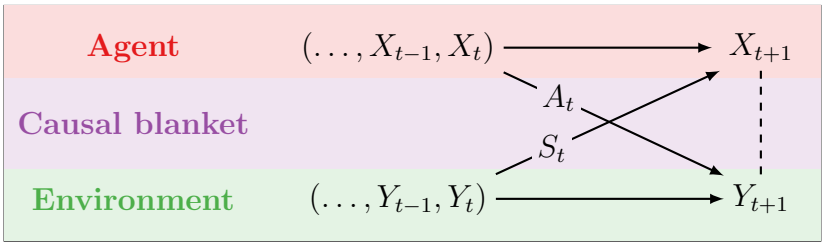

Fig. 2. Causal blanket $\{\boldsymbol{S}, \boldsymbol{A}\}$, which acts as a sufficient statistic mediating the interactions between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Proposition 1. *Given $\boldsymbol{X}, \boldsymbol{Y}$, their CB always exists and is unique (up to an isomorphism). Moreover, their CB is isomorphic to a pair $\{\boldsymbol{S}, \boldsymbol{A}\}$, where $\boldsymbol{A}$ is a minimal D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$, and $\boldsymbol{S}$ is a minimal D-BaSS of $\boldsymbol{Y}$ w.r.t. $\boldsymbol{X}$.*

Proposition 1 has two important consequences: it guarantees that CBs *always* exist, and that they naturally resemble a PALO — as visualised in Fig 2. Please note that this type of PALO formalisation has a rich history, being studied in Refs. [4,5] and variations being considered in Refs. [15,16,26]. In contrast, our framework follows Refs. [3,6] and does not assume active and sensory variables as given, but discovers them directly from the data. As a matter of fact, the "sensory" ($\boldsymbol{S}$) and "active" ($\boldsymbol{A}$) variables of CBs correspond (due to Definition 2) to minimal sufficient statistics that mediate the interdependencies between the past and future of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The construction of CBs imposes no requirements on the system's statistics or its structure — beyond the bipartition, holding also for non-ergodic and also non-stationary systems, and systems with non-Markovian dynamics.

It is also possible to build internal and external states M_t, E_t such that $(M_t, A_t) = X_t$ and $(E_t, S_t) = Y_t$ with great generality. This can be done via an orthogonal completion of the phase space; the details of this procedure will be made explicit in a future publication. In this way, CBs can be thought as suggesting implicit "equations of motion" somehow equivalent to Eq. (2), as shown in Figure 2. However, it is important to remark that this representation does *not* provide counterfactual guarantees for partially observed systems (see Section 1.2).

Example 1. Consider a multivariate stochastic process $\boldsymbol{M}, \boldsymbol{A}, \boldsymbol{E}, \boldsymbol{S}$ whose dynamics follows

$$\begin{aligned} M_{t+1} &= f_{\text{in}}(M_t, A_t, S_t) + N_{\text{in}}, & A_{t+1} &= f_{\text{a}}(M_t, A_t, S_t) + N_{\text{a}}, \\ E_{t+1} &= f_{\text{ex}}(E_t, A_t, S_t) + N_{\text{ex}}, & S_{t+1} &= f_{\text{s}}(E_t, A_t, S_t) + N_{\text{s}}, \end{aligned} \tag{4}$$

with $N_t^{\text{in}}, N_t^{\text{a}}, N_t^{\text{ex}}, N_t^{\text{s}}$ being independent of M_t, A_t, E_t, S_t (note that Eq. 4 corresponds to a discrete-time version of Eq. (2)). Then, by defining $X_t = (M_t, A_t)$ and $Y_t = (E_t, S_t)$, one can show using Definition 2 that that $\{\boldsymbol{S}, \boldsymbol{A}\}$ is the CB of $\boldsymbol{X}, \boldsymbol{Y}$ — as long as the partial derivatives of $f_{\text{in}}, f_{\text{a}}, f_{\text{ex}}, f_{\text{s}}$ with respect to their corresponding arguments are nonzero.

3 Integrated information transcends the blankets

According to Def. 2, CBs don't depend on the joint distribution $p(x_{t+1}, y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t)$, but only on the marginals $p(x_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t)$ and $p(y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t)$. Here we study how meaningful the CB (and the description of the system as a PALO) is when the joint process's dynamics are different from the product of these two marginals.

Let us start by introducing the *synergistic coefficient* $\xi_t \in \mathbb{R}$, which is a random variable given by

$$\xi_t \coloneqq \log \frac{p(X_{t+1}, Y_{t+1}|\overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t)}{p(X_{t+1}|\overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t)\, p(Y_{t+1}|\overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t)} \ . \tag{5}$$

A process $(\boldsymbol{X}, \boldsymbol{Y})$ is said to have *factorisable dynamics* if $\xi_t = 0$ *a.s.* for all $t \in \mathbb{Z}$.

Proposition 2 (Conditional independence of trajectories). *If $\boldsymbol{R}$ is a ReD-BaSS and the dynamics of $\boldsymbol{X}, \boldsymbol{Y}$ is factorisable, then $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{R}$.*

A direct consequence of this Proposition is that a ReD-BaSS does not guarantee statistical independence of $\boldsymbol{X}, \boldsymbol{Y}$ at the trajectory level in non-factorisable systems. Therefore, in such systems there are interactions between $\boldsymbol{X}$ and $\boldsymbol{Y}$ that are not mediated by the CB. Please note that this is not a weakness of the CB construction — which are optimal in capturing all the directed influences, as show by Proposition 1. Instead, this result suggests that non-factorisable systems might not be well-suited to be described as a PALO.

To further understand this, let us explore the integrated information in the system $(\boldsymbol{X}, \boldsymbol{Y})$ using information geometry [19]. For this, consider the manifolds

$$\begin{aligned}
\mathcal{M}_1 &= \left\{q_t : q(x_{t+1}, y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) = q(x_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) q(y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t)\right\} \ , \\
\mathcal{M}_2 &= \left\{q_t : q(x_{t+1}, y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) = q(x_{t+1}|\overleftarrow{\boldsymbol{x}}_t) q(y_{t+1}|\overleftarrow{\boldsymbol{y}}_t)\right\} \ .
\end{aligned}$$

Manifold $\mathcal{M}_1$ corresponds to all systems with factorisable dynamics, and $\mathcal{M}_2$ to all systems where the dynamics of agent and environment are fully decoupled. The information-geometric projection of an arbitrary system p_t onto $\mathcal{M}_2$,

$$\tilde{\varphi}_t \coloneqq \min_{q_t \in \mathcal{M}_2} D(p_t||q_t) \ , \tag{6}$$

has been proposed as a measure of integrated information [2,18]. Using the Pythagoras theorem [1] together with the fact that $\mathcal{M}_2 \subset \mathcal{M}_1$, one can decompose $\tilde{\varphi}_t$ as

$$\underbrace{\tilde{\varphi}_t}_{D(p_t\|q_t^{(2)})} = \underbrace{\mathbb{E}\{\xi_t\}}_{D(p_t||q_t^{(1)})} + \underbrace{\left[\mathrm{TE}(\boldsymbol{A} \to \boldsymbol{Y})_t + \mathrm{TE}(\boldsymbol{S} \to \boldsymbol{X})_t\right]}_{D(q_t^{(1)}||q_t^{(2)})} \ , \tag{7}$$

where $q_t^{(k)} \coloneqq \arg\min_{q_t \in \mathcal{M}_k} D(p_t||q_t)$.[11]

[11] Note that in non-ergodic scenarios the expected values are not calculated over individual trajectories, but over the ensemble statistics that defines the probability.

This decomposition confirms previous results that showed that integrated information is a construct that combines low-order transfer and high-order synergies [17]. Thanks to Lemma 1, Eq. (7) states that the transfer component of $\tilde{\varphi}_t$ (i.e. $D\big(q_t^{(1)}||q_t^{(2)}\big)$) is what is properly mediated by the CB. In contrast, the part of $\tilde{\varphi}$ related to high-order statistics, i.e. $\mathbb{E}\{\xi_t\} = I(X_{t+1}; Y_{t+1}|\tilde{\boldsymbol{X}}_t, \tilde{\boldsymbol{Y}}_t)$, is not accounted by the CB. This last term can either refer to spurious synchronous correlations (due e.g. to sub-sampling), or be due to synergistic dynamics that are a signature of emergent phenomena [23].

In summary, our results suggests that the dynamics of a system $(\boldsymbol{X}, \boldsymbol{Y})$ that is too synergistically integrated are poorly represented as a PALO, even if the CB formally still exists. Additionally, the synergistic component of integrated information can be used as a measure for this mismatch.

4 Conclusion

This manuscript introduced a data-driven method to build PALOs leveraging principles of computational mechanics. Our construction provides an informational interpretation of sensory and actuation variables: sensory (resp. active) variables encode all the changes from "outside" (resp. "inside") that affect the future evolution of the "inside" (resp. "outside"). Our framework is broadly applicable, depending only on the underlying bipartition but not imposing any further conditions on the system's dynamics or distribution. Furthermore, we illustrated how this construction allows one to relate — within a PALO framework — the separation of a system and its environment to the integrated information encompassing the two.

It is to be noted that the CB construction relies on discrete time, which, while being immediately applicable to digitally sampled data, might not be natural in some scenarios. Also, CB theory at this stage does not provide explicit links with probabilistic inference. As shown in Example 1, CBs provide a natural extension of Eq. (2) to the discrete-time case, so one possibility would be to combine them with the MB condition in Eq. (1). The exploration of such "causal Markov blankets" which would satisfy both Eq. (1) and Definition 2 is an interesting avenue for future research.

It is our hope that the CB construction may enrich the toolbox of researchers studying PALOs and help to illuminate further our understanding of the nature of agency.

References

1. Amari, S.i., Nagaoka, H.: Methods of information geometry, vol. 191. American Mathematical Soc. (2007)
2. Ay, N.: Information geometry on complexity and stochastic interaction. Entropy **17**(4), 2432–2458 (2015)
3. Ay, N., Löhr, W.: The Umwelt of an embodied agent — A measure-theoretic definition. Theory in Biosciences **134**(3-4), 105–116 (2015)
4. Bertschinger, N., Olbrich, E., Ay, N., Jost, J.: Information and closure in systems theory. In: Explorations in the Complexity of Possible Life. Proceedings of the 7th German Workshop of Artificial Life. pp. 9–21 (2006)
5. Bertschinger, N., Olbrich, E., Ay, N., Jost, J.: Autonomy: An information theoretic perspective. Biosystems **91**(2), 331–345 (2008)
6. Biehl, M., Polani, D.: Action and perception for spatiotemporal patterns. In: Artificial Life Conference Proceedings 14. pp. 68–75. MIT Press (2017)
7. Biehl, M., Pollock, F.A., Kanai, R.: A technical critique of the free energy principle as presented in "Life as we know it". arXiv:2001.06408 (2020)
8. Bressler, S.L., Seth, A.K.: Wiener–Granger causality: A well established methodology. Neuroimage **58**(2), 323–329 (2011)
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons (2012)
10. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Physical Review Letters **63**(2), 105 (1989)
11. Friston, K., Da Costa, L., Parr, T.: Some interesting observations on the free energy principle. arXiv:2002.04501 (2020)
12. Friston, K.J., Fagerholm, E.D., Zarghami, T.S., Parr, T., Hipólito, I., Magrou, L., Razi, A.: Parcels and particles: Markov blankets in the brain. arXiv:2007.09704 (2020)
13. Grassberger, P.: Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics **25**(9), 907–938 (1986)
14. Kirchhoff, M., Parr, T., Palacios, E., Friston, K., Kiverstein, J.: The Markov blankets of life: Autonomy, active inference and the free energy principle. Journal of The Royal Society Interface **15**(138), 20170792 (2018)
15. Klyubin, A.S., Polani, D., Nehaniv, C.L.: Organization of the information flow in the perception-action loop of evolved agents. In: Proceedings. 2004 NASA/DoD Conference on Evolvable Hardware, 2004. pp. 177–180. IEEE (2004)
16. Klyubin, A.S., Polani, D., Nehaniv, C.L.: Representations of space and time in the maximization of information flow in the perception-action loop. Neural Computation **19**(9), 2387–2432 (2007)
17. Mediano, P.A., Rosas, F., Carhart-Harris, R.L., Seth, A.K., Barrett, A.B.: Beyond integrated information: A taxonomy of information dynamics phenomena. arXiv:1909.02297 (2019)
18. Mediano, P.A., Seth, A.K., Barrett, A.B.: Measuring integrated information: Comparison of candidate measures in theory and simulation. Entropy **21**(1), 17 (2019)
19. Oizumi, M., Tsuchiya, N., Amari, S.i.: Unified framework for information integration based on information geometry. Proceedings of the National Academy of Sciences **113**(51), 14817–14822 (2016)
20. Parr, T., Da Costa, L., Friston, K.: Markov blankets, information geometry and stochastic thermodynamics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **378**(2164), 20190159 (2020)

21. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
22. Pearl, J.: Causality. Cambridge University Press (2009)
23. Rosas, F.E., Mediano, P.A., Jensen, H.J., Seth, A.K., Barrett, A.B., Carhart-Harris, R.L., Bor, D.: Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. arXiv:2004.08220 (2020)
24. Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: Pattern and prediction, structure and simplicity. Journal of Statistical Physics **104**(3-4), 817–879 (2001)
25. Shalizi, C.: Causal architecture. Complexity, and Self-Organization in Time Series and Cellular Automata PhD thesis (Univ Wisconsin–Madison, Madison, WI) (2001)
26. Tishby, N., Polani, D.: Information theory of decisions and actions. In: Perception-action cycle, pp. 601–636. Springer (2011)

A Proofs

Proof (Lemma 1). Let's consider $\boldsymbol{U}$ to be a D-BaSS for $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$. Then, property (ii) of a D-Bass is equivalent to

$$I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} \mid U_t, \overleftarrow{\boldsymbol{Y}}_t) = 0 \ . \tag{8}$$

Using this, one can verify that

$$I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) = I(U_t, \overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) = I(U_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) \ .$$

Here, the first equality holds because U_t is a deterministic function of $\overleftarrow{\boldsymbol{X}}_t$, and the second equality follows from an application of the chain rule and Eq. (8).

Proof (Theorem 1). Consider the function $F(\cdot)$ that maps each $\overleftarrow{\boldsymbol{x}}_t$ to its corresponding equivalence class $F(\overleftarrow{\boldsymbol{x}}_t)$ established by the equivalence relationship $\equiv_p$, and define $M_t = F(\overleftarrow{\boldsymbol{X}}_t)$. As this construction satisfies the requirement of precedence in Def. 1, let us show the sufficiency of $\boldsymbol{M}$. By definition of M_t, it is clear that if $m_t = F(\overleftarrow{\boldsymbol{x}}_t)$ then

$$p(y_{t+1} | \overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) = p(y_{t+1} | m_t, \overleftarrow{\boldsymbol{y}}_t) \ ,$$

which implies that $H(Y_{t+1} | \overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t) = H(Y_{t+1} | M_t, \overleftarrow{\boldsymbol{Y}}_t)$. As a consequence,

$$\begin{aligned} I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) &= H(Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) - H(Y_{t+1} | \overleftarrow{\boldsymbol{X}}_t, \overleftarrow{\boldsymbol{Y}}_t) \\ &= H(Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) - H(Y_{t+1} | M_t, \overleftarrow{\boldsymbol{Y}}_t) \\ &= I(M_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) \ . \end{aligned} \tag{9}$$

From this, sufficiency follows from noticing that

$$\begin{aligned} I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | M_t, \overleftarrow{\boldsymbol{Y}}_t) &= I(\overleftarrow{\boldsymbol{X}}_t, M_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) - I(M_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) \\ &= I(\overleftarrow{\boldsymbol{X}}_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) - I(M_t; Y_{t+1} | \overleftarrow{\boldsymbol{Y}}_t) \\ &= 0 \ . \end{aligned}$$

Above, the first equality is due to the chain rule, the second follows from the fact that M_t is a function of $\overleftarrow{\boldsymbol{X}}_t$, and the third uses Eq. (9).

To finish the proof, let us show that $\boldsymbol{M}$ is minimal. For this, consider another $\boldsymbol{U}$ to be another D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$. As $U_t = G(\overleftarrow{\boldsymbol{X}}_t)$ for some function $G(\cdot)$, $\boldsymbol{U}$ corresponds to another partition of the trajectories $\overleftarrow{\boldsymbol{x}}_t$. If there exists no function f such that $f(U_t) = M_t$, that implies that the partition that corresponds to $\boldsymbol{M}$ is not a coarsening of the partition for $\boldsymbol{U}$, and therefore that there exists $\overleftarrow{\boldsymbol{x}}_t$ and $\overleftarrow{\boldsymbol{x}}'_t$ such that $G(\overleftarrow{\boldsymbol{x}}_t) = G(\overleftarrow{\boldsymbol{x}}'_t)$ while $p(y_{t+1}|\overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) \neq p(y_{t+1}|\overleftarrow{\boldsymbol{x}}'_t, \overleftarrow{\boldsymbol{y}}_t)$. This, in turn, implies that there exists a $\overleftarrow{\boldsymbol{x}}'_t$ such that that $p(y_{t+1}|u_t, \overleftarrow{\boldsymbol{x}}'_t, \overleftarrow{\boldsymbol{y}}_t) \neq p(y_{t+1}|u_t, \overleftarrow{\boldsymbol{y}}_t) = \sum_{\overleftarrow{\boldsymbol{x}}_t} p(y_{t+1}|u_t, \overleftarrow{\boldsymbol{x}}_t, \overleftarrow{\boldsymbol{y}}_t) p(\overleftarrow{\boldsymbol{x}}_t|u_t, \overleftarrow{\boldsymbol{y}}_t)$, showing that $\overleftarrow{\boldsymbol{X}}_t$ is not conditionally independent of Y_{t+1} given $U_t, \overleftarrow{\boldsymbol{Y}}_t$, contradicting the fact that $\boldsymbol{U}$ is a D-BaSS. This contradiction proves that the partition induced by $\boldsymbol{U}$ is a refinement of the partition induced by $\boldsymbol{M}$, proving the minimality of the latter.

Proof (Proposition 1). Let's denote by $\boldsymbol{A}$ the minimal D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$, and $\boldsymbol{S}$ the minimal D-BaSS for $\boldsymbol{Y}$ w.r.t. $\boldsymbol{X}$, which are known to exist and be unique thanks to Theorem 1. Then, by defining $M_t \coloneqq (S_t, A_t)$, one can directly verify that $\boldsymbol{M}$ is a ReD-BaSS for $(\boldsymbol{X}, \boldsymbol{Y})$. To prove its minimality, let us consider another ReD-BaSS of $(\boldsymbol{X}, \boldsymbol{Y})$ denoted by $\boldsymbol{N}$. As $\boldsymbol{N}$ is a D-BaSS of $\boldsymbol{X}$ w.r.t. $\boldsymbol{Y}$, the minimality of $\boldsymbol{A}$ guarantees the existance of a mapping $f(\cdot)$ such that $f(N_t) = S_t$. Similarly, thanks to the minimality of $\boldsymbol{S}$, there is another mapping $g(\cdot)$ such that $g(N_t) = A_t$. Therefore, the function $F(\cdot) = (f, g)$ satisfies $F(N_t) = M_t$, which confirms the minimality of $\boldsymbol{M}$.

Proof (Proposition 2). The proof is based on the principle that if $p(A, B, C) = f(A, C)g(B, C)$, then $A \perp\!\!\!\perp B|C$. Building on that rationale, a direct calculation shows that

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= \prod_{\tau=-\infty}^{\infty} p(x_{\tau+1}, y_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau) \\ &= \prod_{\tau=-\infty}^{\infty} \exp\{\xi_\tau\}\, p(x_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau)\, p(y_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau), \end{aligned} \tag{10}$$

where the second equality[12] uses Eq. (5). Additionally, if, as per assumption of the Proposition, $\boldsymbol{R}$ is a ReD-BaSS of $(\boldsymbol{X}, \boldsymbol{Y})$, then

$$p(x_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau) = p(x_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau, r_\tau) = p(x_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, r_\tau),$$

where the first equality uses the fact that r_τ (by definition) is a function of $(\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau)$, and the second uses the sufficiency of D-BaSS's. Following an analogous derivation, one can show that $p(y_{\tau+1}|\overleftarrow{\boldsymbol{x}}_\tau, \overleftarrow{\boldsymbol{y}}_\tau) = p(y_{\tau+1}|r_\tau, \overleftarrow{\boldsymbol{y}}_\tau)$. Then, with the assumption that the dynamics of $(\boldsymbol{X}, \boldsymbol{Y})$ is factorisable and hence $\xi_t = 0$, it

[12] Note that the infinite products in this proof are just a formal procedure to acknowledge products that can be taken up to arbitrary times.

follows from Eq. (10) that

$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{\tau=-\infty}^{\infty} p(x_{\tau+1}|r_\tau, \overleftarrow{\boldsymbol{y}}_\tau)\, p(y_{\tau+1}|r_\tau, \overleftarrow{\boldsymbol{y}}_\tau) \ .$$

Separating the two product series, this shows that there exist functions $f(\cdot)$ and $g(\cdot)$ such that $p(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{r})g(\boldsymbol{y}, \boldsymbol{r})$, and hence one has $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{R}$, which completes the proof.