UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Tiago Martinelli

# Causal modeling in high-order scenarios: unfolding mechanisms by moving across scales

São Carlos

2023

**Tiago Martinelli**

# Causal modeling in high-order scenarios: unfolding mechanisms by moving across scales

Thesis presented to the Graduate Program in Physics at the Instituto de Física de São Carlos da Universidade de São Paulo, to obtain the degree of Doctor in Science.

Concentration area: Theoretical and Experimental Physics

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

Coadvisor: Prof. Dr. Diogo de Oliveira Soares Pinto

**Corrected version**
**(Original version available on the Program Unit)**

**São Carlos**
**2023**

Martinelli, Tiago
Causal modeling in high-order scenarios: unfolding mechanisms by moving across scales / Tiago Martinelli; advisor Francisco Aparecido Rodrigues; co-advisor Diogo de Oliveira Soares Pinto - corrected version -- São Carlos 2023.
91 p.

Thesis (Doctorate - Graduate Program in Theoretical and Experimental Physics) -- Instituto de Física de São Carlos, Universidade de São Paulo - Brasil , 2023.

1. Causal modeling. 2. Emergent phenomena. 3. Multivariate information theory. I. Rodrigues, Francisco Aparecido, advisor. II. Pinto, Diogo de Oliveira Soares, co-advisor. III. Title.

*To my parents Paulo and Sandra.*

# ACKNOWLEDGEMENTS

To my advisor Prof. Francisco A. Rodrigues for his vote of confidence given to a student lost in doubt about pursuing a more applied doctorate instead of a fundamentalist one. And showing me Pearl's book in our first meeting resulted in this work. I hope I did not fall too short of your expectations. Thanks for the almost fatherly concerns regarding my future academic career.

To my co-supervisor Prof. Diogo O. Soares-Pinto allowing our relationship to go beyond the purely academic realm and thus sharing many worldviews. Thanks for the bureaucratic help allowing me to worry only about doing science.

To my parents, Paulo and Sandra, for their unrestricted love and dedication, fundamental ingredients for me to have the necessary strength and foundation to obtain all the achievements, including this one; for understanding my often "absent" behavior, which attests to the affection and comfort they provide me, without which the dedication to my "inner world", necessary for my career, would not be possible. Love you.

To my brother Igor for his spontaneity and practicality, that always is an example of antagonism to my personality.

To my "dinosaurs" friends Matheus, Chris, and Alex, who have been with me since undergraduate and have collaborated professionally and personally throughout this doctorate. Matheus, dedicated time to several difficult moments. Chris for various technical discussions and pointing out several times when I accumulated a lot of information but transmitted it through a very noisy channel. To Alex, for the moments when every ten words I spoke, he laughed eleven.

To Kirstin for being my first contributor and helping me to leave the abstract realm of causality and discuss its mundane usefulness. To Thomas for helping me with the review of papers and this thesis. Also, for his helpful bits of advice in academic life.

To Professor Yamir Moreno for the opportunity to develop research in his COSNET group. For the hospitality and opportunities given throughout this year abroad.

To the various Spanish colleagues, I had the opportunity to meet and provide a pleasant environment during my scholarship abroad. In particular, Carlos, Alfonso, Alberto, Ari and Mario.

To Gabi, who has been a supportive presence throughout a significant portion of my academic journey. While our paths may have diverged in life, I am thankful for the memories we shared and the impact you had on this period of my life.

To the committee members for their availability and patience in carefully reading and analyzing this work.

To CNPq (Grant No. 140665/2018-8) for the financial support during the first year of my Ph.D. and to FAPESP (Grants No. 2018/12072-0 and 2020/04543-2) for the total financial support afterward.

To the people who directly or indirectly contributed to the realization of this goal.

# ABSTRACT

MARTINELLI, T. **Causal modeling in high-order scenarios:** unfolding mechanisms by moving across scales. 2023. 91p. Thesis (Doctor in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2023.

The big data era advanced the possibility of studying emergent phenomena in the real world, often occurring by systems with high-order, non-trivial interactions. One of the main questions for these complex systems is to understand how their collective organization influences the dynamic processes. Although such a study is fundamental to developing the policies of controlling dynamical processes from changes in the network structure, in practice, the only information available is multivariate data recorded from variables with unknown topology. Such a scenario can be explored using information theory and causality tools to quantify an individual's influence and infer a causal structure among them. In other words, we can make reverse engineering to obtain a causal model via data. However, a methodology to deal with emergent causes when extracting information is an open question. If not performed correctly, it can compromise basic assumptions in causal modeling resulting in a spurious view of the organization of complex systems. This thesis is dedicated to investigating fundamental problems regarding the capture of emergence phenomena from high-order complex systems joining techniques from causal manipulative approaches and multivariate information theory. Based on our results, we defend a paradigm shift when dealing with multivariate data in causal modeling by considering the task of a system description by moving scales as a fundamental issue instead of a mathematical artifice.

**Keywords**: Causal modeling. Multivariate information theory. Emergent phenomena.

# RESUMO

MARTINELLI, T. **Modelagem causal em cenários de alta ordem:** revelando mecanismos movendo-se através de escalas. 2023. 91p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A era do *big data* avançou a possibilidade de estudar fenômenos emergentes no mundo real, muitas vezes ocorrendo por sistemas com interações não triviais de alta ordem. Uma das principais questões para esses sistemas complexos é entender como sua organização coletiva influencia os processos dinâmicos. Embora tal estudo seja fundamental para desenvolver a política de controle de processos dinâmicos a partir de mudanças na estrutura da rede, na prática, a única informação disponível são dados multivariados registrados de variáveis com topologia desconhecida. Tal cenário pode ser explorado usando a teoria da informação e ferramentas de causalidade para quantificar a influência de um indivíduo e inferir uma estrutura causal entre eles. Em outras palavras, podemos fazer engenharia reversa para obter um modelo causal via dados. No entanto, uma metodologia para lidar com causas emergentes ao extrair informações é uma questão em aberto. Se não executada corretamente, pode comprometer suposições básicas na modelagem causal, resultando em uma visão espúria da organização de sistemas complexos. Esta tese é dedicada à investigar problemas fundamentais relacionados a captura de fenômenos de emergência em sistemas complexos de alta ordem conciliando técnicas de teorias de causalidade e informação multivariada. Com base em nossos resultados, defendemos uma mudança de paradigma ao lidar com dados multivariados na modelagem causal, considerando a tarefa de descrição de um sistema ao mover escalas como uma questão fundamental ao invés de um artifício matemático.

**Palavras-chave**: Modelagem causal. Teoria de informação multivariada. Fenômenos emergentes.

# CONTENTS

# 1 INTRODUCTION

Causation-based formalisms aim to entail a more fundamental structure on models than an inference based only on statistical relations. Consider a three-node model of possible causal relationships among them, see in Fig.1. Suppose that we have data records from processes $X$ and $Y$ only. Suppose also that two things always happen: first, the hidden node $Z$ (unknown in the data's record) influences $X$ by a `COPY` operation; second, $Z$ influences $Y$, showing that $X$ and $Y$ share redundant information about $Y$. Such a model is quite general since it can also include more interesting cases: that both $X$ and $Z$ synergically influence $Y$ but neither $X$ nor $Z$ provide a specific influence on their own. The latter is an example of dependent (synergic) causes raising an emergent phenomenon.

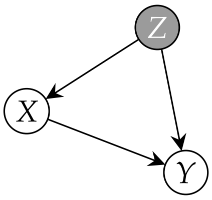


Figure 1 – Graphical model of causal relationships among, nodes $X$ and $Y$, and the confounder (hidden) node $Z$. The great benefit of manipulative approaches to causation is to provide techniques to decide if an arrow is really a cause or only a correlation. (1,2)
Source: By the author.

Now, imagine an ensemble of nodes $\mathbf{X}$ influencing another ensemble $\mathbf{Y}$ with a large number of dependent causal influences. Identifying and distinguishing the different contributions of causal information due to unique against redundant/synergic contributions among them could allow, e.g., to understand better how collective information in the brain contributes to consciousness experience. (3) Besides brain networks, the application domains of synergic causes can also be found in social cognitive phenomena (4), gene regulatory networks (5), and others whose behavior emerges spontaneously from numerous interactions that are not known a priori.

The characterization of information from dependent relationships in causal modeling is an open question in the literature. (6–8) Such a problem is linked with the partition of high-order[1] statistical dependencies in complex systems since without statistical dependency there is no causal dependency. (1) The answer to this problem, then, relies on the intersection of three wide areas: complex systems, causality, and information theories.

The purpose of this thesis is to contribute with an information-theoretic method-

[1] We will reserve the term *high-order* for any complex system that has non-pairwise relationships.

ology to deal with data in the presence of high-order phenomena and noise[2]. We want to defend the following statement,

PROBLEM STATEMENT

To characterize coarse-grainings (Fig.2) as useful strategies to unfold causal mechanisms when dealing with data from complex scenarios in the presence of emergent phenomena.

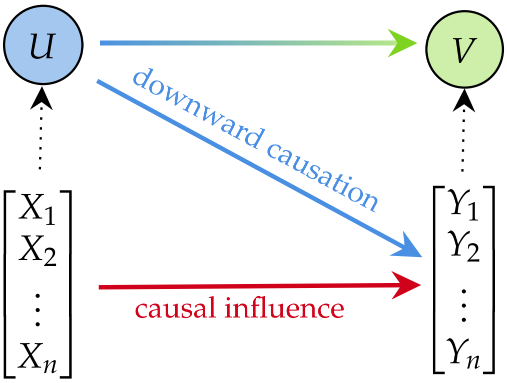


Figure 2 – Coarse-graining[3] the ensembles $\mathbf{X}, \mathbf{Y}$ into macro variables $U, V$, respectively, can show more causal influence beyond micro (individual) components. We will say that we have emergent phenomena when either downward causation (macro-micro causal influence) or causal decoupling (macro-macro causal influence) happens.
Source: Adapted from ROSAS *et al.*(9)

In Sec.1.1, we demonstrate how the rise of emergent phenomena in science and its need for a formal methodology raised the fields of our interest. In Sec.1.2 we list our contributions in such field and how this text is organized.

## 1.1 Statement and motivation

Sixty years ago, digital computers made information readable. Twenty years ago, the internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now, Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the petabyte age. (10)

[2] We will use the term noise to refer to any type of correlation in the data that disrupt the ability to capture meaningful information.

[3] A commonly known example of coarse-graining is in thermodynamics when a macroscale is used, such as the temperature and its relation with the motion of individual particles. The collection of all possible microstates, combinations of particle kinetic energy, is simplified into a single macrostate, temperature. In this case, the mapping of individual particles' states to the whole's temperature is done via a function (average) of kinetic energy.

Every day, we create *2.5 quintillion bytes*[4] of data. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures, and stock market records, to name a few. *This data is big data.*(11) This information era has made possible the storage and analysis of big data to deepen our understanding of the *emergence phenomena* that are at the core of complex systems.(9, 12)

### 1.1.1 Emergence: unpredictability or irreducibility

*Could Laplace's demons exist?*

The term *emergence* was coined in 1875 by the philosopher G. H. Lewes. (13) It comes from the Latin verb *emergo* which means to arise, to rise up, to come up, or to come forth. In Lewes' work, three essential features of emergence are laid out. First, it is a theory about the structure of the natural world; therefore, it has ramifications concerning the unity of science; second, it is a relation between the whole-parts properties; third, the question of emergence is related to the question of the possibility of reduction[5].

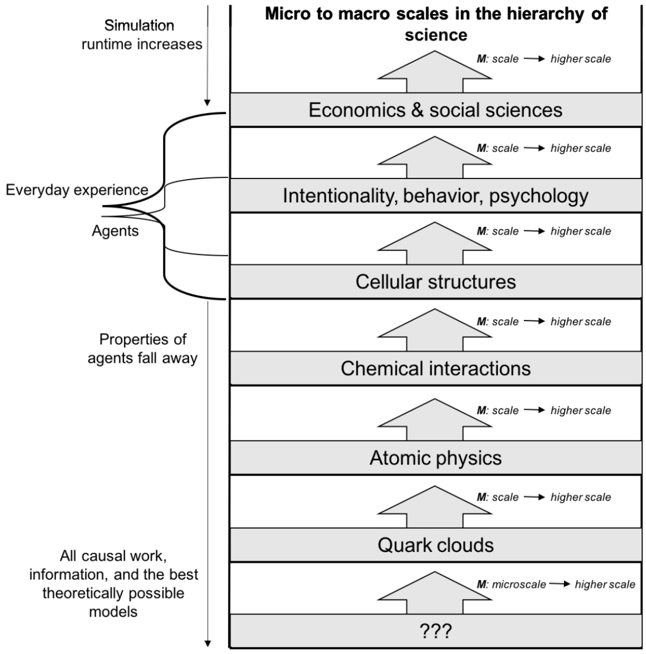

Simulation runtime increases
Micro to macro scales in the hierarchy of science
M: scale → higher scale
Economics & social sciences
M: scale → higher scale
Intentionality, behavior, psychology
Everyday experience
Agents
M: scale → higher scale
Cellular structures
M: scale → higher scale
Properties of agents fall away
Chemical interactions
M: scale → higher scale
Atomic physics
M: scale → higher scale
Quark clouds
All causal work, information, and the best theoretically possible models
M: microscale → higher scale
???

Figure 3 – Reductionist paradigm.

Source: HOEL. (14)

From the reductionist point of view, a biologist studying a cell is really referring to some astronomically complex constellation of quarks, see Fig.3. The reductionist philosophy can be stated clearly using the terminology of information theory: compression of information. Macroscales are useful coarse-grainings being lossless representations of the information contained in the respective structure. Their usefulness stems from the necessity of compression in communication because all systems have limited bandwidth[6] (14). When moving up in scale, the runtime to simulate the full system decreases.

[4] US-based scale defines a quintillion as 1 followed by 18 zeros. British-based scales as 1 followed by 30 zeros. As the source was Forbes magazine, based in New Jersey, we will keep the former.

[5] As we will see during text, the final framework which will rise align with these conditions.

[6] Bandwidth's limit refers to the capacity of a system to transmit data. If these limits are not in place, the device can overload its processing capacity making the system unpredictable.

Given that all systems are describable at the physical microscale, what possible extra gain is there for any coarse-grained macroscale? It would seem natural that some form of Occam's razor applies. That is, the higher-scale descriptions are not necessary at a fundamental level. Unless, the fact that something is irreducible means it exhibits some *causal novelty*, or it contains physical principles not derivable from lower scales. To capture this difference between irreducible and unpredictability, the philosopher David Chalmers (15) drew the distinction between weak and strong emergence.

Weak emergence concerns the type of emergence in which the notions of complexity, self-organization, and non-linearity are central. The core of this position is that a property is emergent if it is systemic, in the sense that none of its smaller parts share, and it is unpredictable given the properties and the laws governing the lower level, the domain from which it emerged. For instance, artificial agents created in Conway's celebrated Game of Life (GoL) (16) simple rules result in highly complex behavior, with recognizable self-sustaining structures, see Fig.4.

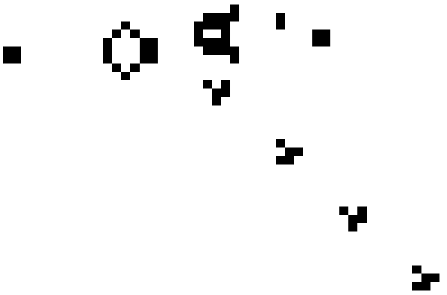


Figure 4 – Gosper glider gun shooting gliders. Simple local rules determine whether a given cell of a $2D$ grid will be ON (black squares) or OFF (white squares) based on the number of ON cells in its immediate neighborhood.
Source: CONWAY'S. (17)

Central to the idea of ontological autonomy of emergence is the concept of *causal novelty.* (13) The principle of strong emergence states that at certain levels of physical complexity, new properties appear that are not found in the parts of the object they emerge from, and cannot be reduced to the fundamental matter from which they emerge. This means that emergent properties have new abilities to cause events that are not explainable by the properties of their underlying parts. Strong emergence, therefore, involves new fundamental properties and new fundamental laws of emergence that involve downward causation, meaning that macroscopic levels can cause events at the microscopic level.

While weak emergence is commonly accepted by some in the scientific community, it is not well-suited to address questions about emergence in situations where the focus is on relationships between parts and wholes. (9) Part of the difficulty in building a deeper understanding of strong emergence was to find a causal role for the macroscales of a system. (18) But a way forward is to consider this issue not from the metaphysical point of view but rather as a problem of causal model choice across scales.

Introduced and refined by Hoel and colleagues (19–21), their work on causal emergence aims to show analytically that macroscopic observables can sometimes exhibit more causal power than microscopic variables. Quantifying the causal power in the micro-to-macro change can be done by combining mathematical tools from information and causal theories, the latter understood within the framework of Pearl's and Woodward's manipulative approach to quantify causation. (1, 2, 22)

Hoel's method is based on a coarse-graining mapping $\mathcal{M} : \mathcal{X} \to \mathcal{V}$ relating macro variables $U, V \in \mathcal{V}$ to the micro ensembles $\mathbf{X}, \mathbf{Y} \in \mathcal{X}$, such that $H(\mathcal{V}) < H(\mathcal{X})$[7]. Causal emergence is declared when the dependency between $U$ and $V$ is *stronger* than the one between $\mathbf{X}$ and $\mathbf{Y}$. To measure this *strength* the framework focuses on a causal capacity, where the distribution $P(X)$ in Fig.5 is chosen to be the maximum entropy distribution $P_{H_{\max}}(X)$, see App.A.2 for further details. Then, the causal capacity searches for the maximum coarse-grained effective information ($EI$):

$$\mathcal{CC} := \max_{M \in \mathcal{M}} EI, \quad EI := I(I_D; E_D). \tag{1.1}$$

aiming to reduce the uncertainty about the future of the system between a manipulated distribution, $P(I_D)$, on the sources set $\mathbf{X}$ and its resultant effect distribution, $P(E_D)$, on the target set $\mathbf{Y}$, see Fig.6.

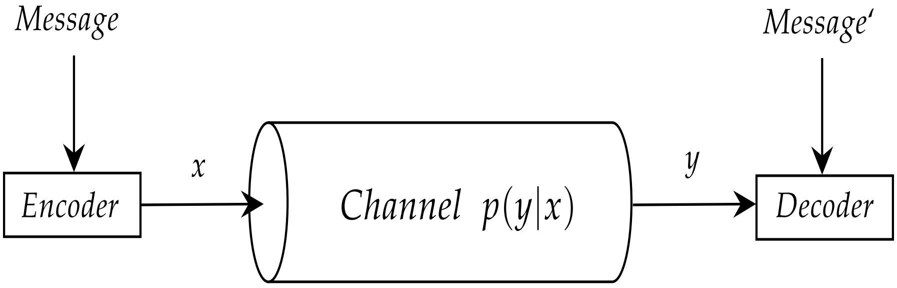


Figure 5 – We can define the capacity of channels to transmit a *message* encoded as information in some probability distribution $P(X)$ as $\mathcal{C} = \max_{P(X)} I(X;Y)$, where $I(X;Y) := H(X) - H(X|Y)$ is the mutual information and $H(\cdot)$ Shannon's entropy. Shannon recognized that the encoding of information for transmission over the channel could change $P(X)$: therefore, due to the natural noise, the channels' capacity is approached by using error-correcting codes to approximate the received *message'*, decoded from information $P(Y)$, to the original one. (23)
Source: By the author.

[7] We will consider dynamical systems as temporal Markov chains of order 1 where $\mathbf{Y} \equiv \mathbf{X}_{t+1}$ and $V \equiv U_{t+1}$ having, then, a stationary distribution.

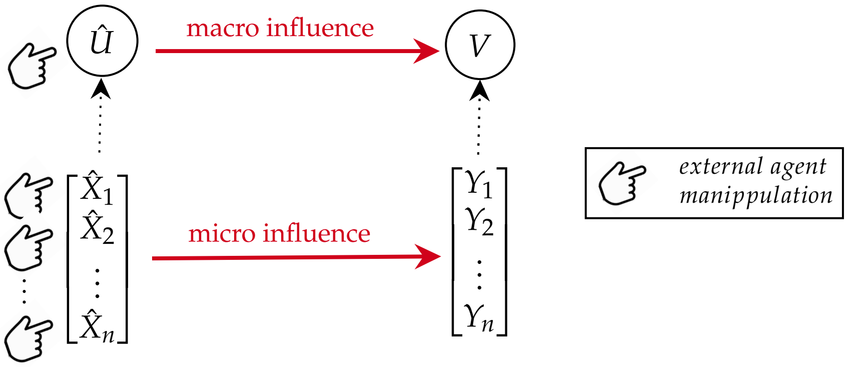


Figure 6 – The manipulative theories of causation argue that to assess the causal structure of a system one has to analyze its response in intervened/manipulated causal models. An external agent "set" a variable $X$ to a specific value, represented above by $\hat{X}$, destroying all incoming influences on it. The causal effect then is the impact from $\hat{X}$ to $Y$ codified by $P(\mathbf{Y}|\hat{\mathbf{X}})$. (24) By setting the entire $\mathbf{X}$ and weighting it according to the maximal entropic distribution $P(I_D) = P_{H_{\max}}(\hat{\mathbf{X}})$, we can measure the (micro) causal influence on the resultant effect distribution on $\mathbf{Y}$ given by $E_D = \sum P(\mathbf{Y}|\hat{\mathbf{X}})P_{H_{\max}}(\hat{\mathbf{X}})$ using mutual information. Applying the same reasoning in coarse-grained variables we have the macro causal influence.
Source: By the author.

### 1.1.2 Agents, special sciences, and multi-information

> *"If it isn't literally true that my wanting is causally responsible for my reaching and my itching is causally responsible for my scratching... if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world."*
>
> *— Jerry Fodor (25)*

In parallel to the program of explaining emergence in causal roots as discussed in Sec.1.1.1, the advances of special sciences and the central role played by agents[8] in these theories challenged the reductionist thinking with Putnam's arguments for multiple realizability. (26) The multiply-realizable property says that different microscopic scales, laws, or mechanisms may lead to the same agent behavior[9].

In an attempt to formalize the specialty of agents (intelligent behaviors), psychologists have argued that a clue could be in analyzing the agent whose adaptive behavior is contingent on multiple signals that interact in complex ways.(4) Indeed, this view was

[8] Agents are generally somewhere above biological mechanisms but below economics on the ladder, Fig.3.

[9] If agents are multiply realizable, then any attempts to link the unique properties of agents to some unique property of microscopic physics are doomed to underdetermination. This culminated in Fodor's argument for the autonomy of the special sciences. (27)

proposed by cognitive scientists as a theory of information processing capturing how combinations of inputs contribute to the final output. (28) The candidate from which to build such a general theory has been the mutual information, see Fig.7.

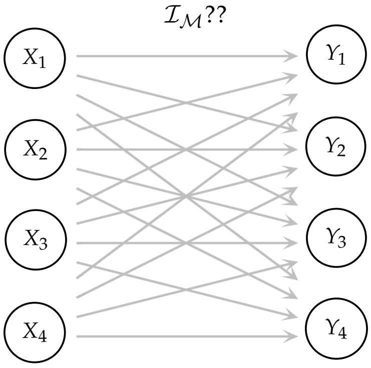


Figure 7 – Aiming to expand the concept of mutual information, the multivariate formalisms work in the informational space of the sources and targets $\{\mathbf{X};\mathbf{Y}\} \mapsto I(\mathbf{X};\mathbf{Y})$ describing emergent phenomena through the informational map $\mathcal{I}_{\mathcal{M}}$. The goal is to isolate the synergic behavior imposing constraints when quantifying the mutual information. (29–32)
Source: By the author.

To illustrate, let's consider an example that embodies the notion of emergent behavior due to systemic, not individual information. (9) For simplicity, we assume that at the initial time, the system is found in a random configuration, i.e. $P_{\mathbf{X}}(\mathbf{x}_{\text{init.}}) = \frac{1}{2^n}$.

**Example 1.1 ($n$-bit `XOR`)**

Consider a system where the parity of $\mathbf{X}$ determines $\mathbf{Y}$, i.e., $Y_2 = \sum_{i=1}^{n} X_i \mod 2$ is the "$n$-bit" `XOR`, and $Y_j$ for $j \neq 2$ is a fair coin flip independent of $\mathbf{X}$ (see Fig.8).

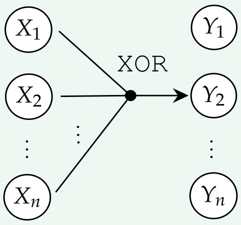


Figure 8 – System's parity determines one element only.
Source: By the author.

In this scenario $\mathbf{X}$ predicts $Y_2$ with perfect accuracy, while it can be verified that $X_i \perp\!\!\!\perp Y_j$ for all $j \in \{1, 3, \ldots, n\}$. In informational terms, one has that $I(X_i, Y_2) = 0$ but $I(\mathbf{X}, Y_2) \neq 0$: the entire system has an effect over a particular element under the proposed evolution rule, even though such an effect cannot be attributed to any individual part. (32)

Despite some efforts to isolate the "systemic" contribution, much of the last half-century has been spent struggling to expand the concept of mutual information to multivariate systems allowing negative information without clear interpretations. (33,34) To overcome it, a nonnegative decomposition of mutual information was proposed under the name of partial information decomposition (PID). (30)

Essentially, the PID formalism provides a method by which the mutual information between the joint state of $n$ sources variables and $m$ target variables[10] can be decomposed,

$$I(\mathbf{X};\mathbf{Y}) = \text{Red}(\mathbf{X};\mathbf{Y}) + \text{Uni}(\mathbf{X};\mathbf{Y}) + \text{Syn}(\mathbf{X};\mathbf{Y}). \tag{1.2}$$

Eq.1.2 proposes the following interpretability for the map $\mathcal{M}$ from Fig.7: the Red term means the information about $\mathbf{Y}$ that is redundantly shared (i.e. an observer could learn the same information about $\mathbf{Y}$ examining any element in this set) among the elements in a well-defined set $\mathcal{R}(\mathbf{X})$[11], Uni refers to the information about $\mathbf{Y}$ that is uniquely present in independent sources $X_i$'s, and Syn is the information about $\mathbf{Y}$ that is only unfolded by the joint states of elements inside the power set $2^{\mathbf{X}}$ considered together.

As we will see, this approach provides a formal mathematical structure for the joint set $\texttt{Red} \cup \texttt{Uni} \cup \texttt{Syn}$ as well as interpretable quantities (partial atoms above) suited to the study of complex information processing. To illustrate, consider the most simple scenario where $\mathbf{X} = \{X_1, X_2\}$ and $\mathbf{Y} = Y$ we can illustrate the *partial atoms* in the diagram of Fig.9.

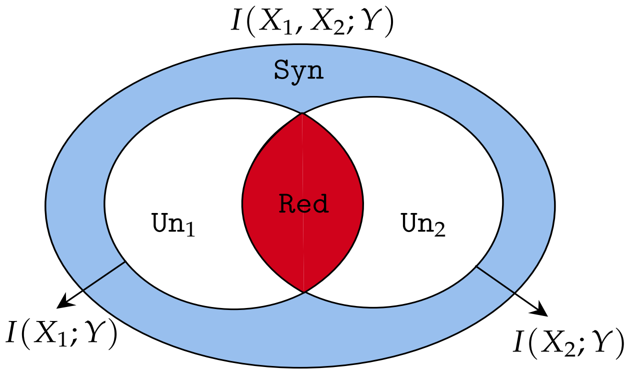


Figure 9 – The $\texttt{Red}$ set (red region) has a single element: the information about $Y$ an observer could learn the same information about $Y$ examining either $X_1$ or $X_2$. The $\texttt{Uni}$ set has two elements: $\texttt{Un}_i$ (white regions) representing the information about $Y$ that is uniquely present in $X_i$ and not in $X_j$, $j = \{1, 2\}$. And, $\texttt{Syn}$ (blue region) a single element: the information about $Y$ that is only unfolded by the joint states of $X_1$ and $X_2$.
Source: By the author.

An inconvenience of such a structure is that the number of atoms, the cardinality of $\texttt{Red} \cup \texttt{Un} \cup \texttt{Syn}$, grows super-exponentially with the number of sources.(30) A recent and promising idea put forward by Rosas & Mediano (9) is to coarse-grain the decomposition

[10] The formalism was expanded to multi-target settings under the name of ΦID. (32)

[11] Later we will provide a mathematical definition for $\mathcal{R}$. From now we are interested in the conceptual idea of formalism.

according to some specific criteria preserving the interpretability synergy, redundancy, and unique information[12]. This allows us to formulate coarse-grained PID decompositions in Eq.1.2 with a small number of atoms that scale linearly with the system's size.

Based on this PID extension, recent investigations purpose to quantify how effectively an agent utilizes information in conscious processing. (3) Giving further support through previous studies using PID where the importance of synergistic information processing in intelligent decision-making has been demonstrated, as seen in strategic behavior in games such as poker.(4)

## 1.2 Contributions and text organization

One of the gains of the $EI$-approach is that it proposes an explanation of how multiply-realizable entities can play a more significant role: through error-correction in causal relationships, making them more informative than their underlying microscale. (35) Despite that, by computing mutual information terms using maximum entropy distributions the approach account for potential transitions the system could do. This is not well-suited to analyze dynamic systems where such transitions can never occur. (36, 37) Also, the way to build the coarse-granings (mappings $\mathcal{M}$) is uncleared for general systems. It would be interesting to use informational criteria agnostic to the system's knowledge.

In light of that, a conciliation with Equation 1.2 could be convenient since both seek to look for a similar answer, that is, to explain emergent phenomena by using informational quantifiers. Whilst PID seems at first glance to define precisely the specific contributions from sources to targets, Eq.(1.2), it remains open the task of incorporating it in the causal machinery among the relationships between $\mathbf{X}$ and $\mathbf{Y}$. In this text we put forward some results in order to narrow the gap between these two approaches by showing how one can benefit from the other, aiming to deal with the concept of emergence based on causal information. Based on our investigations we had the main results,

FIRST CONTRIBUTION

Merge concepts from causation theories and multivariate information theory in order to characterize emergent phenomena. (https://arxiv.org/abs/2203.10665)

SECOND CONTRIBUTION

Show evidencies where coarse-grainings preserve emergent phenomena (when present) and eliminate redundancy revealing causally-relevant information. *(To be published.)*

Also, during the course of this thesis, we had the opportunity to study others topics not incorporated in this text. As part of this work appeared in the following works:

[12] As we will show later, we have used similar reasoning in our results.

- Quantifying quantum reference frames in composed systems: Local, global, and mutual asymmetries. Tiago Martinelli and Diogo O. Soares-Pinto, Phys. Rev. A 99, 042124, 2019. https://doi.org/10.1103/PhysRevA.99.042124;

- Data Study Group team. Data Study Group Final Report: UK Dementia Research Institute. Using machine learning to improve sleep habits in Dementia patients. Zenodo. 2022.https://doi.org/10.5281/zenodo.6798769;

- Roster KO, Martinelli T, Connaughton C, Santillana M, Rodrigues F. Estimating the impact of the COVID-19 pandemic on dengue in Brazil. (Preprint). 2023. doi: https://doi.org/10.21203/rs.3.rs-2548491/v1.

- Martinelli T, Aleta A, Rodrigues F, Moreno Yamir. An informational approach to uncover the age group interactions in epidemic spreading from macro analysis. 2023. http://arxiv.org/abs/2306.00852.

### 1.2.1 Text organization

The present dissertation was organized as follows: this chapter 1 introduced and motivated our problem of interest; chapter 2 gives a contextualization from the area of causal inference. Introduces quickly the role of causality in science focusing on an overview of the manipulative approaches from Pearl and Woodward. (22, 24) We finalize this part by highlighting two points: emergent causes raise the need to deal with stochastic interventions, these which cannot be described in terms of atomic ones as delineated by the do-operator (1,38); and, we argue that $EI$-approach for causal emergence can be seen as a methodology to find the appropriate level of granularity in causal models, which in Woodward's terms it gives the better causal explanation where concepts can be guaranteed in order to apply a manipulative approach.

In chapter 3 we start giving an overview of the PID framework. We clarify the importance of the background context to capture pure synergism in terms of the conditioning operation in mutual information. In what follows, we make use of simple simulated systems, focusing on higher-order interactions, to show quantitatively that genuine causal synergism violates the causal faithfulness assumption and how redundant-dominated systems recover faithfulness due to spuriousness while promoting the failure of causal minimality assumption promoting poor causal models (FIRST CONTRIBUTION). This observation caves the path for a coarse-grained approach to causal modeling. Then, we point out how PID quantifiers can have a causal interpretation by using non-atomical interventions. Finally, we show that the $EI$ measure can be justified in terms of PID by preserving synergism and reducing redundancy (SECOND CONTRIBUTION). Chapter 4 concludes the text by delineating questions to answer in future work.

# 2 ACCESSING CAUSATION IN THE BIG DATA AGE

*Order is not sufficient. What is required, is something much more complex. It is order entering upon novelty; so that the massiveness of order does not degenerate into mere repetition; and so that the novelty is always reflected upon the background of the system.*

*A. N. Whitehead on "Ideal Opposites" in Process and Reality.*

Attempts to analyze causation in the processes of nature come back to the 17th century. A notable philosopher was David Hume, who stipulated that causes are invariably followed by their effects:

> "We may define a cause to be an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second. (39)"

Such causality view, in terms of invariable patterns of succession, started the program referred to as regularity theories of causation. As the area was being developed several well-known problems were faced culminating in the probabilistic approaches to causation. To name a few: the inability to handle spurious correlations and the failure of physical determinism. The former is probably the greatest source of attraction for probabilistic theories of causation. (24, 40) For the latter, if an event $E$ is not determined to occur then no other event $E'$ can be a sufficient condition for $E$ arising imperfect regularities. For this reason, many philosophers desire to develop a theory of causation that does not presuppose determinism.(41)

Probabilistic theories of causation did much to illuminate the relationship between causation and probability. However, despite the mathematical framework, and points of contact with statistics and experimental methodology, it did not give rise to any new computational tools or suggest any new methods for detecting causal relationships.(42) For this reason, the program has largely been supplanted by the causal modeling tools described in the following sections 2.1, 2.2 and 2.4. Within causal modeling programs, the most significant works are those developed in computer science, Pearl (1), and philosophy, Spirtes *et al.* (43) The work of Woodward (2, 22) is also significant for us since it aims to establish a broader philosophical foundation for the concept of intervention, which serves as the basis for a manipulable/interventionist theory of causation. As we will argue, this interventionist character is crucial to deal with the concept of emergent causes.

Manipulability theories state that an event $C$ is considered a cause of another event $E$ only when manipulating $C$ can be used to control $E$. Pearl and Woodward are known for their support of interventionist causation. (41) Pearl's approach emphasizes using global network constraints to correctly identify local causal relations, while Woodward's approach emphasizes using local manipulation to identify single causal relations that together compose global causal structures, see Fig.10. It has been argued in the literature (44) that these distinct perspectives have, indeed, a complementary relationship.

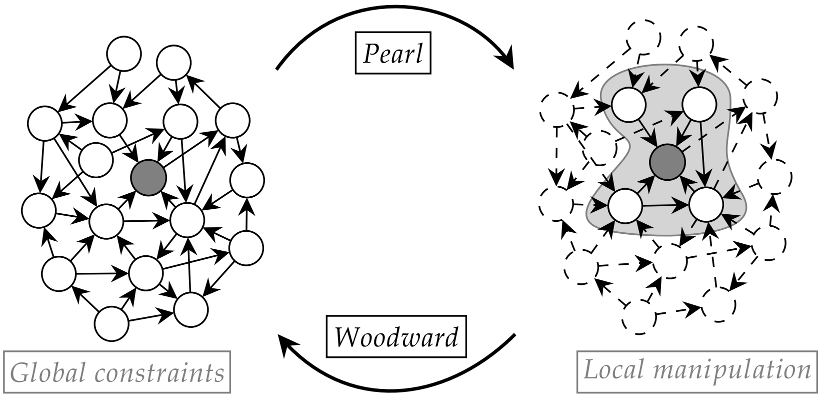


Figure 10 – Pearl's focus is on leveraging global network constraints to correctly identify local causal relations, left-hand side (l.h.s) → right-hand side (r.h.s.) Woodward's focus is on the use of local manipulation to identify single causal relations that then compose into global causal structures, l.h.s ← r.h.s.
Source: By the author.

In this chapter, we begin by briefly reviewing probabilistic causation and its replacement by manipulative approaches, as outlined in HITCHCOCK.(42) We then present Pearl's framework for causal modeling and its significance for enabling interventionism. In what follows, we dedicate some space to discussing information-based causation measures and how they exploit the concept of non-atomic interventions. We highlight that such measures are useful to examine Woodward's philosophical approach to causation clarifying precisely the range of interventions and background conditions to reveal causal relationships. This allows us for a discussion of causal models across different scales, which, when not done correctly, can lead to misinterpretation of significant physical phenomena such as emergent causes. Regarding emergence, we emphasize that the condition of faithfulness/stability is crucial in this aspect. Finally, we suggest that $EI$'s approach can be seen as a first step towards an informational-based methodological approach for finding the optimal causal model scale in Woodward's sense.

## 2.1 A historical path through causality

Probabilistic theories of causation propose that causes alter the probability of their effects. An effect may still occur without a cause or fail to occur with its presence. For example, smoking increases the probability of developing lung cancer, even though some smokers do not develop lung cancer. Therefore, smoking is a cause of lung cancer, not because all smokers develop it, but because smokers are more likely to develop it than non-smokers.(41)

To model systems that do not follow a deterministic behavior, the language of stochastic processes is useful. The shift towards probabilistic theories of causation moves away from deterministic settings and causes relationships to take on a probabilistic form, instead of the traditional "all-or-nothing" approach. Consider the following premise:

> If event $C$ happens today, then event $E$ will happen tomorrow.

Mathematically:

> The conditional probability of event $C$ happening given that $E$ happened before $C$ is 1,
>
> $$P(E|C) = 1. \tag{2.1}$$

In real science, however, things are rarely represented as Eq.(2.1). Normally, scientific laws take a probabilistic form:

> If event $C$ happens today, then event $E$ is very likely to happen tomorrow.
>
> $$P(E|C) > 1 - \epsilon, \forall \epsilon \in [0, 1]. \tag{2.2}$$

One could even have a weaker statement:

> If event $C$ happens today, then event $E$ is more likely to happen tomorrow.
>
> $$P(E|C) > P(E), \tag{2.3}$$

with this equation being the first establishment at a probabilistic theory of causation (42). However, Eq.(2.3) raises some troubles with spurious correlations as discussed by Hans Reichenbach.

### 2.1.1 The Hans Reichenbach's Common Cause Principle

Suppose that events $C$ and $E$ are positively correlated, i.e., that

$$P(C|E) > P(C) \tag{2.4}$$

But suppose that neither $C$ nor $E$ is a cause of the other. This is the situation shown in Fig.11 below.

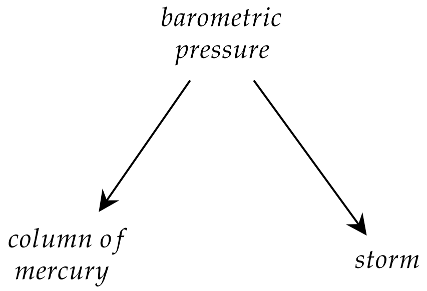


Figure 11 – Here, the variable $C$ is represented by the drop in the level of mercury in a barometer, and variable $E$ is the occurrence of a storm. The atmospheric pressure, variable $B$, is referred to as a common cause (confounding factor). Source: Adapted from HITCHCOCK. (41)

Reichenbach sustained that there will be a common cause, $B$, of $C$ and $E$, satisfying the following conditions (the symbol $\neg$ stands for the logical negation):

**(Reichenbach) Common Cause Principle**

① $P(C, E|B) = P(C|B)P(E|B)$;
② $P(C, E|\neg B) = P(C|\neg B)P(E|\neg B)$;
③ $P(C|B) > P(C|\neg B)$;
④ $P(E|B) > P(E|\neg B)$.

The idea here of that all kinds of correlations between $C$ and $E$ are causally present in a common cause. Conditions ① and ② stipulate that $B$ and $\neg B$ screen off $C$ from $E$. Conditions ③ and ④ follow from $B$ being a cause of $C$ and a cause of $E$. With these conditions, Eq.(2.4) is mathematically entailed using ① until ④. Therefore, probabilistic correlations that are not the result of one event causing another (missing arrow between $C$ and $E$) are derived from probabilistic correlations that do result from a common causal relationship ($B$). (42)

### 2.1.2 The raise of manipulative theories

Manipulative theories argue that without semantics there is no reason to assume that purely probabilistic information should support causal reasoning. The relationship “a cause $C$ raises the probability of its effect $E$” is manipulative in nature, and cannot be captured in the language of probability theory. Therefore, inequalities such as:

$$P(E|C) > P(E) \tag{2.5}$$

are misguided from the start since manipulative raising cannot be reduced to observational raising. The correct inequality, according to a manipulative theory, should read:

$$P(E|\text{do}(C)) > P(E) \tag{2.6}$$

where do($C$) stands for an external agent manipulating on $C$ (remember the discussion involving Fig.6 in introduction). The conditional probability $P(E|C)$ represents a probability resulting from a passive observation of $C$, and rarely coincides with $P(E|\text{do}(C))$. In the example of Fig.11, an observation of the falling barometer indicates a higher probability of a storm, but it does not directly cause the storm. If manipulating the barometer could affect the probability of a storm, then the falling barometer would be considered a cause of the storm.

From here on, we denote random variables with capital letters, $X$, and their associated outcomes using lower case, $x \in \mathcal{X}$ where $\mathcal{X}$ is the space of realizations. Random vectors, of size $n$, will be denoted by bold capital letters, $\mathbf{X} = \{X_{[n]}\}$[1].

## 2.2 Pearl's diagrams as the oracle for interventions

Pearl developed his formalism while writing artificial intelligence programs. He aimed to capture how we learn about the world with only limited actions in a noisy environment.(44) The process of searching for relevant probabilistic data can be exceptionally arduous, even for basic environmental models[2]. Pearl showed that by identifying conditional (in-)dependencies, a Bayesian network or a directed acyclic graph (DAG) (defined formally in the next section) provides a possibility for representing a joint probability distribution $P$ in a more compact form.(1) Afterward, by noting the earliest attempt, due to the geneticist Sewall Wright who makes use of diagrams linking the notion of probabilistic dependence to one of the causal mechanisms, Pearl formalized the language of diagrams utilizing DAGs to uncover explicitly causal structure.

Wright's ideas came back to the 1920's (45, 46) to express mathematically the common understanding that symptoms do not cause diseases. If $X$ stands for a disease variable and $Y$ stands for a certain symptom variable of the disease, Wright would write a linear equation:

$$Y = \beta X + U_Y, \tag{2.7}$$

with $X$ and $Y$ standing for the disease's and symptom's severity, respectively, and $U_Y$ standing for all unknown factors that could affect $Y$. In interpreting this equation one should think of a physical process whereby nature examines the values of $x$ and $u_Y$ and, accordingly, assigns variable $Y$ the value $y = \beta x + u_Y$. Similarly, to explain the occurrence of disease $X$, one could write $X = U_X$, where $U_X$ stands for all factors affecting $X$.

Wright added path diagrams to his equation to depict the inherent directionality of the process. In these diagrams, arrows are drawn from causes to their effects, and the

[1] We will make use of the shorthand notation $[n] := \{1, \ldots, n\}$.

[2] Suppose that one has a table of statistical data of a set containing $n$ variables. If each variable has $k$ possible values, then to exactly specify a probability distribution over all possible combinations of values in the model, one needs $k^n - 1$ parameters.

lack of an arrow implies that Nature assigns values to one variable without considering the other. (47) For example, the absence of an arrow from $Y$ to $X$ represents the claim that symptom $Y$ is not among the factors $U_X$ which affect disease $X$, see Fig.(12-a). The variables $U_X$ and $U_Y$, known as *exogenous variables*, represent background factors that the modeler chooses to leave unexplained. These factors influence, but are not influenced by the other variables, known as *endogenous variables*, in the model. If a correlation is believed to exist between the exogenous variables $U_X$ and $U_Y$, it is common to connect them with a dashed double arrow, as shown in Fig.(12-b). The generalization to nonlinear systems of equations can be seen in Fig.(13-a).

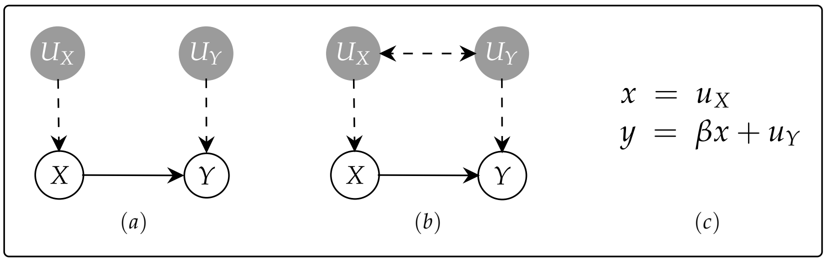


Figure 12 – A simple equation model (c), and its possible associated path diagrams (a) and (b). Unobserved exogenous variables ($U_{(\cdot)}$) are connected by dashed arrows. Missing double dashed arrows between $U_X$ and $U_Y$ (a), represent the assumption of the covariance among two random variables be zero, $\text{Cov}(U_X, U_Y) = 0$ against $\text{Cov}(U_X, U_Y) \neq 0$ in (b) where Cov represents the covariance between two random variables.
Source: By the author.

Notably, and probably the reason for their success in causal modeling, path diagrams provide a formal interpretation and symbolic machinery for analyzing manipulated relationships of the type: "$Y$ would be $y$ had $X$ been $x_0$ in situation $U = u$", denoted $Y_{x_0}(u) = y$. Here $U$ represents the vector of all exogenous variables. The key idea is to interpret the phrase "had $X$ been $x_0$" as an instruction to modify the original model and replace the equation for $X$ by a constant $x_0$, yielding the sub-model, Fig.(13-b).

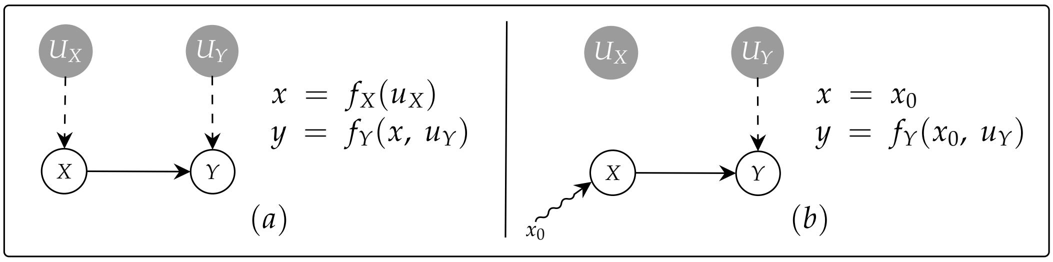


Figure 13 – (a) The path diagram associated with its model's equations. (b) The perturbed diagram associated with the modified model, $Y_{x_0}$, representing the intervention $do(X = x_0)$. The graphical effect of intervening on a variable is the rupture of all arrows pointing to it.
Source: By the author.

This replacement permits the constant $x_0$ to differ from the actual value of $X$ without rendering the system of equations inconsistent, thus yielding a formal interpretation of interventions/counterfactuals in multi-stage models. For example, to compute the effect of setting $x$ to $x_0$, also called the causal effect of $x$ on $Y$, denoted $P(Y|\text{do}(X = x_0))$ or, generically, $P(Y|\text{do}(x_0))$, we solve equation Fig.(13-b) for $Y$ in terms of the exogenous variables, yielding $Y_{x_0} = f_Y(x_0, u_Y)$, and average over $u_Y$. In this simple system, the answer can be obtained without knowing the form of the function $f_Y(x_0, u_Y)$ or the distribution $P(u_Y)$ and is given by: $P(Y_{x_0}) = P(Y|\text{do}(X = x_0)) = P(Y|x_0)$ which is computable from observed samples of $P(x, y)$. This result hinges on the assumption that $U_X$, and $U_Y$ are mutually independent on the topology of the graph, i.e., $\text{Cov}(U_X, U_Y) = 0$.

We can discuss more examples to get a feel of the do operator to reveal causal information when we roll two dices or we study the influence of smoking on lung cancer as seen below. Indeed, example 2.2 elucidates to us that, without causal semantics, purely probabilistic information can give wrong data interpretation according to Eq.(2.12). The observational probability $P(cancer|(smoking)) = 85.25\%$ says wrongly that smoking "causes´´ cancer while the interventional $P(cancer|\text{do}(smoking)) = 47.5\%$ reveals that this is not true. The explanation here could be an unknown genetic factor encoded in the unmeasured confounder $U$ inflating $P(cancer|(smoking))$.

### Example 2.1 (Rolling two dices (38))

Let us consider the structural model representing the setting when two dice are rolled but only the sum and the difference of their values are observed:

$$\mathcal{M} = \begin{cases} X = U_X + U_Y \\ Y = U_X - U_Y \end{cases} \tag{2.8}$$

and $P(U_\alpha = i) = 1/6, \alpha = \{X, Y\}, i = \{1, \dots, 6\}$. Consider the different dice configurations compatible with $Y = 0$: $(U_X, U_Y) = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$. Since each of the $U$'s realization happens with probability $1/36$, the event of the difference between the first and second dice being zero $(Y = 0)$ occurs with probability $1/6$.

If we know that the sum of the two dice is two, i.e., $X = 2$, the probability of the difference of the two dice being zero $(Y = 0)$ becomes certain (probability 1). This is because the only event that is compatible with $(X = 2, Y = 0)$ is $(U_X = 1, U_Y = 1)$. So the knowledge of $X = 2$ makes the relationship between $X$ and $Y$ deterministic. Now, suppose that the observer decides to misreport the sum of the two dice as 2, which can be written as submodel $\mathcal{M}_{X=2}$:

$$\mathcal{M}_{X=2} = \begin{cases} X = 2 \\ Y_{X=2} = U_X - U_Y \end{cases} \tag{2.9}$$

while $P(U_X, U_Y)$ remains unchanged. Note that $Y_{X=2}$ is the same as $Y$; in other words, misreporting the sum of the two dice will not change their difference. This entails the following probabilistic invariance,

$$P(Y = 0|\text{do}(X = 2)) = P(Y = 0). \tag{2.10}$$

In fact, the distribution of $Y$ when $X$ is fixed to $X = 2$ remains the same as before, i.e., $P(Y = 0|\text{do}(X = 2)) = 1/6$. We saw above that knowing that the sum was two meant that, with probability one, their difference had to be zero, $P(Y = 0|X = 2) = 1$. On the other hand, intervening on $X$ will not change $Y$'s distribution, Eq.(2.10); in other words, $X$ does not have a causal effect on $Y$.

### Example 2.2 (Smoking & lung cancer (1))

Let us now examine some hypothetical data to understand how the distinction between observations and interventions can lead to different conclusions about causality. Consider a toy structural model expressed in the causal diagram of Fig.14 with data:

1. 47.5% of the population are nonsmokers with no tar in their lungs, and 10% of these get cancer;
2. 2.5% are smokers with no tar, and 90% get cancer;
3. 2.5% are nonsmokers with tar, and 5% get cancer;
4. 47.5% are smokers with tar, and 85% get cancer.

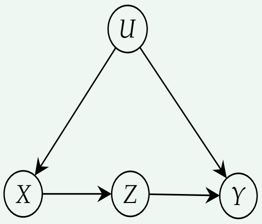


Figure 14 – Causal diagram $G$ for the toy model. $U \equiv$ hidden factor; $X \equiv$ smoking; $Z \equiv$ tar in lungs; $Y \equiv$ lung cancer.
Source: Adapted from PEARL. (1)

When the causal diagram satisfies some criterion relative[a] to $(X, Y)$ and if $P(x, z) > 0$, then the causal effect of $X$ on $Y$ is identifiable and is given by:

$$\mathrm{P}(y|\text{do}(x_0)) = \sum_z \mathrm{P}(z|x_0) \sum_{x'} \mathrm{P}(y|x', z)\mathrm{P}(x') \tag{2.11}$$

$$P(\mathit{cancer}|\text{do}(\mathit{smoking})) = 47.5\% \quad \text{and} \quad P(\mathit{cancer}|(\mathit{smoking})) = 85.25\% \tag{2.12}$$

[a] These are known as back-door and front-door criteria, see App.B.1 for further details.

### 2.2.1 Causal Markov, Faithfulness, and Minimality

Pearl's account of causality, which is based on probabilistic causal models (Def.2.1), involves two premises: the causal Markov and faithfulness assumptions. These premises are applied at both the graphical and statistical levels, and they serve as a global constraint to ensure that the probability distributions are appropriate for describing causal structures.

The way to represent the causal structure among the causal processes behind variables in $\mathbf{X}$ is by directed acyclic graphs (DAGs) $G$. Formally, a directed graph $G$ is a tuple on the variable set $\mathbf{X}$ and the set of ordered pairs $(i,j) \in \mathcal{E}$ of variables in $\mathbf{X}$. A path in a directed graph is a non-repeating sequence of arrows that have endpoints in common. A directed path is a path in which all the arrows point in the same direction. A directed graph is acyclic, and hence a DAG, if there is no directed path from a variable to itself. The directionality encapsulates the cause-effect mechanism among nodes and the acyclicity guarantees the impossibility to have effects before causes.

The relationships in the DAG are often described using the language of genealogy. The node $j$ is called a parent of the node $i$ (the children) just in case there exists the arrow $j \to i$. If $i$ is a parent of $j$, we call any other parent $i'$ of $j$ the spouse of $i$. The node $i$ is called the descendant of $j$, (and $i$ is called an ancestor of it) if there exists a directed path $i \to \cdots \to j$ in $G$. The set of all parents, children, spouses, descendants, and ancestors of $i$ will be denoted, respectively, by $\mathrm{PA}_i$, $\mathrm{CH}_i$, $\mathrm{SP}_i$, $\mathrm{DE}_i$, and $\mathrm{AN}_i$.

**Definition 2.1 (Causal Models)**

A causal model $\mathcal{M}$ is a 2-tuple $\langle G_{\mathbf{u}}, P_{\mathbf{u}} \rangle$, where $G_{\mathbf{u}} = (\mathbf{X}, \mathcal{E})$ is directed acyclic graph (DAG) with edges $\mathcal{E}$ that indicate the causal connections among a set of nodes $\mathbf{X}$ and a given set of background conditions (state of exogenous variables) $\mathbf{U} = \mathbf{u}$ encoded in $P(\mathbf{u})$. The nodes in $G_{\mathbf{u}}$ represent a set of associated random variables with the probability function $P_{\mathbf{u}} = P_{\mathbf{u}}(\mathbf{X})$ given by

$$P_{\mathbf{u}}(\mathbf{X}) = \prod_i P(X_i|\mathrm{PA}_i), \quad P(X_i|\mathrm{PA}_i) = \sum_{u_i} P(X_i|\mathrm{PA}_i, u_i)P(u_i). \tag{2.13}$$

where $\mathrm{PA}_i$ defines the parents for any node $X_i \in \mathbf{X}$, see Fig.15-(B) for an illustration. For a causal graph, there is the additional requirement that the edges $\mathcal{E}$ capture causal dependencies (instead of only correlations) between nodes[a],

$$P(\mathbf{X}) = \prod_i P(X_i|\mathrm{PA}_i) = \prod_i P(X_i|\mathrm{do}(\mathrm{PA}_i)). \tag{2.14}$$

This means that the decomposition holds even if the parent variables are actively set into their state as opposed to passively observed in that state, causal Markov condition ($CMC_{\text{Factorization}}$).(1,38)

[a] The fact that the background variables $\mathbf{U}$ are conditioned to a particular state $\mathbf{u}$ throughout the causal analysis they will, otherwise, not further considered in $P_{\mathbf{u}}(\mathbf{X})$.

**Remark 1.** *In some scenarios, causal sufficiency is a strong assumption. In that case, causal diagrams are supplemented by semi-markovian models to account for the existence of unobserved confounders (common causes). In that case, Eq.(2.14) can be achieved considering a specific ordering of the parents set. (38)*

Pearl developed a criterion called directional separation ($d$-separation, for short) to let us inspect graphically a causal model and to conclude when two random variables in the model cannot tell us anything about the value of each other. Such criterion, which will be called $CMC_{\text{d-separation}}$ states that $d$-separation is sufficient for conditional independence,

$$(X_i \perp\!\!\!\perp_G Y|Z) \implies (X \perp\!\!\!\perp_P Y|Z). \tag{2.15}$$

The concept of $d$-separation is defined via the concept of blocked paths in a DAG. In a DAG, a path between $i_1$ and $i_n$ is blocked by a set $S$ (neither containing $i_1$ nor $i_n$) whenever there is a node $i_k$ in the path, such that one of the following possibilities holds:

(i) $i_k \in S$ and $i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$;

(ii) $i_{k-1} \rightarrow i_k \leftarrow i_{k+1}$ and neither $i_k$ nor $\text{DE}_{i_k}$ are in $S$.

We say that two disjoint subsets of vertices $X$ and $Y$ are $d$-separated by a third (also disjoint) subset $S$ if every path between nodes in $X$ and $Y$ is blocked by $S$. The set of all nodes that d-separates $X$ from $Y$ is called the Markov Blanket of $Y$, denoted by $MB_Y$, see Fig.15-(B) for a graphical illustration.

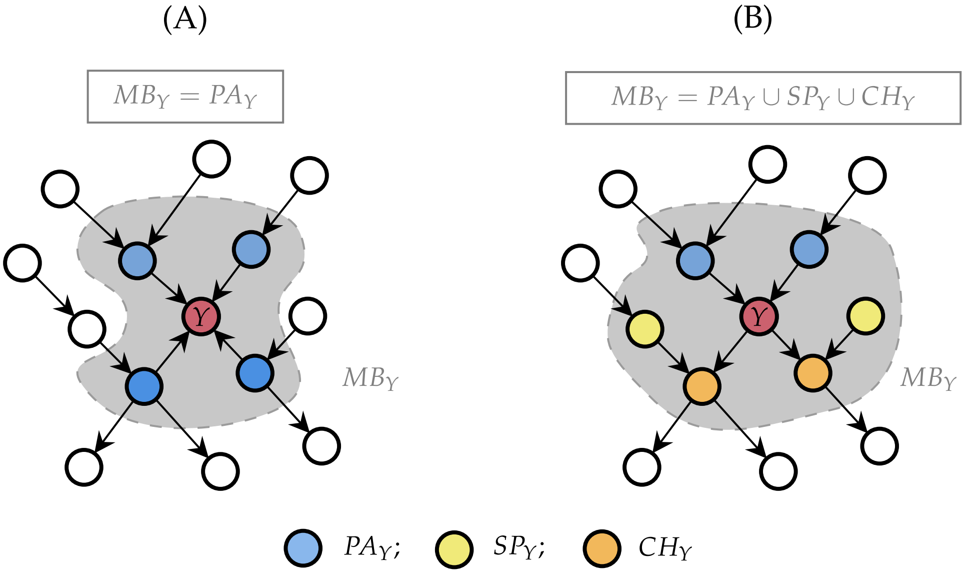


Figure 15 – Consider any set of (or single) nodes $X, Y, Z$ in $\mathbf{X}$ denoted by white, red, and colored nodes, respectively, in the figures above. Given a path $X - -Y$, the node $X$ is d-separated from $Y$ given $Z$, denoted by $X \perp\!\!\!\perp_G Y|Z$ if $Z \in MB_Y$. Markov Blanket, $MB_Y$, for nodes: (A) time-ordered and (B) not time-ordered. Source: By the author.

It turns out that as long as the joint distribution has a density and causal sufficiency[3] is satisfied, $CMC_{\text{Factorization}}$, eq.(2.14), and $CMC_{\text{d-separation}}$, eq.(2.15), are equivalent. (40) In that case, we will denote both conditions without the subscription, $CMC$.

Of interest to us will be dynamical causal models where a particular state of the causal model corresponds to a system evolving over consecutive $\tau + 1$ time steps for a discrete dynamical system of interacting elements. For this we will define (7, 21):

### Definition 2.2 (Dynamical causal models)

A dynamical causal model $\mathcal{M}_t$ defines a partition of its nodes $X \in \mathbf{X}$ into $k$ temporally ordered steps, $X = \{X_{t-1}, X_{t-2}, \ldots, X_{t-k}\}$, where $\text{PA}(X_{t-k}) = \emptyset$ and the parents of each successive step are fully contained within the $k$ previous step, $\text{PA}(X_t) \subseteq X_{t-\tau}$, $\tau = \{1, \ldots, k\}$. Because time is explicit in $G$ and we assume that there is no instantaneous causation[a], then the earlier variables, $X_{t-\tau}$, influence the later variables, $X_t$. Together, these assumptions imply that,

$$P(\mathbf{X}_t|\mathbf{X}_{t-\tau}) = \prod_i P(X_t^{(i)}|X_{t-\tau}^{(i)}) = \prod_i P(X_t^{(i)}|\text{do}(X_{t-\tau}^{(i)})), \tag{2.16}$$

i.e., nodes at time $t$ are conditionally independent given the state of the nodes at time $t - \tau$.

[a] We assume here that $\mathbf{U}$ contains all relevant background variables, any statistical dependencies between $X_{t-\tau}$ and $X_t$ are causal dependencies, and cannot be explained by latent external variables (causal sufficiency condition). This avoids "instantaneous causation" between variables meaning that $\mathcal{M}_t$ fulfills the temporal Markov property. (48)

**Remark 2.** *Note that for dynamical causal models $\mathcal{M}_t$, due to time-ordering in the variables, the Markov blanket of a target is given by its parent's set solely, see Fig.15-(A).*

### Example 2.3 (Directed Ising model)

An example of a dynamical causal model with $k = 1$ (Markov systems with one step memory) is a directed Ising model where the system of $n+1$ spins with Hamiltonians having only interactions of order 2,

$$H_2(\mathbf{X}, Y) = -Y \sum_{i=1}^{n} J_i X_i, \tag{2.17}$$

where $J_i$ are the interaction coefficients and $X_i$ is in the past of $X_{n+1} \equiv Y$.

[3] The causal sufficiency condition assumes that $\mathbf{U}$ contains all relevant background variables, any statistical dependencies among measured variables are causal dependencies, and cannot be explained by latent external variables.(7)

### 2.2.2 Faithfulness and Minimality

So far, we discussed the causal Markov conditions, which enables us to read off statistical independencies from the graph structure. The opposite direction allows us to infer graphical dependencies from statistics present in data. Such assumptions are known as Faithfulness and Minimality.

**Definition 2.3 (Faithfulness & Minimality)**

Consider the causal model $\mathcal{M} = \langle P, G \rangle$, the target $Y$ and all disjoint subsets of nodes (or single nodes) $X, Y, Z$ in the DAG $G$ with the symbols $\perp\!\!\!\perp_P$ and $\perp\!\!\!\perp_G$ standing for conditional statistical independence and d-separation, see Fig.15. Assume that the joint distribution has a probability distribution $P$. Then,
(CF) $P$ is faithful with respect to the DAG $G$ if: $(X_i \perp\!\!\!\perp_P Y|Z) \implies (X \perp\!\!\!\perp_G Y|Z)$. Faithfulness is an ad-hoc assumption claiming the opposite of the causal Markov assumption, Eq.(2.15);
(CM) $P$ satisfies causal minimality with respect to $G$ if it is Markovian with respect to $G$, but not to any proper subgraph[a] of $G$.

[a] Let $G = (V, \mathcal{E})$ be a graph with $V := (1, \ldots, n)$ and corresponding random variables $\mathbf{X} = (X_1, \ldots, X_n)$. A graph $G_1 = (V_1, \mathcal{E}_1)$ is called a subgraph of $G$ if $V_1 = V$ and $\mathcal{E}_1 \subseteq \mathcal{E}$; we then write $G_1 \leq G$. If additionally, $E_1 \subset \mathcal{E}$, then $G_1$ is a proper subgraph of $G$, $G_1 < G$.

The principles of faithfulness and minimality can be clarified using Pearl's analogy. (1) Consider the picture of the chair in Fig.16. Suppose that, we have to decide between two theories as follows:

$T_1$: The object in the picture is a chair.

$T_2$: The object in the picture is either a chair or two chairs positioned such that one hides the other.

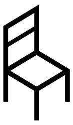


Figure 16 – Chair's analogy to clarify the faithfulness/stability and minimality as principles in causal models.
Source: By the author.

The faithfulness principle rules out $T_2$ a priori, saying that it would be quite unlikely two objects align themselves so as to have one perfectly hide the other. Such an alignment would be unstable relative to slight changes in environmental conditions. The minimality principle prefers $T_1$ over $T_2$ because the set of positions composed of single objects is a proper subset of positions composed of two or fewer objects and, unless we have evidence to the contrary, we should prefer the more specific theory (Occam's Razor).

The following formulation can be shown as equivalent to Def.2.3 and is of further help to our interests, see App.B.2 for the proof.

### Definition 2.4 (Faithfulness & Minimality)

Since both (CF') and (CM') are causal links constraints due to statistical dependence criteria, we can restate them using the mutual information quantity $I(\cdot)$:
(CF') $P$ is faithful with respect to the DAG $G$ iff $I(X_i; Y|\mathbf{Z}) \neq 0 \ \forall \mathbf{Z} \subset \mathrm{PA}_Y \notin \mathrm{DE}_Y$;
(CM') $P$ satisfies causal minimality with respect to $G$ if and only if $\forall Y$, we have that $I(X_i; Y|\mathrm{PA}_Y \setminus X_i) \neq 0$, $\forall X_i \in \mathrm{PA}_Y$.

Consider the DAG from Fig.17. By choosing $\mathbf{Z} \equiv \emptyset$ in (CF') we have that $I(X_i; Y) \neq 0$, i.e., node $X_i$ $\forall i$ presents an observable effect regardless of the information about other causal parents $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$ (49) — stable under the background parents (the reason that is also called stability, Def.2.4.1 from PEARL(1)). On the other hand, (CM') says that a distribution is minimal with respect to a causal graph if and only if there is no node that is conditionally independent of any of its parents, given the remaining parents. In some sense, all the parents are "active" (40). Suppose now, we are given a causal model, for example, in which causal minimality is violated. Then, one of the edges is "inactive". This is in conflict with the definition of (CF'), we have then

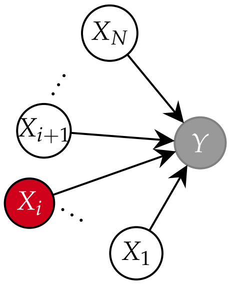


Figure 17 – DAG $G$ illustrating the causal influence from $PA_Y = \{X_{[n]}\}$ to $Y$.
Source: By the author.

### Proposition 2.1 (Faithfulness implies causal minimality)

*If $P$ is faithful and Markovian with respect to $G$, then causal minimality is satisfied.*

We get, then, the following picture of causation from Pearl's schema. The basic axioms of probability theory let us derive conditional dependencies and independencies from statistical data, and they set certain constraints on the conditional probability relations among the variables that characterize the system data. As a result, we develop graphical representations that are DAGs satisfying the global constraints — the causal Markov condition and faithfulness/minimality. By assuming that they hold, we can provide a causal interpretation of the Bayesian networks such that they are "oracles for interventions".

By postulating that local interventions are possible, Pearl assumes that the act of setting a variable to a determined value can deterministically override the causal mechanisms of the model. Thus, an intervention provides information by disrupting only the local mechanisms associated with that node. (44) It is an influence "that originates from outside the probability space" of the model. (1)

The intervention replaces the original influencing causal mechanisms (parents) with a mechanism that determines the effect variable value $X = x$ with probability 1. The graphical consequence is that arrows into the intervened variable are disrupted, and a new probability distribution is associated with the perturbed graph. Such an intervention is called an atomic intervention. Recall the joint distribution of the unperturbed graph is:

$$P(X_1, \dots, X_n) = \prod_j P(X_j|\text{PA}_j). \tag{2.18}$$

The joint distribution for a perturbed graph, for example, when $X_j$ is set to $x'_j$, is:

$$P(X_1, \dots, X_n|\text{do}(x'_j)) = \prod_{i \neq j} P(X_i|\text{PA}_i) \cdot \delta_{x_j, x'_j}. \tag{2.19}$$

The term on the left is to be read as the joint probability over all the variables $X_1$ to $X_n$, given that we set the value of $X_j$ to $x'_j$.

## 2.3 Assessing causal structure with information theory

The problem of quantifying causal influences has received considerable attention starting in communities of epidemiology and economics by means of statistical methods, those summarized in PEARL.(24) Later, a new information-theoretic (IT) paradigm, measured in units of bits, appeared in the study of complex systems (50, 51), such as neuroscience (52) and, recently, molecular biology. (53) Nonetheless, the difference in perspective between these two approaches is not clearly explained in the development of IT methodologies.

To illustrate this difference, consider using a simple example of a two-node graph $X \to Y$, where $X$ represents whether an individual has won the lottery and $Y$ represents that individual's average monthly spending. A statistical measure such as the average causal effect (ACE) would answer the question "What is the effect of winning the lottery on spending?" by comparing the average spending of lottery winners ($X = 1$) and non-winners ($X = 0$): $\mathbb{E}[Y|\text{do}(X = 1)] - \mathbb{E}[Y|\text{do}(X = 0)]$. We would expect this to be quite large. On the other hand, an IT approach would quantify the effect of $X$ on $Y$ by the mutual information (MI), $I(X; Y)$. Addressing a different question, "What is the effect of the lottery on spending?" it considers the effect of the random variable representing whether one wins the lottery on spending. However, as the odds of winning the lottery are low, an IT measure indicates that the lottery has a negligible effect on spending. Therefore, statistical measures consider the effect of a specific cause, whereas IT measures the effect at a systemic level. (51)

How to make information theory causal? Given the node $Y$ and its parent set $\mathrm{PA}_Y = \{X_j\}_i^n$ the object that plays a central role is the causal mutual information[4]:

$$I\left(\widehat{X}_j;Y\right) := \sum_{\widehat{x}_j} P\left(\widehat{x}_j\right) \sum_{y} P\left(y|\widehat{x}_j\right) \log_2 \left[\frac{P\left(y|\widehat{x}_j\right)}{\sum_{\widehat{x}'_j} P\left(\widehat{x}'_j\right) P\left(y|\widehat{x}'_j\right)}\right]. \tag{2.20}$$

This measure quantifies the number of bits of information conveyed by ‘doing’ $X$ regarding $Y$. Note that it is not a symmetric measure. Changes to $X$ providing information about $Y$, do not imply that changes to $Y$ provide information about $X$, see Fig.18-(A). The use of post-interventional distributions eliminates the effect of upstream dependencies of $X$, thereby removing any relationship between $X$ and $Y$ caused by confounding variables.

By neglecting a crucial interacting factor, such as $Z$ in Fig.18-(B), during an intervention on $X$, we may find that the cause $X$ has no causal influence. However, controlling for a relevant background variable would make $X$ causally influent. To express the independence between $X$ and $Y$ when intervening on $X$ potentially controlling some background conditions $\widehat{\mathbf{Z}} = (\widehat{X}_2, \widehat{X}_3, \ldots, \widehat{X}_n)$, we have the causal version of conditional mutual information (CMI)[5]

$$I\left(\widehat{X}_j;Y|\widehat{\mathbf{Z}}\right) := \sum_{\widehat{\mathbf{z}}} P\left(\widehat{\mathbf{z}}\right) \sum_{\widehat{x}_j} P\left(\widehat{x}_j|\widehat{\mathbf{z}}\right) \sum_{y} P\left(y|\widehat{x}_j,\widehat{\mathbf{z}}\right) \log_2 \left[\frac{P\left(y|\widehat{x}_j,\widehat{\mathbf{z}}\right)}{\sum_{\widehat{x}'_j} P\left(\widehat{x}'_j|\widehat{\mathbf{z}}\right) P\left(y|\widehat{x}'_j,\widehat{\mathbf{z}}\right)}\right], \tag{2.21}$$

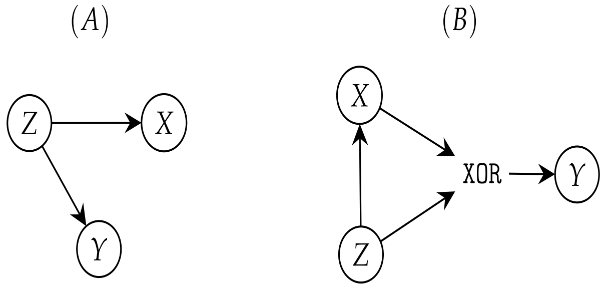


Figure 18 – (A) Intervening on $X$ gives no information about $Y$ since there is no causal link between them: $I(\widehat{X};Y) = 0$ bit. By contrast, intervening on $Z$ does give information about $Y$: $I(\widehat{Z};Y) = 1$ bit. Finally, intervening on $Y$ does not bring any information about the value of its cause $Z$: $I(\widehat{Y};Z) = 0$ bit reflecting the asymmetry of the measure. (B) Dependent causes $X$ and $Z$ with a common effect $Y$ where $Z$ takes the value $\{0,1\}$ with equal probability and $X = \mathtt{COPY}(Z)$. The law $Y := X \,\mathtt{XOR}\, Z$ means that if $X \neq Z$, then $Y = 1$, otherwise $Y = 0$. As intervening on $X$ depends on the value of $Z$, $Y$ takes values $\{0,1\}$ with equal probability resulting in $I(\widehat{X};Y) = 0$ bit. However, when we control $Z$ using an independent intervention, intervening on $X$ gives full control over the value of $Y$: $I(\widehat{X};Y|\widehat{Z}) = 1$ bit, assuming that interventions to set $X$ or $Z$ to $\{0,1\}$ are equally likely.
Source: By the author.

[4] To not overload, depending on the context, we will freely interchange the notation for interventions either as do$(\cdot)$ or as a hat on the variable in question, $\widehat{(\cdot)}$.

[5] We can identify the conventional MI/CMI as Eqs.2.20/2.21 in the observational regime.

Despite the example in Fig.18-(B), controlling for a background variable can also lead to decreasing causal CMIs, which occur in communication scenarios that use redundancy, see Example 2.4.

**Example 2.4 (Error correcting code)**

Let $E$ and $D$ be binary variables that we call encoder and decoder communicating over a channel that consists of the bits $B_1, \ldots, B_{2k+1}$ provided that $E$ is uniformly distributed. Using the simple repetition code, all $B_j$ are just copies of $E$. Then $D$ is set to the logical value that is attained by the majority of $B_j$. This way, $k$ errors can be corrected, that is, removing $k$ or less of the links $B_j \to D$ has no effect on the joint distribution.

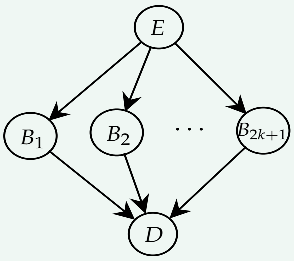


Figure 19 – Causal structure of an error-correcting scheme: the encoder generates $2k+1$ bits from a single one. The decoder decodes the $2k+1$ bit words into a single bit again.
Source: JANZING.*et al.* (54)

Consider any $B_i, B_j$ inside the majority set, then $I(\widehat{B}_i; D|\widehat{B}_j) = 0$ even though $I(\widehat{B}_{i_1}; D|\widehat{B}_{i_2}, \ldots, \widehat{B}_{i_k}) = 1$ bit and $I(\widehat{B}_{i_1}; D|\widehat{B}_{i_2}, \ldots, \widehat{B}_{i_l}) = 0$ with $l \geq k$ and $i_m$ representing any permutation of the indices $i$.

This highlights the following scenario: having full control over the parent set, we can manipulate them in order to prevent any redundant information sharing. By carefully selecting strategies for $P(I_D)$ we can, in principle, obtain accurate quantifiers of emergent phenomena. Indeed, in example 2.4 above, if we consider $P(I_D) = \{0, \ldots, 0, {}^1\!/\!(k+1), \ldots, {}^1\!/\!(k+1)\}$ with $k$ zeros we avoid the redundancy and capture the downward causation from the $k+1$ nodes on $D$. We will give more examples in Secs.2.4 and 2.5 and expand the concept of redundancy in the next chapter.

### 2.3.1 Moving beyond atomic interventions

It is important to notice that causal MIs and CMIs in Eqs.(2.20) and (2.21) are measured according to *distributions of interventions or stochastic interventions (55)*, $P(I_D) = \{P(\widehat{x}), P(\widehat{\mathbf{z}})\}$. One might accept that is natural to define $P(Y|\widehat{x})$, but not $P(\widehat{x})$. To clarify the interpretation of these objects, remember that interventions override any other causal influence on the manipulated variable and set its value with probability 1, see Eq.(2.19). Note that, the contribution of each individual value of a causal variable in Eqs.(2.20) and (2.21) of that variable is its contribution to a weighted sum.

Therefore, the probabilities of particular interventions, e.g. $P(\widehat{x})$, can be thought of as weights in an approach where measures like Eqs.(2.20), (2.21) assume the contributions of each causal value are independent of the contribution of other causal values as the examples in Fig.18. This last point should not be necessarily true since the interventional probabilities can be chosen by weighting the contribution of each causal value (or each arrow in the causal mapping) according to the better strategy for the external agent in question being then a desirable feature of these objects. By rewriting MI as a Kullback-Leibler divergence (see appendix A) in Eq.(2.20):

$$I\left(\widehat{X}_j;Y\right) \equiv \mathbb{E}_{\widehat{P}}\left[D_{KL}\left(P\left(Y|\widehat{x}_j\right)\middle|\middle|\sum_{\widehat{x}'_j} P\left(\widehat{x}'_j\right)P\left(Y|\widehat{x}'_j\right)\right)\right], \tag{2.22}$$

we can view the causal mutual information as an average over $\widehat{P} := P\left(\widehat{x}_j\right)P\left(Y|\widehat{x}_j\right)$ comparing the effect of an atomic intervention $\widehat{x}$ with a stochastic (i.e. non-atomic) intervention — a probability distribution over some applied set of them. Note that the interventional distribution fixes $\widehat{P}$, see Refs.(1, 55) for further discussions on stochastic interventions.

There are many ways to weigh the causal contributions according to different strategies. Here, to settle in stochastic interventions we list some of the options that appeared in the literature.

① **Information Flow.** Given that we can exchange intervention for observation we start with $P(I_D)$ being chosen as the empirical distribution $P_{\text{obs.}}(I_D)$. Ay & Polani (50) used this strategy where the goal was to measure causal influence 'as it flows' in a complex system. In this case[6],

$$\widehat{P} \sim P\left(x'_j\right)P\left(Y|\widehat{x}'_j\right), \quad \{P\left(\widehat{x}'_j\right)\} \sim P_{\text{obs.}}(I_D), \tag{2.23}$$

and, can interpret Eq.(2.22) as an average over the $\widehat{P}$ comparing the causal effect of $x$ on $Y$ with a counterfactual distribution wherein nature was allowed to run its course under injection of noise[7]. The information flow addresses a very clear causal question: "How much would we expect, on average, performing the atomic intervention $P(Y|\widehat{x})$ to change the course of the counterfactual distribution $\sum_{x'_j} P\left(x'_j\right)P\left(Y|\widehat{x}'_j\right)$?"

② **Causal strength.** In order to measure the strength of causal arrows, or the flow of causation on specific paths, Janzing *et al.*(54) introduced a communication scenario, where

[6] When considering more controlled variables, the information flow has to be averaged according to the probability that this flow occurs in the system.

[7] Noise injection here is an atomic intervention, meaning it breaks all causal influences directly impacting the manipulated variable and only those influences.

edges in a DAG play the role of channels that can be locally corrupted by interventions. This gives their interventional distribution as exhibited in Def.2.5, see Fig.20-(A) for an illustration.

**Definition 2.5 (Corrupting causal arrows (54))**

Let $\langle G, P\rangle$ be a causal model. Let $S \subset G$ be a set of arrows. Set $\mathrm{PA}_Y^S$ as the set of those parents $X_i$ of $Y$ for which $(\mathrm{PA}_Y^S, Y) \in S$ and $\mathrm{PA}_Y^{\bar{S}}$ those for which $(\mathrm{PA}_Y^S, Y) \notin S$. Then,

$$P_S(Y|\mathrm{pa}_Y^{\bar{S}}) := \sum_{\mathrm{pa}_Y^S} P(Y|\mathrm{pa}_Y^{\bar{S}}, \mathrm{pa}_Y^S) P_\Pi(\mathrm{pa}_Y^S), \tag{2.24}$$

where $P_\Pi(\mathrm{pa}_Y^S)$ denotes for a given $Y$ the product of marginal distributions of all variables in $\mathrm{PA}_Y^S$.

The causal strength $\mathfrak{C}_S$ is then defined as the average, over $\widehat{P} = P(\mathrm{pa}_Y)$, of the relative entropy distance between the empirical and the non-atomic intervention, Eq.(2.24),[8]

$$\mathfrak{C}_S := \mathbb{E}_{\widehat{P}}\left[D_{KL}\left(P\left(Y|\mathrm{pa}_Y\right) \middle|\middle| P_S(Y|\mathrm{pa}_Y)\right)\right]. \tag{2.25}$$

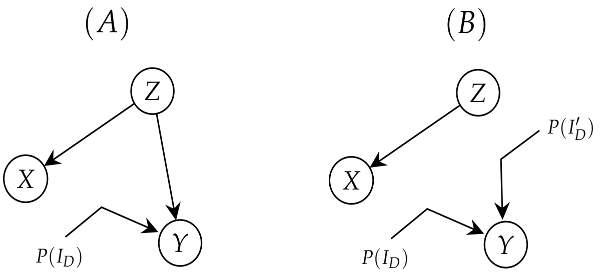


Figure 20 – (Information flow versus Causal strength) (A) Deletion of $X \to Y$. The conditional $P(Y|X,Z)$ is weighting by $P(I_D) \equiv P(X)$. The interventional distribution reads $P_S = \sum_{x'} P(y|z,x')P(x')$. (B) Deletion of $\{X \to Y, Z \to Y\}$. The conditional $P(Y|X,Z)$ is weighted with the product distribution $P(I_D)P(I'_D) \equiv P(X)P(Z)$ instead of the joint $P(X,Z)$ as in information flow since the latter would require communication between the open ends. We obtain $P_S = \sum_{x',z'} P(y|x',z')P(x')P(z')$.
Source: By the author.

Note that causal strength answers a slightly different question: "How much would we expect, on average, corrupting a subset $S$ with the strategy given by Eq.(2.25) from the parents $\mathrm{PA}_Y$ to change the empirical distribution $P(Y|\mathrm{PA}_Y)$?" Therefore, even stochastically manipulating a subset $S$ the quantifier $\mathfrak{C}_S$ has a systemic interpretation looking for the change in the whole parent's set $\mathrm{PA}_Y$.

[8] In Eq.(2.25), $P_S(Y|\mathrm{pa}_Y)$ depends solely on the reduced set of parents $\mathrm{pa}_Y^{\bar{S}} \equiv \mathrm{pa}_Y \setminus \mathrm{pa}_Y^S$ only, but for convenience of the notation we kept the formal dependence on all $\mathrm{pa}_Y$.

**Remark 3.** *Note the difference between causal strength and information flow strategies. When considering multiple sources interventions, causal strength is corrupted locally by weighting using marginal empirical distributions from the sources instead of weighting using joint empirical distributions used in information flow, see Fig.20-(B). Weighting with marginal inputs is particularly relevant for the following example: let $X$ and $Z$ be binary with $X = Z$ and $Y = X\,\texttt{XOR}\,Z$. Then, the cutting had no impact if we would keep the dependencies. Indeed, in this case $\mathfrak{C}_{X\to Y} = \mathfrak{C}_{Z\to Y} = 1$ bit and $IF_{X\to Y} = IF_{Z\to Y} = 0$ bit with $IF$ standing for information flow.*

③ **Causal specificity/Causal power/Effective information.** Being rediscovered under different names, such measures consider the probability of interventions satisfying Jayne's maximal entropy principle (56): stochastic interventions are assumed to be uncorrelated with any other non-downstream variables and are equiprobably distributed. (53) In this case,

$$\widehat{P} \sim P\left(\widehat{x}_j'\right) P\left(Y|\widehat{x}_j'\right), \quad \{P\left(\widehat{x}_j'\right)\} \sim P_{H_{\max}}(I_D), \tag{2.26}$$

since the maximum entropy distribution is maximally agnostic about the behavior of the system, we can interpret Eq.(2.22) as a measure of capacity/power of the system providing a baseline for comparing causal powers. (57) By writing the counterfactual effect distribution as

$$P(E_D) = \sum_{\widehat{x}' \in I_D} P(\widehat{x}')P(Y|\widehat{x}'), \tag{2.27}$$

we have, in a discrete finite system, the causal specificity (53), causal power (57), or effective information (19, 52) can be written in a compact way,

$$EI := I(I_D; E_D). \tag{2.28}$$

In the next section, we will make use of Eq.(2.28) to clarify in a quantitative way the philosophical path of causation taken by Woodward opening the door to talk about causation across scales.

### 2.4 The James Woodward Criteria for Interventionist Causation

Woodward's project aims to provide a clear and precise understanding of the conditions under which manipulating variables can reveal causal relationships. This notion aligns closely with Pearl's calculus of interventions. These conditions set a local constraint, meaning they only apply to a specific subset of variables, on the statistical dependencies between interventions and relevant variables in the model.

According to Woodward,

> "$X$ causes $Y$ if and only if there are background circumstances $\mathcal{B}$ such that, if some (single) intervention that changes the value of $X$ (and no other variable) were to occur in $\mathcal{B}$, then $Y$ or the probability distribution of $Y$ would change". (22)

The general idea is to find a variable $I$, which represents a way to modify the value of the cause variable $X$, and that satisfies the following conditions to be considered an intervention variable on $X$ in relation to $Y$:

**Definition 2.6 (Woodward's interventions (2))**

(I1) $I$ causes $X$;

(I2) $I$ breaks the relation between $X$ and the rest of its causes. That is, $X$ ceases to depend on the values of other variables that cause $X$ and instead depends only on the value taken by $I$;

(I3) Any directed path from $I$ to $Y$ goes through $X$. That is, $I$ is not directly or indirectly causally related to $Y$;

(I4) $I$ has an origin independent of the variables that are being investigated. In particular, $I$ is statistically independent with any causes of $Y$ that do not lie on the causal path $I - X - Y$.

Some important consequences follow as a result of these conditions that provide a more complete notion of causation according to the interventionist account and elucidate better how the approach should deal with emergent phenomena as we exhibit in the following two sections.

### 2.4.1 Invariance, stability, and modularity

Woodward's criterion of causation relies on 'change-relating' generalizations, where at least one intervention upon $X$ will produce some change in $Y$. Change-relating generalizations provide causal explanations by being invariant under interventions rather than because they hold widely in nature. Then, the first natural condition is invariance under intervention meaning simply that the relationship between variables $X$ and $Y$ continues to have when interventions are made on $X$. Secondly, there will likewise be a range of background conditions under which possible interventions on $X$ will bring about relevant changes in $Y$ while other possible interventions will not. Intuitively, $X$ is a stable cause[9] of $Y$ if it continues to cause $Y$ across some range of values of other variables in $\mathcal{B}$. Whilst

[9] Note here, the slight distinction with Pearl's definition of stability/faithfulness. Woodward talks about stability in an interventional regime while Pearl discusses stability in a statistical sense.

invariance concerns the relationship between $X$ and $Y$, stability concerns the relationship between other variables and that relationship.

Finally, for a given set of functional relations between a set of variables to correctly represent the causal facts concerning some system, the interventionist account requires that the functional relations are modular; that is, an intervention $I$ on some variable $X$ does not alter the functional relation between the effect $Y$ and any of its causes that are not on a directed path from $X$ to $Y$.(58) Modularity requires that some functional relation is invariant and stable over some range of interventions and background conditions and any other functional relations in the system remain unchanged when an intervention is carried out. (59)

The concepts of invariance, stability, and modularity are relative. A causal relationship between $X$ and $Y$ may only hold under certain ranges of possible interventions and background conditions. To represent causal relations, we must specify a level of granularity for the variables and relationships of a system such that: (i) we can intervene in the system according to the above criteria, (ii) the functional relationships between the variables are sufficiently modular, and (iii) there are appropriate ranges of invariance and stability under which the functional relationships hold. These properties may appear at finer or coarser levels of granularity. (44)

> “The choice of grain associated with the causal analysis of a situation is intimately related to the contrastive character of causal claims. As we alter the grain, we alter the potential contrastive foci that are available”. (22)

The ability for a system to be specified by causal relationships depends on the level of detail (coarse-graining) chosen. Indeed, a causal relationship may exist across a wide range of invariance, but not provide the level of precision (bijectivity) associated with the idea of specificity: “a functional relationship might be invariant and involve discrete variables but not be $1-1$ (injective) or onto (surjective)” – that is, it might fail to be bijective. (22)

How to find the optimal level of causally-relevant details, then? We will defend that the following information-based principle is a good way to formalize quantitatively how Woodward’s mapping between the cause and effect should approximate a bijection,

**Definition 2.7 (Causal explanatory power)**

The causal explanatory power of a system is a search over coarse-grainings mappings $\mathcal{M}$ for the system in a way that maximizes causal information, Eq.(2.22).

Note how the principle above remembers Eq.(1.1) from Introduction. The only change is keeping free the strategy in quantifying causal information Eq.(2.22), in other words, without using the MEP on it. Indeed, we will discuss in the next chapter how this will be important to analyze dynamical systems that cannot achieve the maximum probability distribution. In what follows, in the rest of this chapter, we will assume systems where maximum probability distribution exists and is given by the uniform one.

### 2.4.2 The specificity and proportionality in causal information

The intuitive idea behind causal information is that interventions on $X$ can be used to specify any one of a large number of values of $Y$, providing what Woodward terms “fine-grained influence” (22) over the effect variable, see Fig.21.

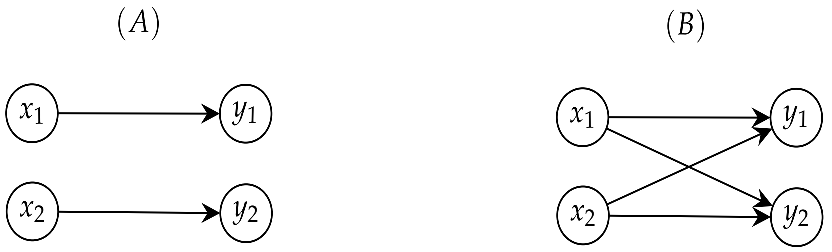


Figure 21 – Causal mappings showing a difference in bijection between causal values and effect values. Complete ignorance is given by the maximum entropy distribution: $H(Y) = -\sum_{i=1}^{2} {}^{1}\!/\!{}_{2} \log_2({}^{1}\!/\!{}_{2}) = 1$ bit. (A) After knowing $x_1$ or $x_2$, the effect is fully specified and the conditional entropy is $H(Y|\widehat{X}) = -\sum_{i=1}^{2} p(\widehat{x}_i) \sum_{j=1}^{2} p(y_i|\widehat{x}_j) \log_2 p(y_i|\widehat{x}_j) = -\sum_{i=1}^{2} {}^{1}\!/\!{}_{2} \sum_{j=1}^{2} 1 \log_2(1) = 0$ bit. The information gained by knowing the cause is $I(\widehat{X};Y) = H(Y) - H(Y|\widehat{X}) = 1$ bit. (B) Here, knowing $x_1$ or $x_2$ does not specifies the effect, $I(\widehat{X};Y) = 0$.
Source: By the author.

It is important to note that increasing the entropy of the cause variable will not necessarily lead to an increase in causal information if it does not result in additional entropy in the effect variable, as shown in Fig.22.

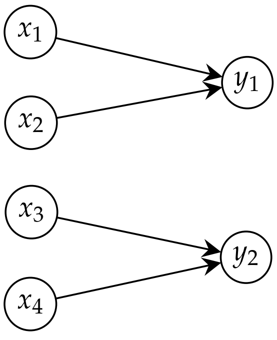


Figure 22 – Illustration of different values of the cause leading to the same outcome as in Fig.21, $H(Y) = 1$ bit. Despite two values of the cause leading to the same effect, intervening to set the value of the cause fully specifies the value of the effect as in Fig.21-(A). Then, the difference in uncertainty about the effect between before and after intervening to set the value of the cause is the same: $I(\widehat{X};Y) = H(Y) - H(Y|\widehat{X}) = 1$ bit.
Source: By the author.

Whether a system can be specified at all as being constituted by causal relations will depend upon the particular coarse-graining that is chosen. As Woodward points out, simply increasing the number of values of a cause variable is not sufficient unless these additional values are mapped onto distinct values of the effect. Such a fact implies identifying the range of invariance from the respective cause, which involves aggregating all values that make the same impact on the effect, as illustrated in Fig.23.

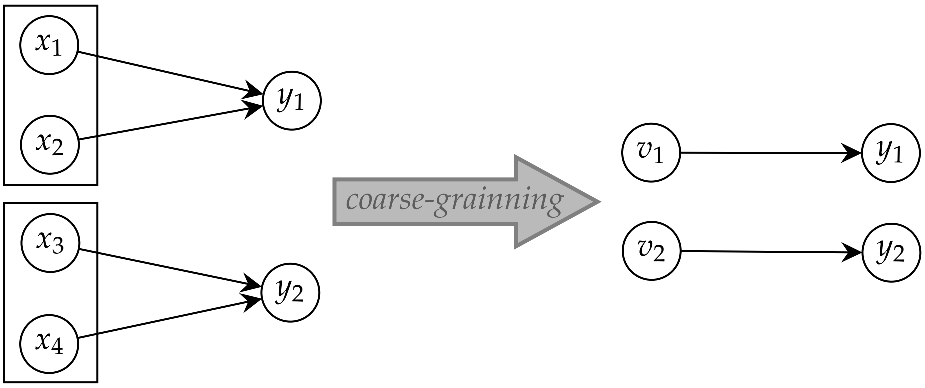


Figure 23 – The coarse-graining of the causal variable from Fig.22 reduces the entropy from 2 bits to 1 bit. Yet, it maintains specificity, $I(\widehat{V};Y) = H(Y) - H(Y|\widehat{V}) =$ 1 bit.
Source: By the author.

This brings us to the concept of proportionality: causes should be just enough for their effects, "neither omitting too much relevant detail nor containing too much irrelevant detail". (22) To ensure that irrelevant information is not included, we can minimize the entropy of the cause variable by grouping together values that have the same impact while maintaining its specificity. (60)

In the last example, it is the values of one of the variables rather than the variables themselves that are coarse-grained. Choosing how finely or coarsely to discretize one variable and choosing which variables will be coarse-grained are different tasks aiming the same goal. To illustrate, consider the `XOR` process where two sources $\{X, Z\}$ regulates $Y$ with equal probabilities $P(X) = P(Z) = 1/2$ for all values of $X$ and $Z$, see Fig.24.

The reason for the instability, Fig.24-(A), of the causal relationship between $X$ and $Y$ in the presence of a background variable $Z$ is their interdependence (synergism), here captured by $I(\widehat{X};Y|\widehat{Z})$, Fig.24-(B), as we will demonstrate in the next chapter using the PID formalism. By coarse-graining interdependent causes, Fig.24-(C), we can uncover emergent influence. As the role of the environment becomes significant we have to define properly the rules of coarse-grainings to avoid redundancy and preserve synergism in the relationships among the causes, remember example 2.4. Based on this digression, then, we justify the principle of *causal explanatory power*, Def.2.7, that operates at the whole systemic level (causes & effects).

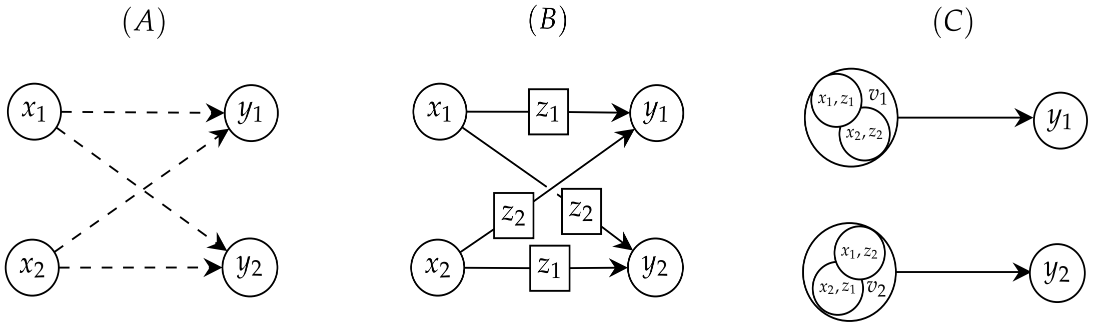


Figure 24 – Example of emergent causes $X$ and $Z$ where $Y = X$ XOR $Z$ and $X, Z \sim$ Ber(1/2). (A) If the background $Z$ is not controlled, the cause $X$ is totally non-specific. Any intervention $\widehat{x}_1$ or $\widehat{x}_2$ can equiprobably lead to $y_1$ or $y_2$ giving $I(\widehat{X}; Y) = 0$ (dashed line arrows). (B) Once we fixed the background, represented by boxes, $X$ is entirely specific: $I(\widehat{X}; Y|\widehat{Z}) = 1$ (continuous line arrows). (C) Coarse-graining variables accordingly can also give full specificity, $I(\widehat{V}; Y) = 1$.
Source: By the author.

Woodward's approach to causation begins with modular interventions and, uses constraints on statistical relationships between the intervention variable, system variables, and background variables to identify local causal connections. From there, one can determine conditional dependencies and independences across a broader scope of variables (global constraints) and then assemble the local causal relationships into DAGs.

This perspective offers a rich and fundamentally important structure to us. The level of grain used to describe the model plays a crucial role in discussing causal relationships, and regarding the description, some causal assumptions could fail as we will see in the next chapter. In particular, we will see why both descriptions (B) and (C) from Fig.24 do not remain causally optimal in more general scenarios.

## 2.5 Accessing causal structure in dynamical systems

Let us consider a device regularly recording data from a composed system of $n$ parts over time. We will consider two-time points of the evolution of the system, denoted as $t$ and $t'$, with $t < t'$ where the corresponding dynamics are encoded in the transition probability $P(\mathbf{X}_{t'}|\mathbf{X}_t)$. Consider a dynamical system with the temporal Markov property of order 1. When the interventional distribution with maximal entropy, $P_{H_{\max}}(I_D)$, is applied at the time step $t$, the distribution of states transitioned into at $t' = t+1$, $P(E_D)$, is given by Bayes' rule which we will denote as before $\mathbf{Y} \equiv \mathbf{X}_{t'}$.

Note that a transition probability matrix (TPM) can be obtained for a system with specified elements and mechanisms provided that the probability distributions are guaranteed. In this section and the next chapter, we will consider discrete dynamical systems where the maximal entropic distribution exists: such as systems of interconnected logical gates, Markov chains defined as state transitions, or even Ising models. A virtue of

$EI$ is that it can quantify the causal architecture of the TPM in terms of meaningful components of the dynamical system telling how effectively (deterministically and uniquely) causes produce effects in the system, and how selectively causes can be identified from effects.

### 2.5.1 The deterministic and degeneracy coefficients

Effective information (EI)/specificity can be rewritten as

$$\text{EI} = \text{determinism} - \text{degeneracy}. \tag{2.29}$$

In this perspective, determinism is defined as the absence of randomness or noise in causal relationships. It is determined by the level of certainty in state transitions that the information gained provides:

$$\text{determinism} := \log_2(n) - \mathbb{E}_{\widehat{x}' \in I_D}\big[H(Y|\widehat{x})\big] \tag{2.30}$$

where, the expression inside the average in the second term,

$$H(Y|\widehat{x}) = -\sum_{y \in \mathcal{Y}} P(y|\widehat{x}) \log_2 P(y|\widehat{x}), \tag{2.31}$$

is zero when a cause has only one possible effect, $P(y|\widehat{x}) = 1$, and maximal, $\log_2 n$, when a cause can lead to any of $n$ possible effects randomly. The determinism of a system is defined as the degree to which a cause has a specific effect, see Fig.25 (shaded red region).

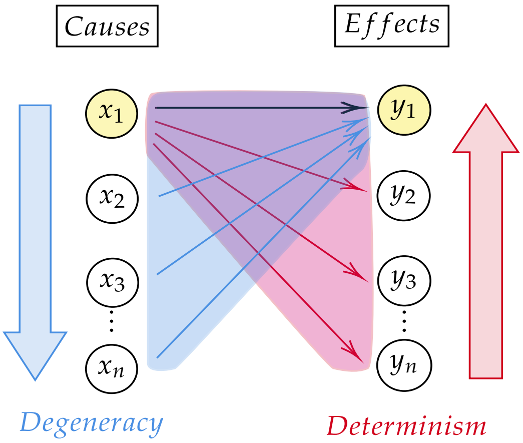


Figure 25 – The relationship between a set of causes $\{x_i\}$ to a set of effects $\{y_i\}$ can be evaluated in terms of determinism and degeneracy. Causes and effects are assumed to be temporally ordered. High determinism refers to how certain the cause $x$ is to bring about the effect $y$ (decrease of the red region). In contrast, high degeneracy looks at whether multiple causes can lead to the same effect (increase of the blue region).
Source: By the author.

Soft determinism is an indication of randomness or noise in the system. The difference between the maximum possible entropy and the actual entropy observed in the system gives a measure of how well we can predict the future of $Y$ compared to the worst-case scenario[10].

The degeneracy coefficient of a system measures the degree to which certain effects are more likely to occur than others. To define it, consider the entropic term,

$$H(E_D) = \sum_{\widehat{x},y} P(\widehat{x})P(y|\widehat{x}) \log_2 \left[ \sum_{\widehat{x}'} P(\widehat{x}')P(y|\widehat{x}') \right], \tag{2.32}$$

which is zero when the probability of all effects are the same, regardless of the causes and maximal when certain effects are favored because they are caused by a greater number of causes, making those causes less essential, see Fig.25 (region blue). Then, the degeneracy coefficient for the system is defined as:

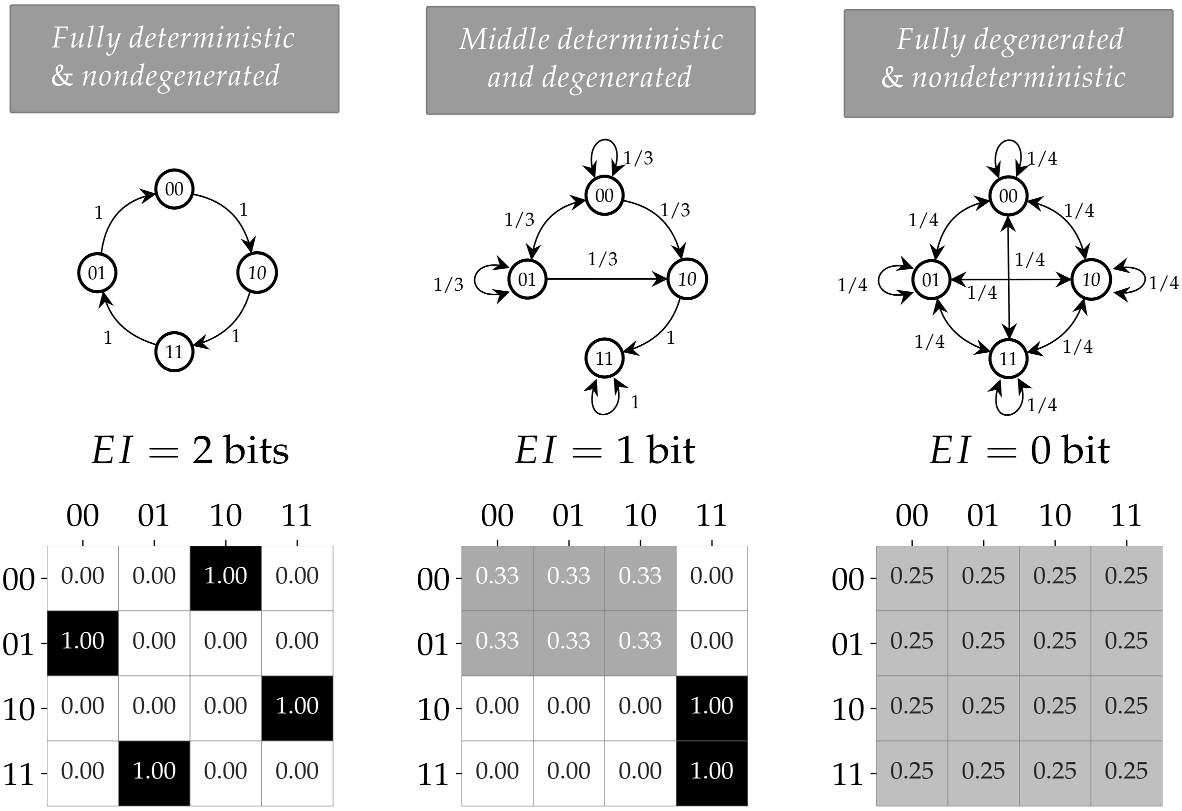
$$\text{degeneracy} := \log_2(n) - H(E_D). \tag{2.33}$$

Degeneracy can be understood as the amount of information about the past that is lost when multiple causes lead to the same effect. High degeneracy is an indication of attractor dynamics. (20) To demonstrate the deterministic and degeneracy coefficients' ability to accurately quantify causal structure, consider the TPMs ($t$ by $t+1$) of three Markov chains, each with $n = 4$ states $[00, 01, 10, 11]$, see Fig.26.

From left to right in Fig.26 we have differences in $EI$ due to decreasing determinism and increasing degeneracy in the systems, respectively, with a value that ranges between 0 and 1. In the first model, every state ultimately determines both the past and future, while the states of the second one only constrain the past and future to some degree. The last model is not constrained at all, and the probability of any state-to-state transition is $1/n$. This affects the effective information values of the three models: $EI(1) = 2$ bits, $EI(2) = 1$ bit, and $EI(3) = 0$ bits.

Note how determinism and degeneracy explain how causal specificity or $EI$ approximates Woodward's bijection concept for a good mapping describing the cause and effect. The best scenario of explanation is when determinism is maximal and degeneracy is minimal which corresponds to a bijective mapping, see yellow nodes in Fig.25. Also, nonzero degeneracy implies in failure of an injective mapping. Despite that determinism can be related to the question of "how surjective the map is". Therefore, to achieve Woodward's criteria for an optimal description of a causal model – bijection – we should search for a description where we maximize determinism while minimizing degeneracy. Below, we show how coarse-grainings analysis across scales in dynamic systems answers this question.

[10] Note that, determinism considers the transitions of a cause to the effect set and not a specific transition, although it is possible to calculate the contribution of each transition.


Fully deterministic & nondegenerated
Middle deterministic and degenerated
Fully degenerated & nondeterministic
00
01
10
11
1
1/3
1/4
EI = 2 bits
EI = 1 bit
EI = 0 bit

|  | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 01 | 1.00 | 0.00 | 0.00 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 1.00 |
| 11 | 0.00 | 1.00 | 0.00 | 0.00 |

|  | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 00 | 0.33 | 0.33 | 0.33 | 0.00 |
| 01 | 0.33 | 0.33 | 0.33 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 1.00 |
| 11 | 0.00 | 0.00 | 0.00 | 1.00 |

|  | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 00 | 0.25 | 0.25 | 0.25 | 0.25 |
| 01 | 0.25 | 0.25 | 0.25 | 0.25 |
| 10 | 0.25 | 0.25 | 0.25 | 0.25 |
| 11 | 0.25 | 0.25 | 0.25 | 0.25 |

Figure 26 – Markov chains with different levels of deterministic and degeneracy coefficients with their respective TPMs, representing probabilities in gray-scale.
Source: By the author.

### 2.5.2 Causal analysis across scales

A full micro causal model represents a system with the highest level of detail in space and time for all elements and states, represented as micro $= \{\mathbf{X}\}$. However, according to Woodward's framework, a causal model can be modeled at different levels of detail or focus on different subsets. These levels are generated by features macro $= \{V\}$ calculated through $P(V_t|\mathbf{X}_t)$, which are based on the underlying system satisfying, then, the supervenience condition: once the base level of detail is established, all higher-level models are determined. This relationship can be expressed through statistical independence between $V_t$ and $\mathbf{X}_{t'}$ when $\mathbf{X}_t$ is given, forming a Markov chain. (9) This includes both deterministic processes where $V_t = \mathcal{M}(\mathbf{X}_t)$ as well as coarse-grainings affected by observational noise.

By evaluating the effective information $EI$ overall levels of the coarseness of $\mathbf{X}$, one can identify the scale at which the model becomes more causally specific. This proposes a method to identify the scale that contains the most explanation of the causal relationships between the objects, measured in bits,

$$EI(\text{macro}) - EI(\text{micro}). \tag{2.34}$$

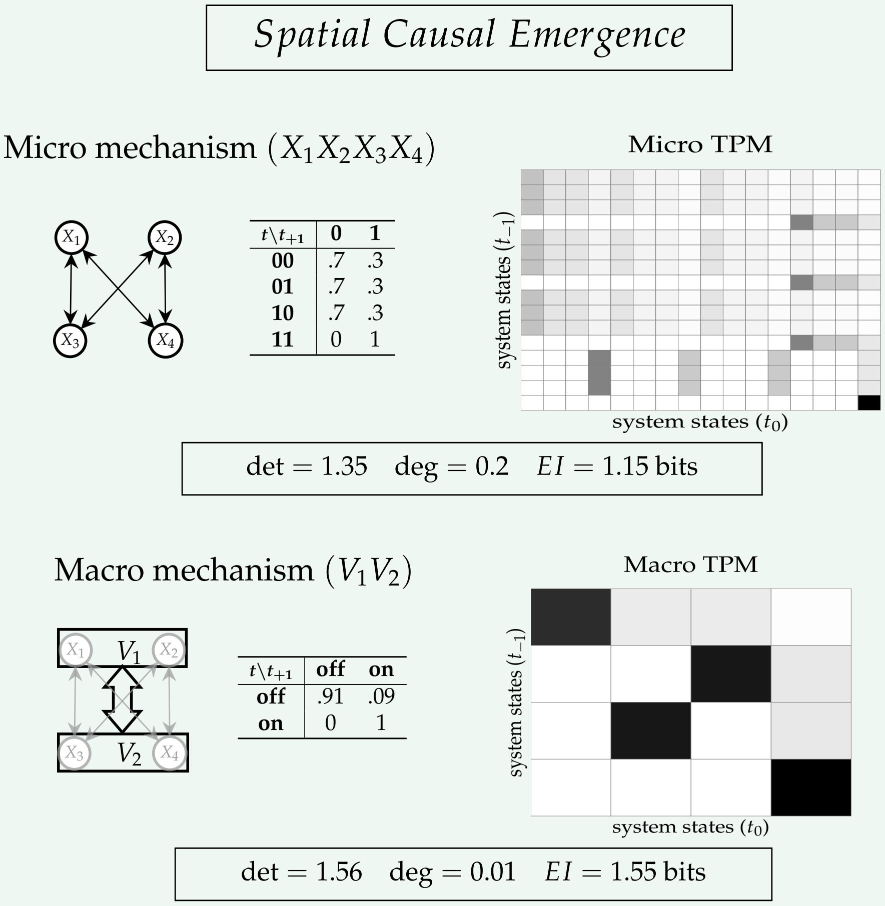
To quantify $EI(\text{macro})$, we have to deal with macro interventions. A possible coarse-grained intervention strategy that we can choose is:

$$\text{do}\Big(V = v_j\Big) := \frac{1}{n} \sum_{x_i \in v_j} \text{do}\big(X = x_i\big). \tag{2.35}$$

where $n$ is the number of microstates $x_i$ mapped into $V$. In other words, a uniform average over a set of micro-interventions. In Hoel's analysis, causal emergence is claimed when the macro-level $EI$ beats the micro-level (19), i.e., when Eq.(2.34) is positive. See example 2.5 for an illustration.

### Example 2.5 (Interconnected AND gates (19))

Consider the system of four binary elements $\mathbf{X} = \{X_1X_2X_3X_4\}$ in Fig.27-(top-left). Each micro mechanism is an AND-gate (two inputs) over some intrinsic noise. The $16 \times 16$ TPM was constructed by setting the system into all possible microstates from [0000] to [1111] with equal probability, Fig. 27-(top-right). At the micro level $\mathbf{X}$, effective information $EI(\mathbf{X}) = 1.15$ bits, out of maximally 4 bits.



| $t\backslash t_{+1}$ | 0 | 1 |
|---|---|---|
| 00 | .7 | .3 |
| 01 | .7 | .3 |
| 10 | .7 | .3 |
| 11 | 0 | 1 |

det = 1.35 deg = 0.2 $EI$ = 1.15 bits

| $t\backslash t_{+1}$ | off | on |
|---|---|---|
| off | .91 | .09 |
| on | 0 | 1 |

det = 1.56 deg = 0.01 $EI$ = 1.55 bits

Figure 27 – Spatial causal emergence (counteracting indeterminism). (top-left) The fine-graining $\mathbf{X}$ of the system is composed of identical noisy micro-mechanisms. (top-right) The micro TPM. (bottom-left) The coarse-graining $\mathbf{V}$ and its macro mechanism. (bottom-right) The macro TPM. By raising determinism and reducing degeneracy, the macro beats the micro: $EI(\text{macro}) - EI(\text{micro}) = 0.40$ bits.
Source: Adapted from HOEL; ALBANTAKIS; TONONI. (19)

The macro mechanism Fig.27-(bottom-left), composed of two elements $\{V_1, V_2\}$, each with states $\{\texttt{on}, \texttt{off}\}$, is a coarse-graining of $\mathbf{X}$ as defined by the mapping $\mathcal{M}$ in Fig.28 for $\{X_1, X_2\}$ to $V_1$, ($\{X_3, X_4\}$ to $V_2$ is symmetric).

The $4 \times 4$ TPM in Fig.27-(bottom-right) was obtained by setting the system into all possible macro states from $[\texttt{off}, \texttt{off}]$ to $[\texttt{on}, \texttt{on}]$ with equal probability. In the macro level, $EI(\mathbf{V}) = 1.55$ bits, higher than $EI(\mathbf{X}) = 1.15$ bits demonstrating that in this case, the macro constitutes the optimal causal model of the system. In this example, the gain in $EI$ at the macro level comes primarily (91%) from counteracting noise, $\det(\mathbf{X})/\log_2(16) = 0.34$ and $\det(\mathbf{V})/\log_2(4) = 0.78$; and less so (9%) from reducing degeneracy, $\deg(\mathbf{X})/\log_2(16) = 0.05$ and $\deg(\mathbf{V})/\log_2(4) = 0.006$.

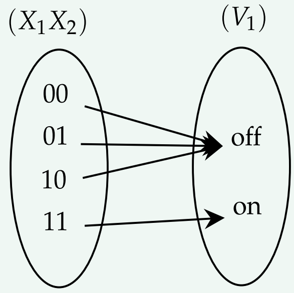


Figure 28 – Mapping $\mathcal{M}$.
Source: HOEL; ALBANTAKIS; TONONI. (19)

As we will discuss in the next chapter, we will reserve the term emergence for processes where there exists informational synergism according to the PID formalism. Then, we will argue based on the results of the next chapter that $EI$'s increase is not solely due to capturing synergism but to noise reduction manifested by redundancy in the systems. Indeed, the latter seems more crucial to raise the causal explanatory power for the model in question.

# 3 HIGH-ORDER INTERTENDENPENCIES BY THE CAUSAL LENS

*Each in his own opinion*
*Exceeding stiff and strong,*
*Though each was partly in the right,*
*And all were in the wrong!"*

*John Godfrey Saxe's in "The Blind Men and the Elephant"*

In Pearl's framework, a crucial task is discovering the DAGs structure to implement manipulative techniques and uncover causality. This can be achieved by causal discovery algorithms (CDTs) (1, 43) and their variants incorporating dynamic causal models.(7, 49, 61–66). To infer causal links from multivariate time-dependent data these algorithms rely on conditional mutual information based on the concept of causal faithfulness/stability, Def.2.4. (40,48,67) As already pointed out by James *et al.* (6), but not taken forward to the causal concept, the mutual information analysis can be blinded to high-order interactions. As we will prove, for synergic dependencies, causal faithfulness can be violated provoking the failure of a huge class of CDTs.

Remember that, from Woodward's path to capturing causal influence, the drawback of using mutual information (MI) for evaluating systems with more than two variables leads to a limited understanding of the distribution of information among dependent causes, needing to incorporate the background context $\mathcal{B}$ to capture emergence by using the conditional operation in MIs. In some scenarios, we saw the appearance of redundant causes in $\mathcal{B}$ showing that the simple task of conditioning could fail according to the growth of $\mathcal{B}$. By applying coarse-grainings we argued that the size of $\mathcal{B}$ can be reduced while not losing the capture of emergent properties as seen by using the EI approach. However, when the coarse-graining mapping $\mathcal{M}$ is unknown, it is unclear how we should operate, calling, then, for new informational techniques to investigate these multivariate scenarios.

The attempt to establish foundations for a multivariate informational analysis comes back to information theory originator Claude Shannon (68), who made use of Garrett Birkhoff's lattice theory (69, 70), a subclass of order theory[1]. Recently, these ideas have been refined, and one particular formalism is gaining strength in the scientific community, the *Partial Information Decomposition* (PID). (30) The multivariate analysis

[1] Order theory is a branch which investigates the intuitive notion of order using binary relations.

can be complex, in the sense that, one interest could be the informational contribution from one source to many targets; many sources to one target; or, from many sources to many targets, see Fig.29.

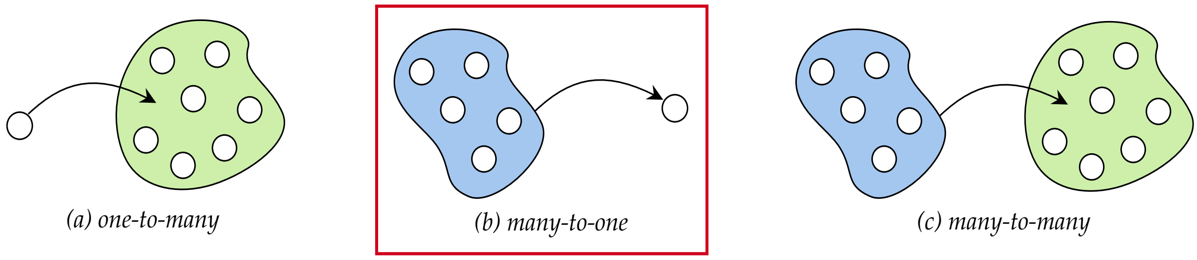


Figure 29 – In the terminology of causation, scenarios (a) and (c) are examples of causal decoupling and (b) downward causation, remember Fig.2.
Source: By the author.

In this chapter, motivated by the problem of identifying dependent causes, we will center on the (b) many-to-one (downward causation) scenario. We will assume no confounders among the parent set $\mathbf{X}$ being able to change the notation $P(Y|\mathbf{X})$ and $P(Y|\text{do}(\mathbf{X}))$ freely when necessary. We will discuss briefly in Sec.3.4 how to operate when this is not the case. We use simple simulated systems, including higher-order interactions ones, to show quantitatively that genuine causal synergism violates faithfulness. We do so by claiming the importance of the conditioning operation in mutual information to capture pure synergism by formalizing such a concept using partial information decomposition theory (PID). (30) Then, we connect this with the causal concepts of faithfulness and minimality. (1)

By comparing different structural organizations of non-pairwise systems we also show that causal faithfulness is recovered when redundancy dominates the system in question. Such a phenomenon manifests when the conditional mutual information (CMI) starts to fail, raising a trade-off between faithfulness and minimality in terms of levels of redundancy and synergy. These results clarify a long-standing discussion about the regime of the faithfulness condition for the traditional CDTs in high-order scenarios proposing a review of them (7, 49, 61–66, 71) when dealing with high-order interdependencies. All the content of this chapter is new, which can be found in MARTINELLI *et al.* (72), except Sec.3.4 which is part of an ongoing work.

Finally, in Sec.3.4, we argue how coarse-grainings should be done according to the informational mapping constrained by PID: it preservers the synergic set while reducing the redundant set. We give a "proof-of-principle" example showing how the $EI$ quantifier remains invariant using that reasoning. This gives an informational principle to operate on higher scales in order to elucidate emergent phenomena and eliminate noise raised by unnecessary redundancies. We dedicate the rest of this section to discussing how constraints in the PID quantifiers can be seen as strategies of non-atomic interventions.

## 3.1 Partial Information Decomposition

The relationship between parts and wholes is a fundamental aspect of nature, present at all levels of space and time, from atoms in molecules to planets in solar systems. It transcends just scientific fields and is commonly seen in everyday languages, such as referring to a president as part of the government or a slice of cake as part of the whole cake. Its widespread presence makes it an intuitive concept to understand.

Asking how to decompose joint mutual information into its components is similar to asking “how to slice a cake?” There are multiple ways to do so and therefore, no single answer to the question. To clarify the question, a criterion must be established for how the joint mutual information should be decomposed. The work of Williams & Beer (30) was to realize that such combinations of information should satisfy a parthood distribution (73) $f : 2^{[n]} \to \{0, 1\}$ s.t.,

1. $f(\{\emptyset\}) = 0$ (There is no information in the empty set);
2. $f(\{1, \dots, n\}) = 1$ (All information is in the full set);
3. For any two collections of source indices $a, b : a \subseteq b$, then $f(a) = 1 \implies f(b) = 1$;

using conditions above it means that the number of atoms is equal to the number of monotonic Boolean functions minus two. Such a sequence is a very famous sequence in combinatorics called the Dedekind numbers which counts the number of antichains in a distributive lattice (70), see Def.3.1.

The fact that antichains form a lattice (70) gives a natural hierarchical structure of partial order over the elements of $\mathcal{A}(\mathbf{X})$,

$$\forall \alpha, \beta \in \mathcal{A}(\mathbf{X}),\ \alpha \preceq \beta \iff (\forall B \in \beta :\ \exists A \in \alpha,\ A \subseteq B). \tag{3.1}$$

**Definition 3.1 (The lattice of antichain)**

Consider the sources $\mathbf{X} = \{X_{[n]}\}$. The set

$$\mathcal{A}(\mathbf{X}) = \{\alpha \in \mathcal{P}^{+}(\mathcal{P}^{+}(\mathbf{X})) : A_1 \not\subset A_2,\ \forall A_1, A_2 \in \alpha\}, \tag{3.2}$$

is called the lattice of antichains of the sources $\mathbf{X}$[a] where $\mathcal{P}^{+}(S) = \mathcal{P}(S) \setminus \{\emptyset\}$ denotes the set of nonempty subsets of $S$.

[a] Henceforth, we will denote sets of $\mathcal{A}(\mathbf{X})$, corresponding to collections of sources, omitting the brackets with a dot separating the sets within an antichain, and the groups of sources are represented by their variables with respective indices concatenated. For example, $X_1 \cdot X_2X_3$ represents the antichain $\{\{X_1\}\{X_2, X_3\}\}$, see Fig.30.

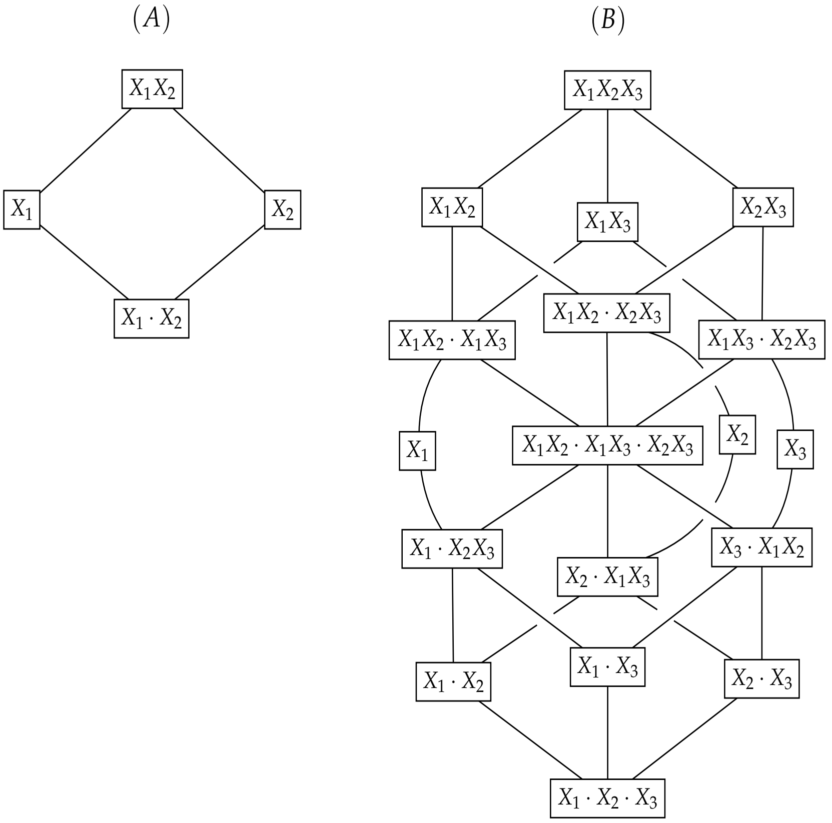


Figure 30 – Illustration of antichains $\mathcal{A}(\mathbf{X})$ for: (A) $\mathbf{X} = \{X_1, X_2\}$ and (B) $\mathbf{X} = \{X_1, X_2, X_3\}$.
Source: By the author.

The cardinality of $\mathcal{A}(\mathbf{X})$ for $\mathbf{X} = n$ is given by the $(n-2)$-th Dedekind number, which for $n = 2, 3, 4, \ldots$ is $1, 4, 18, 166, 7579, \ldots$ being super-exponential according to $|\mathbf{X}|$.

To attribute an informational character to $\mathcal{A}(\mathbf{X})$, one then assigns a quantity of shared information to each element and defined by the mapping: $\alpha \mapsto I_\cap(\alpha; Y)$[2]. This should quantify the amount of information shared by each set of sources within an antichain $\alpha$ about the target, see Fig.31. Using the parthood distribution conditions Williams & Beer (30) proposed the following axioms that such a measure should follow[3]

(S) (symmetry) $I_\cap(\alpha; Y)$ is unchanged under permutations of $\alpha$;

(SR) (self-redundancy) $I_\cap(\{\alpha\}; Y) = I(\{\alpha\}; Y)$, where $\{\alpha\}$ is a singleton;

(M) (monotonicity) $I_\cap(\alpha_1; Y) \leq I_\cap(\alpha_2; Y), \forall \alpha_1 \preceq \alpha_2$.

[2] The $\cap$ symbol refers to the idea that the redundant information of collections $\{A_1, A_2, \ldots, A_m\}$ captures intersecting information contained in $\{A_1, \text{and } A_2, \text{and } \ldots, \text{and } A_m\}$.

[3] Supported by the idea that any measure of redundancy as intersecting information should satisfy the same basic properties of set intersection; namely, commutative (symmetric): $X \cap Y = Y \cap X$; idempotent (self-intersection): $X \cap X = X$; and monotonic: $(X_1 \cap \ldots \cap X_{k-1} \cap X_k) \subseteq (X_1 \cap \ldots \cap X_{k-1})$ with equality if $X_{k-1} \subseteq X_k$.

When ascending the lattice, the redundancy function $I_\cap(\alpha;Y)$, monotonically increases, being a cumulative measure of information where a higher element provides at least as much information as a lower one. (30) The inverse of $I_\cap(\alpha;Y)$ called the partial information functions (PI-functions) and denoted by $I_\partial$ measures the partial information contributed uniquely by each particular element of $\mathcal{A}(\mathbf{X})$. This partial information will form the atoms into which we decompose the total information that $\mathbf{X}$ provides about $Y$. For a collection of sources $\alpha \in \mathcal{A}(\mathbf{X})$, the PI-functions are defined implicitly by the Möbius inverse of $I_\cap$,

$$I_\partial(\alpha;Y) = I_\cap(\alpha;Y) - \sum_{\beta \prec \alpha} I_\partial(\beta;Y). \tag{3.3}$$

Decomposing the MI into PI's and applying this to the XOR process discussed above, with $\mathbf{X} = \{X_1, X_2\}$ and a single target variable $Y$ we have that,

$$\begin{aligned} I(X_1;Y) &= I_\partial(X_1;Y) + I_\partial(X_1 \cdot X_2;Y), && (3.4)\\ I(X_2;Y) &= I_\partial(X_2;Y) + I_\partial(X_1 \cdot X_2;Y), && (3.5)\\ I(X_1,X_2;Y) &= I_\partial(X_1;Y) + I_\partial(X_2;Y) + I_\partial(X_1 \cdot X_2;Y) + I_\partial(X_1X_2;Y) && (3.6) \end{aligned}$$

From the equations above we can see that the causal link in `XOR` process discussed above from $\{X_1, X_2\}$ to $Y$ is due to the synergic term $I_\partial(X_1X_2)$.

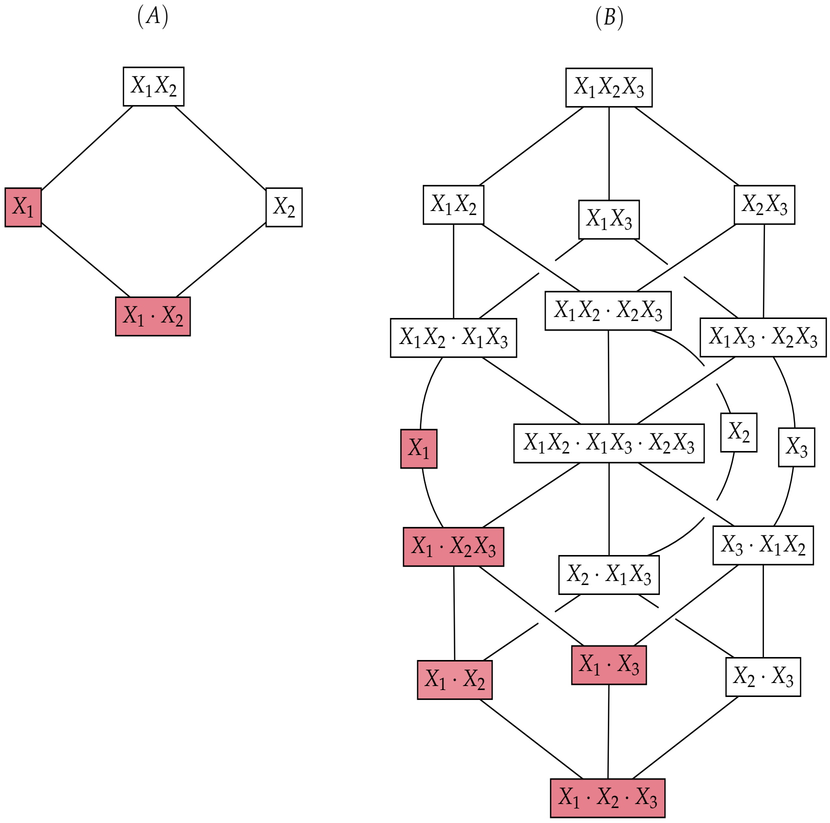


Figure 31 – (red boxes) Cumulative partial atoms in $I_\cap(\{\{X_1\}\},Y)$ for (A) $\mathbf{X} = \{X_1, X_2\}$ and (B) $\mathbf{X} = \{X_1, X_2, X_3\}$.
Source: By the author.

### 3.1.1 Isolating partial atoms

While the PID framework provides the structure on which information can be decomposed, it fails to provide the specific keystone necessary to calculate it: the $I_\cap$ that forms the base of the PI lattice. Indeed, Eqs.3.4, 3.5, and 3.6 form an under-determined system of three equations with four unknown quantities. Given an appropriate function with which to compute any of these three, the rest are trivial. Williams & Beer proposed the specific information as a plausible function, denoted as $I_{WB}$:

$$I_{WB}(\{A_1, A_2, \ldots, A_k\}; Y) := \sum_{y \in \mathcal{Y}} P(Y) \min\{I(A_1; y), I(A_2; y), \ldots, I(A_k; y)\}. \tag{3.7}$$

The term $\min_{A_i \in \alpha} I(A_i; y)$ calculates the minimal amount of information any $A_i \in \alpha \equiv \{A_1, A_2, \ldots, A_k\}$ provides about the specific state $Y = y$. Across all $y \in \mathcal{Y}$, $I_{WB}$ quantifies the expected minimum amount of information that the atom $\alpha$ will unfold about $Y$, see example 3.1 for an illustration.

**Example 3.1 (Quantifying redundancy – disturbed XOR)**

Consider the XOR operation discussed previously with some noise in source $X_2 \sim \text{Ber}(p = 3/4)$ giving the output:

| $X_1$ | $X_2$ | $Y$ | $P$ |
|---|---|---|---|
| 0 | 0 | 0 | 3/8 |
| 1 | 0 | 1 | 1/8 |
| 0 | 1 | 1 | 3/8 |
| 1 | 1 | 0 | 1/8 |

Using Eq.3.7 in Eqs.3.4, 3.5, and 3.6 we have that the disturbed XOR produces[a]: $I_\partial(X_1X_2; Y) = 0.811$ bits of synergy, and $I_\partial(X_1; Y) = 0.188$ of unique information from $X_1$.

[a] We used the dit (Discrete Information Theory) package (74) for the computation of the partial atoms.

**Remark 4.** *Even after extensive work, there is no general agreement with a particular measure. (75, 76) In what follows our results are flexible meaning that they do not depend on a specific quantifier for partial atoms. Instead, we focus in to isolate groups of atoms with common properties. This means that any quantifier that meets these properties can be used.*

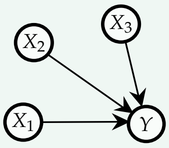
Note that MI obviously satisfies the PID three axioms; it is well known that CMI does not satisfy monotonicity (77) being difficult to isolate specific atoms in the informational lattice. However, as we will show, CMI monotonically increase (decrease) high-order synergic (redundant) terms a key property, to explain faithfulness violation in the presence of synergic properties.

## 3.2 Emergent causes from partial atoms

The existence of the causal link in the `XOR` process, even though $I(X_i;Y) = 0$ for $i = 1, 2$, is known as a violation of the causal faithfulness condition, see Def.2.4. Note, however, that the concept of causal minimality is still satisfied since conditional independence plays a role here. Also, if $P(X_1) \neq P(X_2)$ as Example 3.1 above, faithfulness is not violated anymore raising the common claim that its violations are rather pathological. (7, 49, 66) Below, we consider a simple causal process showing that when studying synergism, faithfulness violations are not rare corner cases, but can be prevalent in the space of probability distributions.

### Example 3.2 (Failure of faithfulness)

Consider $\mathbf{X} = \{X_i\}_{i=1}^{3}$ as three independent binary sources and the target $Y$ being the logical `OR` process between $X_1$ `XOR` $X_2$ and $X_3$ with probabilities distribution varying according to $0 \leq f \leq 1$ given by the table below:



$Y = (X_1$ `XOR` $X_2)$ `OR` $X_3$

| $X_1$ | $X_2$ | $X_3$ | $Y$ | *Prob.* |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | $f/8$ |
| 1 | 0 | 0 | 1 | $f/8$ |
| 0 | 1 | 0 | 1 | $3f/8$ |
| 1 | 1 | 0 | 0 | $3f/8$ |
| 0 | 0 | 1 | 1 | $(1-f)/8$ |
| 1 | 0 | 1 | 1 | $(1-f)/8$ |
| 0 | 1 | 1 | 1 | $3(1-f)/8$ |
| 1 | 1 | 1 | 1 | $3(1-f)/8$ |

By computing the MIs and CMI × $f$, see figure below, we can see that $I(X_2;Y) = 0$, but $I(X_2;Y|X_1,X_3) \neq 0$ , $\forall f > 0$, showing a simple system where faithfulness is violated but minimality is not, in a non-pathological way.

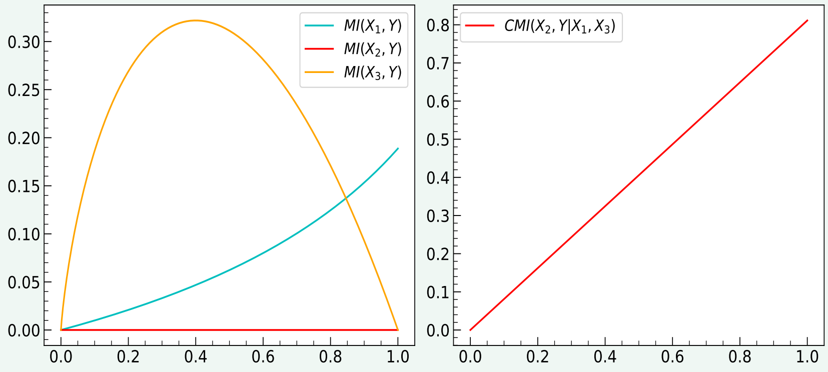


Figure 32 – (left) $I(X_i)$ for $i = 1, 2, 3$ and (right) $I(X_2|X_1,X_3)$ varying $0 \leq f \leq 1$.
Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

In Example 3.2 the robustness of minimality inside the context of the synergism seems to be related to the need to consider conditioned independencies. This point motivates us to view the conditioning operation under the PID eyes. The natural question, therefore, raises: Is there a subset $\mathcal{S}(\mathbf{X}) \subset \mathcal{A}(\mathbf{X})$, in which its elements form a lattice such that one can isolate only synergic nodes in $\mathcal{S}$?

To do so, we will look deeper into the PID approach in the search for $\mathcal{S}(\mathbf{X})$. To start, we propose Def.3.2 which gives a formal definition for $\mathcal{S}(\mathbf{X})$ allowing us to connect it with the conditioning operation, Prop.3.1 with proof in App.C.

**Definition 3.2 (The unique, redundant, and synergic sets)**

Let be a set of sources $\mathbf{X} = \{X_{[n]}\}$, $\mathbf{Z} \subseteq \mathbf{X}$, and the downset[a] of $\mathbf{Z}$ inside $\mathcal{A}(\mathbf{X})$, denoted by $\downarrow\!\mathbf{Z}$. We say that the subset $\mathcal{B}(\mathbf{Z}) \subset \mathcal{A}(\mathbf{X})$, for $k \in [n]$, is:

(a) **unique:** this set is represented by the singletons in $\downarrow\!\mathbf{Z}$ with size $k = 1$ (pairwise behavior). In this case, we denote $\mathcal{B}(\mathbf{Z}) \equiv \mathcal{U}(\mathbf{Z})$;

(b) **synergic of order** $k$**:** this set is represented by the singletons in $\downarrow\!\mathbf{Z}$ with size $k > 1$. Then, $\mathcal{B}(\mathbf{Z}) \equiv \mathcal{S}^{(k)}(\mathbf{Z})$;

(c) **redundant of order** $k$**:** this set is represented by all non-singletons $\beta = \{\{B_1\}, \{B_2\}, \ldots\}$ in $\downarrow\!\mathbf{Z}$ s.t. there exists at least one $B_j \in \beta$ with size $k$. Then, $\mathcal{B}(\mathbf{Z}) \equiv \mathcal{R}^{(k)}(\mathbf{Z})$;

When considered all the orders $k$ we will omit the superscript, having then $\mathcal{S}(\mathbf{X}) = \bigcup_k \mathcal{S}^{(k)}(\mathbf{X})$ and $\mathcal{R}(\mathbf{X}) = \bigcup_k \mathcal{R}^{(k)}(\mathbf{X})$, respectively. Note that $\mathcal{A}(\mathbf{X}) = \mathcal{U}(\mathbf{X}) \sqcup \mathcal{S}(\mathbf{X}) \sqcup \mathcal{R}(\mathbf{X})$. (For an illustration of these sets see Figs in App.C.1).

[a] The down set of $\alpha$, $\downarrow\!\alpha$, means that $\beta \preceq \alpha, \forall \beta \in \downarrow\!\alpha$ where $\alpha, \beta \in \mathcal{A}(\mathbf{X})$.

**Proposition 3.1 (Synergic property of the conditioning set)**

*Consider a node* $X_j \in \mathbf{X}$*,* $\mathbf{Z} \subseteq \mathbf{X}_{\setminus X_j}$ *where* $|\mathbf{Z}| = k - 1$ *for* $k \in [n]$*. The CMI* $I(X_j; Y \mid \mathbf{Z})$ *captures the unique set* $\mathcal{U}(X_j)$ *and the synergic set* $\mathcal{S}(\{X_j, \mathbf{Z}\}) = \bigcup_k \mathcal{S}^{(k)}(\{X_j, \mathbf{Z}\})$*. Furthermore, the set* $\mathcal{S}^{(k)}(\{X_j, \mathbf{Z}\})$ *increases monotonically on* $I(X_j; Y \mid \mathbf{Z})$ *according to* $|\mathbf{Z}|$*.*

To elucidate Prop.3.1 we consider a scenario where the non-pairwise relationships between $\mathbf{X}$ and $Y$ can be expressed as a Gibbs distribution (Full simulation details are reported in App.C.1). To start, we show the importance of the conditioned set size, $|\mathbf{Z}|$, to capture information contribution from non-pairwise terms. To do so, we consider a system of $n + 1$ spins, with Hamiltonians having interactions only of order $k$,

$$H_k(\mathbf{X}) = -X_{n+1} \sum_{|\boldsymbol{\alpha}|=k-1} J_{\boldsymbol{\alpha}} \prod_{i \in \boldsymbol{\alpha}} X_i, \tag{3.8}$$

where $J_{\boldsymbol{\alpha}}$ are the interaction coefficients and the sum runs over all collections of indices $\boldsymbol{\alpha} \subseteq [n]$ of cardinality $|\boldsymbol{\alpha}| = k - 1$.

For these systems, we calculate the average normalized CMI, $\overline{CMI}$, to measure the strength of the high-order statistical effects beyond pairwise interactions (Fig.33). Our results confirm that to get information from a causal influence of order $k$ we have to account for conditioned sets of proportional size. Furthermore, the pairwise interaction regime is the only case where MIs (CMIs with $|\mathbf{Z}| = \emptyset$) are nonzero showing the violation of faithfulness for non-pairwise interactions.

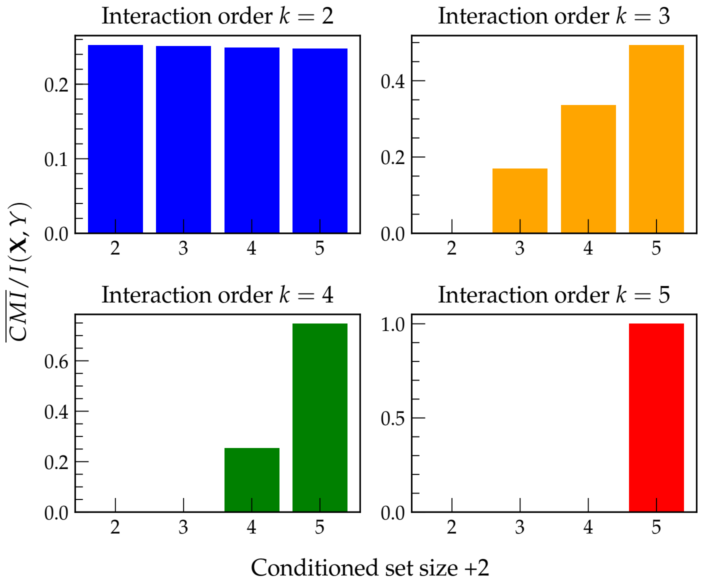


Figure 33 – Growth of the $\overline{CMI}$ according to the conditioned set size. Here, we consider systems of size $n+1 = 5$ with interaction orders $k = 2, 3, 4, 5$ obeying Eq.(3.8) where the last node, $X_5$, was considered as a target. The calculation of the $\overline{CMI}$ was made over all permutations of the set $\{X_i\}_{i=1}^{4}$ in the CMI formula. Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

Note that, by viewing $\mathcal{A}(\mathbf{X})$ as a *causal* informational lattice, $\mathcal{A}(\mathrm{PA}_Y)$, where $\mathrm{PA}_Y \equiv \mathbf{X}$ emphasizes that $\mathbf{X}$ is the set of parents of $Y$, and using the terminology of Def.3.2 we can identify the concepts of faithfulness and minimality in the PID language, see Prop.3.2 with proof in App.C.2. This allows us to clarify why faithfulness fails and minimality is necessary to capture high-order causal dependencies.

**Proposition 3.2 (Faithfulness & Minimality)**

*Consider $X \in PA_Y$, then the causal link $X \to Y$ satisfies,*

- *(a)* ***causal faithfulness*** *if $X$ has, necessarily, nonzero informational atoms belonging on $\mathcal{U}(X)$ or $\mathcal{R}^{(1)}(X)$, i.e., are independent from the others parents;*
- *(b)* ***causal minimality*** *if $X$ has, necessarily, nonzero informational atoms belonging on $\mathcal{U}(X)$, $\mathcal{S}^{(k\geq 2)}(X)$ or $\mathcal{R}^{(k\geq 2)}(X)$ $\forall k$, i.e., could depend on the others parents.*

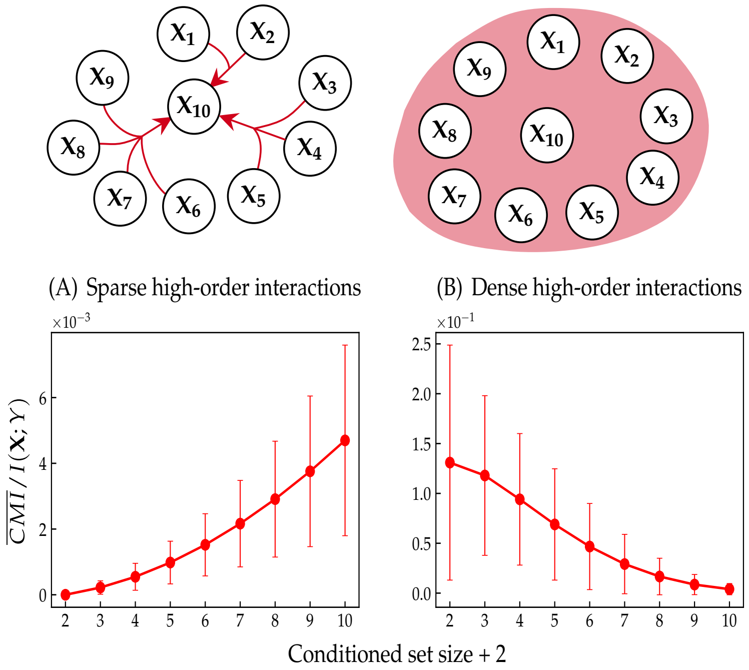
Prop.3.2-(a,b) straightly enlighten in an informational way how minimality includes faithfulness. The set $\mathcal{R}^{(1)}(X)$ can be viewed as having redundancy in the presence of faithful causes, i.e., there exists $B$ with $|B| = 1$, $\forall \beta$ in $\mathcal{R}^{(2)}(X)$, see Def.3.2-(b). On the other hand, $\mathcal{R}^{(k\geq 1)}(X)$ relaxes this requirement allowing redundancy among non-pairwise (synergic) causes.

## 3.3 The illusion of faithfulness or the deluge of redundancy?

Here, we analyze deeper why faithfulness can become optimal because of spurious correlations instead of genuine regularities. To answer this question, we fix the system size and investigate how a change in the organization of the interactions impacts the structure of the informational antichains and, consequently, the computation of MIs and CMIs. We will consider Hamiltonians with interactions up to order $k$,

$$H_k(\mathbf{X}) \;\;=\;\; -\sum_{i=1}^{n+1} J_i X_i - \sum_{i=1}^{n} \sum_{j=i+1}^{n+1} J_{ij} X_i X_j - \sum_{|\boldsymbol{\alpha}|=k} J_{\boldsymbol{\alpha}} \prod_{i\in\boldsymbol{\alpha}} X_i, \qquad (3.9)$$

Firstly, we model non-pairwise components *exclusively* in the Hamiltonian, which means that if a node has the interaction of order $k$, it cannot interact anymore, represented by the causal directed hyper-graph in Fig.(34)-A. For the second case, we relax the exclusivity condition, represented graphically by a dense cloud of connectivity, Fig.(34)-B.



(A) Sparse high-order interactions

(B) Dense high-order interactions

Figure 34 – Simulation of two non-pairwise systems of size $n + 1 = 11$ following Eq.(3.9) with different organization. (A) This model has spins with only exclusive interactions of order $k = 3, 4, 5$. (B) Here, we allowed all possible interactions of order $k = 3, \ldots, 10$. Again, the last spin, $X_{n+1}$, is the target node $Y$, and the calculation of $\overline{CMI}$ was done as explained in the previous figure.
Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

Our results show that when the Hamiltonian only possesses exclusive non-pairwise interactions, faithfulness still fails in capturing causal influence. Also, the monotonic increase of the $CMI$ according to the conditioning set size is preserved showing how minimality remains robust to identify high-order of synergism, Fig.34-(A). However, when we relax the exclusivity condition, even without pairwise interactions, the MI's are nonzero anymore. More intriguing, faithfulness is satisfied while minimality fails being a contradiction according to Prop.15, Fig.34-(B).

We argue that such behavior is due to the appearance of a specific type of redundancy, induced by the functional properties of the system. Indeed, as generated independently, we would expected that $I(X_i, X_j) = 0$ for any sources $X_i, X_j \in \mathrm{PA}_Y$, which is the case for the system of Fig.34-(A). However, it occurs that in the system of Fig.34-(B), we have $I(X_i, X_j) \neq 0$. The existence of these correlations among the sources spuriously inflates the calculation of $MI$s while it provokes the decrease of CMIs according to the conditioning set size.

Such redundancy is similar to the concept of mechanistic nature. (78,79) In its simplest form, this type of redundancy occurs when the sources are generated independently and are related to the functional properties of the system as well. While its importance has been recognized, how to define or quantify this type of redundancy inside PID is an open question. (80)

Regarding that, we relax the strong synergism condition from Def.3.2 by considering the synergic dominant set as

$$\mathcal{S}_d = \left\{ P_{Y|\mathbf{X}} \middle| X_i \perp\!\!\!\perp Y,\ \forall i \in [n]. \right\}. \tag{3.10}$$

This set can be viewed as a stochastic mapping from $\mathbf{X}$ to $Y$. Such mapping can be interpreted as unfolding information from the output $Y$ using the dataset $\mathbf{X}$ while keeping the constraints $X_i \perp\!\!\!\perp Y,\ \forall i \in [n]$. The redundancy dominant set, denoted by $\mathcal{R}_d$, is defined exactly as the complementary set of $\mathcal{S}_d$ in the respective antichain, see Fig.35 for an illustration. In Prop.C.2 from Appendix C.2 we showed that $\mathcal{R}_d$ encapsulates all redundancies of order 1 (using the terminology of Def.3.2), which means single-source redundancies being, then, a proposal to capture mechanistic redundancies.

Based on that and supported by previous work (81, 82) we can extend naturally Eq.3.10 to the $k$-synergic dominant set considering high-orders constraints as follows,

$$\mathcal{S}_d(k) = \left\{ P_{Y|\mathbf{X}} \middle| X_i \perp\!\!\!\perp Y|\mathbf{Z},\ \mathbf{Z} \subseteq \mathbf{X}_{\setminus X_i} \text{ where } |\mathbf{Z}| = k-1\ \forall i \in [n] \right\}. \tag{3.11}$$

These mappings can be seen as $k$-unfaithful mappings since violate faithfulness condition by construction, see Def.2.4-(FM'). This digression can be summarized in the following result,

**Proposition 3.3 (Causal premises via Redundant/Synergic dominance)**

*The prevalence of synergic and redundant dominant sets, $\mathcal{S}_d$ and $\mathcal{R}_d$, respectively, in $\mathcal{A}(\mathbf{X})$ for high-order scenarios raise a trade-off between causal faithfulness and causal minimality conditions.*

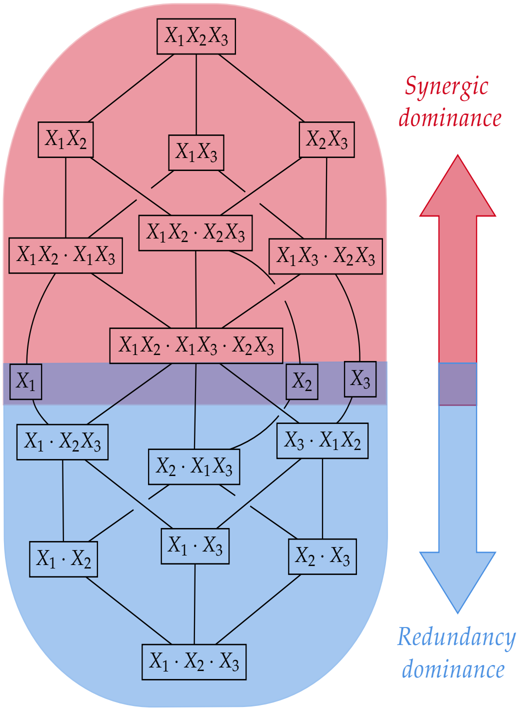


Figure 35 – Representation of the synergic and redundant dominant sets $\mathcal{S}_d$ and $\mathcal{R}_d$ in the informational antichain $\mathcal{A}(\mathbf{X})$ for $\mathbf{X} = \{X_1, X_2, X_3\}$.
Source: By the author.

Hence, in light of Proposition 3.3, it can be concluded that redundant dominant systems do not serve as optimal causal models. This is because faithfulness is achieved through spurious correlations rather than the existence of significant causal patterns, despite the violation of minimality.

As a final remark, we highlight that the reason for imposing the faithfulness assumption is primarily that the probability of finding distributions that violate it for a given graph $G$ is extremely low, almost negligible. (43) However, when dealing with a limited sample size, errors in estimation may occur. Robins *et al.* (83) demonstrated that many causal discovery algorithms, and the PC algorithm in particular, are not uniformly consistent due to the possibility of constructing a sequence of distributions that are faithful but still very similar to an unfaithful distribution. (84)

Our mappings from Eq.3.11 are represented by the same Pseudo-Independent Relations (PIR) in Def.3 from Lemeire *et al.*(85) where they also showed that PIR is one of the main reasons for faithfulness failure. Therefore, it should be interesting to investigate the prevalence in the space of probability distributions of the mappings from Eq.3.11

from the high-order point of view according to the size of $\mathbf{X}$. For this, we could study the behavior of $k$-synergic capacities defined as

$$S_k(\mathbf{X};Y) := \sup_{P_{Y|X_{[n]}} \in \mathcal{S}_d(k)} I(\mathbf{X};Y). \tag{3.12}$$

which carries the principle of synergic disclosure (81): it is possible for $\mathbf{X}$ to carry information about a feature $Y$ while revealing no information about any of the constraints in the dataset.

## 3.4 Towards an unifying framework

From Prop.3.3 we saw that redundancy dominant systems are not optimal causal models since violating minimality despite faithfulness being satisfied. How the $EI$-approach, which was designed to find optimal causal models, should behave in these scenarios? Here we investigate it by analyzing the behavior of determinism and degeneracy coefficients, Eqs.(3.13) and (3.14) below, respectively, for the systems of Fig.34. Based on simulations, Sec.3.4.2, we will argue for a link between the degeneracy coefficient and redundant dominated systems.

$$\text{determinism} := \mathbb{E}_{P(\widehat{x})}\left[D_{KL}\Big(P(Y|\widehat{x})\Big|\Big|P(\widehat{x})\Big)\right], \tag{3.13}$$

$$\text{degeneracy} := \mathbb{E}_{P(\widehat{x})}\left[D_{KL}\Big(\sum_{\widehat{x}'} P(\widehat{x}')P(y|\widehat{x}')\Big|\Big|P(\widehat{x})\Big)\right]. \tag{3.14}$$

Note that Eqs.(3.13) and (3.14) above expand Eqs.2.30 and 2.33 in the sense that it is relaxing $P(\widehat{x})$ to not be the maximum entropic distribution.

### 3.4.1 PID quantifiers as interventions

To tighten the link between the degeneracy coefficient and redundancy we should be able to interpret PID quantifiers as interventional strategies. Indeed, the PID quantifiers depend on the joint probability distribution $P(\mathbf{X}, Y)$ which, at first instance, makes not straight how to interpret them interventionally. However, note that $P(\mathbf{X}, Y) = P(Y|\mathbf{X})P(\mathbf{X})$. Even though the distribution $P(Y|\mathbf{X})$ is built on purely observational data generally understood in the Granger-causal sense, the $P(\mathbf{X})$ is chosen according to an agent's strategies resembling distributions of interventions, Sec.2.3.1.

This is the case for the redundant measure $I_{WB}$, Eq.(3.7), where we could interpret as an agent choosing strategies to the family of *specific mutual informations*,

$$I(\mathbf{X}; y), \tag{3.15}$$

given by $\min\{I(A_1; y), I(A_2; y), \ldots, I(A_k; y)\}$ where $A_i$ are the elements of the informational antichain $\mathcal{A}(\mathbf{X})$ and averaged over $P(Y)$. Although the specific values $\{p(y)|\mathbf{x}_i)\}$[4]

[4] Yet, we could go to the interventional regime, $\{p(y)|\text{do}(\mathbf{x}_i))\}$ straightly when the presence of confounders is relevant.

are in the observational regime, the overall information measure is constrained by an external agent's strategy despite its non-straight character.

The same is true for synergic measures from PID literature (31, 81), in particular, for the $k$-synergic capacities, Eq.(3.11), where an external agent put constraints in order to isolate *synergic* contributions from specific atoms in the informational antichain. Our point is that if we interpret these strategies as non-atomic interventions, we can put redundancy/synergic measures and causal measures as discussed in Sec.2.3.1 on an equal level.

Now, if we come back to Eqs.(3.13) and (3.14) we have that maximum determinism means $P(Y|\widehat{x}) \sim 1$ for every $\widehat{x}$ while minimum degeneracy $\{P(y|\widehat{x})\} \sim \delta_{\widehat{x},x'}$. We argue that this is achievable when one reduces redundancy in the system and captures causallly-relevant information (unique and synergic) as the examples in the next section.

### 3.4.2 Coarse-grainings strategies as the conversion of information

Analyzing the systems of Fig.34-(A), our findings are that due to the systemic character, $EI$ preserves synergism and remains constant across scales not differentiating orders of synergic behavior (deg=0), see Table 1. This is because in this scenario we have the contextual independence: $V_i \perp\!\!\!\perp Y | V_j$ for $i \neq j = \{1, 2, 3\}$, see Fig.36-top for an illustration. Such search is revealed by applying coarse-grains in groups stable under background contexts, remember Sec.2.4.2.

SPARSE CONNECTIONS

Table 1 – Calculation of Effective information ($EI$), determinism (det), and degeneracy (deg) (average over 100 simulations) for the system of Fig.34-(A). Here, the maximal entropic distribution is the uniform one, confirmed by the algorithm of the dit package (74). The measures remain constant over the coarse-graining strategies in Fig.36-(top).
Source: By the author.

| $(J_\alpha = 0.2)$ | | | | $(J_\alpha = 0.4)$ | | | |
|---|---|---|---|---|---|---|---|
| $N$ | $EI$ | det | deg | $N$ | $EI$ | det | deg |
| 4 | 0.06 | 0.06 | 1e-16 | 4 | 0.19 | 0.19 | 1e-16 |
| 7 | 0.50 | 0.50 | 1e-14 | 7 | 0.76 | 0.76 | 1e-14 |
| 11 | 0.93 | 0.93 | 1e-16 | 11 | 0.84 | 0.84 | 1e-16 |

For the case of redundancy dominance scenarios Fig.34-(B), we just analyzed the micro case since the noisy character of the dynamics gave a non-simple rule for system reduction. Despite that, as a proof-of-principle, we can see from Table 2 that det$\neq 0$ for these systems revealing the possibility of the entropic reduction for the set $\mathbf{X}$, see Fig.36-(bottom). This non-zero degeneracy coefficient tells us that redundant-dominant systems are not better explanatory in a causal sense. This is in accordance with our results from

Sec.3.3 where we could see that these systems raise a contradiction in the causal premises of prevalence of faithfulness with violation of minimality.

DENSE CONNECTIONS (proof-of-principle)

Table 2 – Calculation of Effective information ($EI$), determinism (det), and degeneracy (deg) coefficients (average over 100 simulations) for the system of Fig.34-(B). Here, the maximal entropic dynamics transitions do not converge to the uniform. To overcome it we constructed a baseline distribution that matches pairwise marginals with one another maximizing the total entropy using the dit package (74). Due to the time complexity of the algorithm, we kept the investigation for $N = 4, 5, 6$.
Source: By the author.

| ($J_\alpha = 0.2$) | | | | ($J_\alpha = 0.4$) | | | |
|---|---|---|---|---|---|---|---|
| $N$ | $EI$ | det | deg | $N$ | $EI$ | det | deg |
| 4 | 0.16 | 0.20 | 0.04 | 4 | 0.36 | 0.58 | 0.22 |
| 5 | 0.27 | 0.33 | 0.06 | 5 | 0.52 | 0.31 | 0.29 |
| 6 | 0.40 | 0.46 | 0.06 | 6 | 0.63 | 0.91 | 0.28 |

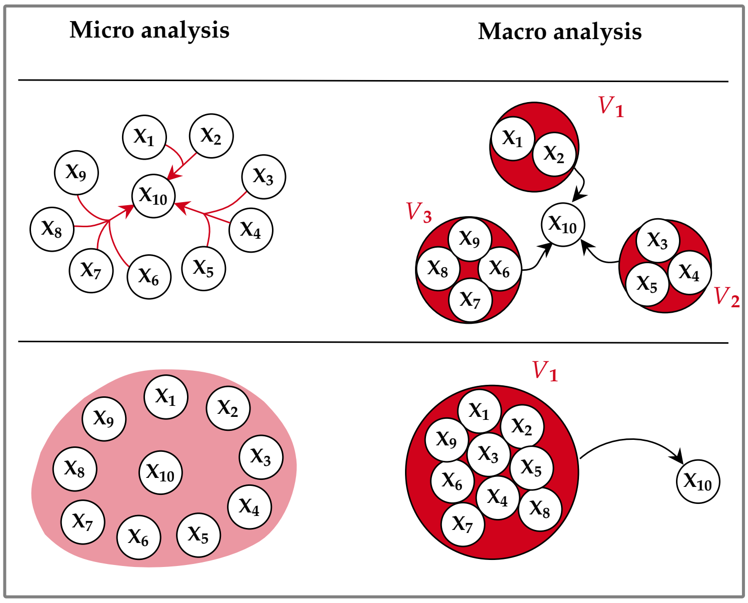


Figure 36 – Proposal of coarse-graining strategies for systems of Fig.34-(A) and (B).
Source: By the author.

Note that both systems are examples of downward causation mechanisms, in the sense that a group of sources is influencing a single target. In this sense, the determinism coefficient raises due to the size system instead of a reduction in transitions with the effects repertoire. Therefore, causal emergence in $EI$'s approach can be seen as the conversion of causally-irrelevant information in the form of redundancy (uncertainty of state transitions) to causally-relevant information while keeping synergic influence during the procedure.

# 4 CONCLUSION

The debate on causal emergence has led to the criticism that this concept is not compatible with materialism (https://philpapers.org/browse/downward-causation), as higher-level patterns are seen as just outcomes of dynamics at a lower level and therefore lack material instantiation and agency, making them uncausable. By using the framework of elementary dynamical systems, BUTTERFIELD (86) defines emergence as any kind of behavior that is novel in higher-level scales and reduction (coarse-grainings) as a deduction. Then, he was able to deduce emergent behavior by taking the limit, $N \to \infty$, for a given parameter $N$. However, the point here is that this infinite limit is not *physically real* (87). Instead, he claims that emergence (novelty) occurs before we get to the limit, i.e. for finite $N$. And it is this behavior which is physically real.

Our results in this thesis align with those argued by BUTTERFIELD, also similarly discussed by FLACK (88) resolving this conflict by a causal-informational approach, which defends how adaptive systems identify regularities by coarse-grainings and use them to guide behavior. Indeed, in this work we:

① Highlighted the use of non-atomic interventions to identify emergence in causal modeling. Also, we identified the connection of the $EI$-approach with Woodward's task for finding the optimal explanatory causal model (Chapter 2.4);

② Showed that faithfulness fails to capture synergism when the latter is appropriately defined using PID. (72) And, its assumption in high-order scenarios is more related to the appearance of spurious regularities than genuine causal influences where this specific spuriousness can be linked with the concept of (mechanistic) redundancy (Section 3.3);

③ We provide an interpretation in Pearls's interventionist sense for the quantifiers of redundancy/synergism in the PID framework (Section 3.4.1).

④ We connected the concept of effective information using PID according to levels of redundant and synergic information (Section 3.4.2);

An important property of coarse-grainings is an integration over component behavior. As we showed, such property is the reason for producing emergent behavior and is formally captured by the synergic atoms in the informational decomposition of the background inputs (parents set). This can define an informational principle to identify the coarse-grainings when the dynamical law is not known a priori or complicated in a way such that the maximum entropic distribution is never achieved making the $EI$-based approach unsuitable for these scenarios. By doing so we can unfold causal mechanisms according to the better explanatory scales where the premises are satisfied. Indeed, we can see that causal emergence (in Hoel's sense) cannot occur when only (noiseless) synergism is dominant and can occur when redundancies are dominant in the system, Sec.3.4.2. Such an argument is sustained in a recent data-driven analysis where only noise and redundancies from the data are eliminated, without integrating the compressed inputs-outputs (89). Even though this seems to reveal relevant causal information (physical patterns) we should be careful since it also reveals an important point in the approach: the unnecessary presence of high-order phenomena (synergic atoms in the PID sense).

A possible path to identify algorithmically the groups of synergic influence by correct coarse-grainings strategies in the causal domain could be to incorporate the so-called context-specific independence (CSI) (90), which is the independence that holds in a certain value of conditioned variables, i.e., the context. It has been shown that the presence and knowledge of such independence lead to more efficient probabilistic inference by exploiting the local structure of the causal models. (91) The XOR operation follows such relations. (92) Also, it allows the identification of causal effects, which would not be possible without any information about CSI relationships. (93) Further investigations of this approach as well others (94) to detect synergic causal influence in large data sets we leave for future work.

Also, it would be interesting to investigate the concept of causal emergence in the quantum realm. Since there, the concept of causality, as worked here, seems to suffer a significant change giving interesting consequences in quantum information processing. (95–97) A possible path could be to promote EI/PID formalisms to the quantum level and to investigate the distinct learning phases of quantum neurons. This could answer how synergic learning operates compare to similar classical investigations. (98–101)

# REFERENCES

1 PEARL, J. **Causality**: models, reasoning and inference. Cambridge: Cambridge University Press, 2009.

2 WOODWARD, J. **Making things happen**: a theory of causal explanation. Oxford: Oxford University Press, 2003.

3 LUPPI, A. I. *et al.* What it is like to be a bit: an integrated information decomposition account of emergent mental phenomena. **Neuroscience of Consciousnesss**, v. 2021, n. 2, p. niab027, 2021.

4 FREY, S.; ALBINO, D. K.; WILLIAMS, P. L. Synergistic information processing encrypts strategic reasoning in poker. **Cognitive Science**, v. 42, n. 5, p. 1457–1476, 2018.

5 CHAN, T. E.; STUMPF, M. P.; BABTIE, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. **Cell Systems**, v. 5, n. 3, p. 251 – 267.e3, 2017.

6 JAMES, R. G.; BARNETT, N.; CRUTCHFIELD, J. P. Information flows? a critique of transfer entropies. **Physical Review Letters**, v. 116, n. 23, p. 238701, 2016.

7 RUNGE, J. Causal network reconstruction from time series: from theoretical assumptions to practical estimation. **Chaos**, v. 28, n. 7, p. 075310, 2018.

8 NOVELLI, L. *et al.* Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. **Network Neuroscience**, v. 3, n. 3, p. 827–847, 2019.

9 ROSAS, F. E. *et al.* Reconciling emergences: an information-theoretic approach to identify causal emergence in multivariate data. **PLOS Computational Biology**, v. 16, n. 12, p. 1–23, 2020.

10 ANDERSON, C. **The end of theory**: the data deluge makes the scientific method obsolete. 2008. Available at: https://www.wired.com/2008/06/pb-theory/. Access at: 24 July 2022.

11 MARR, B. **How much data do we create every day**: the mind-blowing stats everyone should read. 2018. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=32dc511d60ba. Access at: 24 Jan. 2023.

12 BATTISTON, F. *et al.* Networks beyond pairwise interactions: structure and dynamics. **Physics Reports**, v. 874, p. 1–92, 2020.

13 VINTIADIS, E. Emergence. *In*: **Internet encyclopedia of philosophy**. 2022. Available at: https://iep.utm.edu/emergence/. Access at: 24 Jan. 2023.

14 HOEL, E. P. Agent above, atom below: how agents causally emerge from their underlying microphysics. *In*: AGUIRRE, A.; FOSTER, B.; MERALI, Z. (ed.). **Wandering towards a goal**: how can mindless mathematical laws give rise to aims and intention? Cham: Springer International Publishing, 2018. p. 63–76.

15 CHALMERS, D. J. Strong and weak emergence. *In*: DAVIES, P.; CLAYTON, P. (ed.). **The re-emergence of emergence**: the emergentist hypothesis from science to religion. Oxford: Oxford University Press, 2006.

16 CONWAY, J. The game of life. **Scientific American**, v. 223, n. 4, p. 4, 1970.

17 CONWAY'S Game of Life. Available at: https://en.wikipedia.org/wiki/Conway\%27s_Game_of_Life. Access at: 24 Jan. 2023.

18 KIM, J. **Mind in a physical world**: an essay on the mind-body problem and mental causation. Massachusetts: MIT Press, 1998.

19 HOEL, E. P.; ALBANTAKIS, L.; TONONI, G. Quantifying causal emergence shows that macro can beat micro. **Proceedings of the National Academy of Sciences**, v. 110, n. 49, p. 19790–19795, 2013.

20 HOEL, E. P. When the map is better than the territory. **Entropy**, v. 19, n. 5, 2017.

21 ALBANTAKIS, L. *et al.* What caused what? a quantitative account of actual causation using dynamical causal networks. **Entropy**, v. 21, n. 5, 2019.

22 WOODWARD, J. Causation in biology: stability, specificity, and the choice of levels of explanation. **Biology and Philosophy**, v. 25, n. 3, p. 287–318, 2010.

23 COVER, T. M.; THOMAS, J. A. **Elements of information theory**. New York: Wiley-Interscience, 2006 (Wiley series in telecommunications and signal processing).

24 PEARL, J. Causal inference in statistics. an overview. **Statistics Surveys**, v. 3, p. 96–146, 2009.

25 FODOR, J. A. **A theory of content and other essays**. Cambridge, MA: MIT Press, 1990.

26 PUTNAM, H. Minds and machines. *In*: HOOK, S. (ed.). **Dimensions of minds**. New York, USA: New York University Press, 1960. p. 138–164.

27 FODOR, J. A. Special sciences (or: The disunity of science as a working hypothesis). **Synthese**, v. 28, n. 2, p. 97–115, 1974.

28 BEER, R. D.; WILLIAMS, P. L. Information processing and dynamics in minimally cognitive agents. **Cognitive Science**, v. 39, n. 1, p. 1–38, 2015.

29 SCHNEIDMAN, E. *et al.* Network information and connected correlations. **Physical Review Letters**, v. 91, p. 238701, 2003.

30 WILLIAMS, P. L.; BEER, R. D. **Nonnegative decomposition of multivariate information**. **CoRR**, abs/1004.2515, 2010. Available at: http://dblp.uni-trier.de/db/journals/corr/corr1004.html#abs-1004-2515. Access at: 20 July 2019.

31 GRIFFITH, V.; KOCH, C. Quantifying synergistic mutual information. *In*: PROKOPENKO, M. (ed.). **Guided self-organization**: inception. Berlin: Springer, 2014. p. 159–190.

32 MEDIANO, P. A. M. *et al.* **Beyond integrated information:** a taxonomy of information dynamics phenomena. 2019. Available at: https://arxiv.org/abs/1909.02297. Access at: 24 July 2020.

33 MCGILL, W. J. Multivariate information transmission. **Psychometrika**, v. 19, n. 2, p. 97–116, 1954.

34 LUCE, D. R. Whatever happened to information theory in psychology? **Review of General Psychology**, v. 7, n. 2, p. 183–188, 2003.

35 HOEL, E. P. *et al.* Can the macro beat the micro? integrated information across spatiotemporal scales. **Neuroscience of Consciousness**, v. 2016 1, p. niw012, 2016.

36 BARRETT, A. B.; MEDIANO, P. A. M. **The phi measure of integrated information is not well-defined for general physical systems**. 2019. Available at: https://arxiv.org/abs/1902.04321. Access at: 24 July 2022.

37 BARRETT, A. B.; SETH, A. K. Practical measures of integrated information for time-series data. **PLOS Computational Biology**, v. 7, n. 1, p. 1–18, 2011.

38 BAREINBOIM, E. *et al.* **1 on Pearl's hierarchy and the foundations of causal inference**. 2021. Available at: https://causalai.net/r60.pdf. Access at: 24 Jan. 2023.

39 HUME, D. **An enquiry concerning human understanding**. 1748. Available at: https://socialsciences.mcmaster.ca/~econ/ugcm/3ll3/hume/enquiry.pdf. Access at: 22 July 2020.

40 PETERS, J.; JANZING, D.; SCHÖLKOPF, B. **Elements of causal inference**: foundations and learning algorithms. Cambridge: The MIT Press, 2017. (Adaptive computation and machine learning series).

41 HITCHCOCK, C. Causal models. *In*: ZALTA, E. N. (ed.). **The Stanford encyclopedia of philosophy**. Summer 2019. Stanford University: Metaphysics Research Lab, 2019.

42 HITCHCOCK, C. Probabilistic causation. *In*: ZALTA, E. N. (ed.). **The Stanford encyclopedia of philosophy**. Fall 2018. Stanford University: Metaphysics Research Lab, 2018.

43 SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, prediction, and search**. 2nd. ed. Massachusetts: MIT press, 2000.

44 EVANS, P. W.; SHRAPNEL, S. **The two sides of interventionist causation**. 2017. Available at: http://philsci-archive.pitt.edu/12906/. Access at: 20 July 2021.

45 WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v. 20, n. 7, p. 557–585, 1921.

46 WRIGHT, S. The method of path coefficients. **The Annals of Mathematical Statistics**, v. 5, n. 3, p. 161–215, 1934.

47 ILLARI, P. M.; RUSSO, F.; WILLIAMSON, J. **Causality in the sciences**. Oxford, UK: Oxford University Press, 2011. (Adaptive computation and machine learning series).

48 EICHLER, M. Causal inference with multiple time series: principles and problems. **Philosophical Transactions of the Royal Society A:** mathematical, physical and engineering sciences. v. 371, n. 1997, p. 20110613, 2013.

49 SUN, J.; TAYLOR, D.; BOLLT, E. Causal network inference by optimal causation entropy. **SIAM Journal on Applied Dynamical Systems**, v. 14, n. 1, p. 73–106, 2015.

50 AY, N.; POLANI, D. Information flows in causal networks. **Advances in Complex Systems**, v. 11, n. 01, p. 17–41, 2008.

51 SCHAMBERG, G. *et al.* Direct and indirect effects—an information theoretic perspective. **Entropy**, v. 22, n. 8, 2020.

52 TONONI, G.; SPORNS, O.; EDELMAN, G. Measures of degenercy and redundancy in biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, p. 3257–62, 04 1999.

53 GRIFFITHS, P. E. *et al.* Measuring causal specificity. **Philosophy of Science**, v. 82, n. 4, p. 529–555, 2015.

54 JANZING, D. *et al.* Quantifying causal influences. **Annals of Statistics**, v. 41, n. 5, p. 2324–2358, 2013.

55 CORREA, J.; BAREINBOIM, E. A calculus for stochastic interventions: causal effect identification and surrogate experiments. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 06, p. 10093–10100, 2020.

56 JAYNES, E. T. Information theory and statistical mechanics. **Physical Review Letters**, v. 106, p. 620–630, 1957.

57 HOPE, L. R.; KORB, K. B. An information-theoretic causal power theory. *In*: ZHANG, S.; JARVIS, R. (ed.). **Advances in artificial intelligence**. Berlin: Springer, 2005. p. 805–811.

58 HAUSMAN, D. M.; WOODWARD, J. Independence, invariance and the causal markov condition. **The British Journal for the Philosophy of Science**, v. 50, n. 4, p. 521–583, 1999.

59 HAUSMAN, D. M.; WOODWARD, J. Modularity and the causal markov condition: A restatement. **The British Journal for the Philosophy of Science**, v. 55, n. 1, p. 147–161, 2004.

60 POCHEVILLE, A.; GRIFFITHS, P. E.; STOTZ, K. **Comparing causes**: an information-theoretic approach to specificity, proportionality and stability. 2018. Available at: https://arnaud.pocheville.science/pdf/comparing-information-theoretic-web.pdf. Access at: 24 Jan. 2023.

61 LIZIER, J. T.; RUBINOV, M. Inferring effective computational connectivity using incrementally conditioned multivariate transfer entropy. **BMC Neuroscience**, v. 14, Supl. 1, 2013.

62 RUNGE, J. *et al.* Escaping the curse of dimensionality in estimating multivariate transfer entropy. **Physical Review Letters**, v. 108, p. 258701, 2012.

63 RUNGE, J. **Detecting and quantifying causality from time series of complex systems**. 2014. Dissertation (PhD) — Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2014.

64 CHICHARRO, D.; PANZERI, S. Algorithms of causal inference for the analysis of effective connectivity among brain regions. **Frontiers in Neuroinformatics**, v. 8, p. 64, 2014.

65 HLINKA, J.; KOŘENEK, J. Causal network discovery by iterative conditioning: comparison of algorithms. **Chaos**, v. 30, n. 1, 2018.

66 NOVELLI, L. *et al.* Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. **Network Neuroscience**, v. 3, n. 3, p. 827–847, 2019.

67 EICHLER, M. Graphical modeling of multivariate time series. **Probability Theory and Related Fields**, v. 153, n. 1, p. 233–268, 2012.

68 SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, n. 3, p. 379–423, 1948.

69 SHANNON, C. The lattice theory of information. **Transactions of the IRE Professional Group on Information Theory**, v. 1, n. 1, p. 105–107, 1953.

70 BIRKHOFF, G. **Lattice theory**. New York: American Mathematical Society, 1940. (American mathematical society colloquium publications, 2).

71 GLYMOUR, C.; ZHANG, K.; SPIRTES, P. Review of causal discovery methods based on graphical models. **Frontiers in Genetics**, v. 10, 2019.

72 MARTINELLI, T.; SOARES-PINTO, D. O.; RODRIGUES, F. A. **Shaking the causal tree**: on the faithfulness and minimality assumptions beyond pairwise interactions. 2022. Available at: https://arxiv.org/abs/2203.10665. Access at: 24 Jan. 2023.

73 GUTKNECHT, A. J.; WIBRAL, M.; MAKKEH, A. Bits and pieces: understanding information decomposition from part-whole relationships and formal logic. **Proceedings:** mathematical, physical and engineering sciences. v. 477, n. 2251, 2021.

74 JAMES, R. G.; ELLISON, C. J.; CRUTCHFIELD, J. P. dit: a python package for discrete information theory. **Journal of Open Source Software**, v. 3, n. 25, p. 738, 2018.

75 JAMES, R. G.; EMENHEISER, J.; CRUTCHFIELD, J. P. Unique information via dependency constraints. **Journal of Physics A:** mathematical and theoretical. v. 52, n. 1, p. 014002, 2018.

76 JAMES, R. G.; CRUTCHFIELD, J. P. Multivariate dependence beyond shannon information. **Entropy**, v. 19, n. 10, 2017.

77 YEUNG, R. W. **Information theory and network coding**. Berlin: Springer, 2008.

78 HARDER, M.; SALGE, C.; POLANI, D. Bivariate measure of redundant information. **Physical Review E**, v. 87, p. 012130, 2013.

79 JAMES, R. G.; EMENHEISER, J.; CRUTCHFIELD, J. P. **A Perspective on unique information:** directionality, intuitions, and secret key agreement. 2018. Available at: https://escholarship.org/uc/item/25k3c7ns. Access at: 22 Feb. 2020.

80 INCE, R. A. A. Measuring multivariate redundant information with pointwise common change in surprisal. **Entropy**, v. 19, n. 7, 2017.

81 ROSAS, F. E. *et al.* An operational information decomposition via synergistic disclosure. **Journal of Physics A:** mathematical and theoretical. v. 53, n. 48, p. 485001, 2020.

82 RASSOULI, B.; ROSAS, F. E.; GüNDüZ, D. Data disclosure under perfect sample privacy. **IEEE Transactions on Information Forensics and Security**, v. 15, p. 2012–2025, 2020.

83 ROBINS, J. M. *et al.* Uniform consistency in causal inference. **Biometrika**, v. 90, n. 3, p. 491–515, 2003. Available at: http://www.jstor.org/stable/30042062. Access at: 24 Jan. 2023.

84 UHLER, C. *et al.* Geometry of the faithfulness assumption in causal inference. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 41, n. 2, p. 436–463, 2013. Available at: http://www.jstor.org/stable/23566569. Access at: 24 Jan. 2023.

85 LEMEIRE, J. *et al.* Conservative independence-based causal structure learning in absence of adjacency faithfulness. **International Journal of Approximate Reasoning**, v. 53, n. 9, p. 1305–1325, 2012. Available at: https://www.sciencedirect.com/science/article/pii/S0888613X12000801.

86 BUTTERFIELD, J. Laws, causation and dynamics at different levels. **Interface Focus**, v. 2, n. 1, p. 101–114, 2012. Access at: 24 Jan. 2023.

87 BUTTERFIELD, J. Less is different: emergence and reduction reconciled. **Foundations of Physics**, v. 41, n. 6, p. 1065–1135, 2010.

88 FLACK, J. C. Coarse-graining as a downward causation mechanism.**Philosophical Transactions of the Royal Society A:** mathematical, physical and engineering sciences. v. 375, n. 2109, p. 20160338, 2017.

89 MARTINELLI, T. *et al.* **An informational approach to uncover the age group interactions in epidemic spreading from macro analysis**. 2023. Available at: https://arxiv.org/abs/arXiv:2306.00852v1. Access at: 24 Jan. 2023.

90 BOUTILIER, C. *et al.* Context-specific independence in Bayesian networks. *In:* INTERNATIONAL CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE. 12. **Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence**. Portland: Morgan Kaufmann, p. 115–123, 1996.

91 INWOO, H. *et al.* **On discovery of local independence over continuous variables via neural contextual decomposition**. 2023. Available at: https://openreview.net/forum?id=-aFd28Uy9td. Access at: 24 Jan. 2023.

92 MOKHTARIAN, E. *et al.* **Causal effect identification with context-specific independence relations of control variables**. 2021. Available at: https://proceedings.mlr.press/v151/mokhtarian22a/mokhtarian22a.pdf. Access at: 24 Jan. 2023.

93 TIKKA, S.; HYTTINEN, A.; KARVANEN, J. **Identifying causal effects via context-specific independence relations**. 2019. Available at: https://proceedings.neurips.cc/paper/2019/file/d88518acbcc3d08d1f18da62f9bb26ec-Paper.pdf. Access at: 24 Jan. 2023.

94 REING, K.; STEEG, G. V.; GALSTYAN, A. Discovering higher-order interactions through neural information decomposition. **Entropy**, v. 23, n. 1, 2021. Available at: https://www.mdpi.com/1099-4300/23/1/79. Access at: 24 Jan. 2023.

95 ORESHKOV, O.; COSTA, F.; BRUKNER, Č. Quantum correlations with no causal order. **Nature Communications**, v. 3, n. 1, 2012.

96 AGRESTI, I. *et al.* Experimental test of quantum causal influences. **Science Advances**, v. 8, n. 8, p. eabm1515, 2022.

97 ARAÚJO, M.; COSTA, F.; BRUKNER, Č. Computational advantage from quantum-controlled ordering of gates. **Physical Review Letters**, v. 113, p. 250402, 2014.

98 SHWARTZ-ZIV, R.; TISHBY, N. **Opening the black box of deep neural networks via information**. 2017. Available at: https://arxiv.org/pdf/1703.00810. Access at: 24 Jan. 2023.

99 TAX, T.; MEDIANO, P.; SHANAHAN, M. The partial information decomposition of generative neural network models. **Entropy**, v. 19, p. 474, 09 2017.

100 EHRLICH, D. A. *et al.* **Partial information decomposition reveals the structure of neural representations**. 2022. Available at: https://arxiv.org/abs/2209.10438. Access at: 24 Jan. 2023.

101 MATTSSON, S.; MICHAUD, E. J.; HOEL, E. **Examining the causal structures of deep neural networks using information theory**. 2020. Available at: https://arxiv.org/abs/2010.13871. Access at: 24 Jan. 2023.

102 JANZING, D. Causal versions of maximum entropy and principle of insufficient reason. **Journal of Causal Inference**, v. 9, n. 1, p. 285–301, 2021.

103 FROGNER, C.; POGGIO, T. **Fast and flexible inference of joint distributions from their marginals**. 2019. Available at: https://proceedings.mlr.press/v97/frogner19a.html. Access at: 24 Jan. 2023.

104 MATSUDA, H. Physical nature of higher-order mutual information: intrinsic correlations and frustration. **Physical Review E**, v. 62, n. 3 Pt A, p. 3096—3102, 2000.

105 ROSAS, F. E. *et al.* **Disentangling high-order mechanisms and high-order behaviours in complex systems**. 2022. Available at: https://arxiv.org/abs/2203.12041. Access at: 24 jan. 2023.

# APPENDIX

# APPENDIX A – INFORMATION THEORY

## A.1 Entropy, mutual information, and relative entropy

Information theory provides us with tools to measure uncertainty and to measure the reduction of that uncertainty. Importantly, for our purposes, it tells us how information about the value of one variable can reduce the uncertainty about the value of another, related, variable. The simplest case occurs when a discrete variable has only two values, which can then be known by answering a single question (e.g. by yes or no). The answer is said to convey one unit of information (a bit). If the set of possible values for the variable now contains $2^n$ equally likely elements, we can remark that $n$ dichotomous questions ($n$ bits) are needed to determine the actual value of the variable. The quantity of information contained in knowing the actual value is thus $n = \log_2(2^n)$. If we adopt a probabilistic framework where each possible value has equal probability $P = 1/2^n$, we can say that knowing any actual value of the variable brings $-\log_2(P)$ bits of information. When the values are not equiprobable, the average information gained by knowing the actual value of the variable is measured as an average over the probabilities of the different values. This quantity is the Shannon entropy of the probability distribution of the variable, defined as:

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log P(x_i) \tag{A.1}$$

where $x_i$ represent values of the variable $X$ and $N$ is the number of different values. Entropy measures the uncertainty about the value of the variable and is always non-negative.

### A.1.1 Mutual Information

If $X$ and $Y$ are two random variables (with respectively $N$ and $M$ different values, noted $x_i, y_i$, we can define the entropy of the couple $(X, Y)$:

$$H(X,Y) = -\sum_{i=1}^{N} \sum_{j=1}^{M} P(x_i, y_j) \log P(x_i, y_j). \tag{A.2}$$

With these two quantities, one is able to define the conditional entropy, representing the amount of uncertainty remaining on $Y$ when we already know $X$:

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) \qquad \text{(A.3)} \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{M} P(x_i) P(y_j|x_i) \log P(y_j|x_i). \end{aligned}$$

And, in a similar way, the mutual information, that is, the amount of redundant information present in $X$ and $Y$ is obtained by:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \end{aligned} \tag{A.4}$$

Mutual information can be thought of as the amount of information that one variable, $X$, contains about the other, $Y$ (normalized variants of mutual information are available). Conditional entropy is null, and mutual information is maximal when $Y$ is completely determined by $X$. Note that conditional entropy is generally asymmetric, $H(X|Y) \neq H(Y|X)$, while mutual information is always symmetric, $I(X;Y) = I(Y;X)$.

### A.1.2 Relative Entropy

Let's examine two probability distributions, denoted as $P$ and $Q$. Typically, $P$ represents the observed data, measurements, or a known probability distribution. On the other hand, $Q$ represents a theory, a model, a description, or an approximation of $P$. The Kullback-Leibler divergence or relative entropy is a measure that quantifies the average disparity in the number of bits needed to encode samples from $P$ using a code optimized for $Q$ instead of one optimized for $P$.

For discrete probability distributions $P$ and $Q$ defined on the sample space $\mathcal{X}$, the relative entropy from $Q$ to $P$ is defined to be (23)

$$D_{KL}(P||Q) := -\sum_{i=1}^{N} P(x_i) \log \frac{P(x_i)}{Q(x_i)}. \tag{A.5}$$

## A.2 Maximum entropic principle

Among the most prominent principles to assign priors are the *principle of insufficient reason* (PIR) and the *maximum entropic principle* (MEP) (56). PIR assigns uniform probabilities to a set of possible outcomes whenever the knowledge about the outcomes is invariant under permutations. MEP, which generalizes PIR, chooses a prior that maximizes entropy subject to the known constraints. It is known that both principles result in paradoxical probability updates for joint distributions of cause and effect. For a discussion of it, we refer to JANZING. (102)

The principle of insufficient reason, also called *Laplace's principle of insufficient reason* or *principle of indifference* (56), states that in the absence of any relevant evidence, agents should distribute their credence (or "degrees of belief") equally among all the possible outcomes under consideration. More explicitly, PIR advises considering all possible alternatives in a random experiment equally likely. For the simple example where we know that one of $n$ turns contains a ball, PIR considers each of the urns as an equally likely location and assigns $P(j) = 1/n$ for to each $j = 1, \ldots, n$.

Inferring underdetermined probability distributions by maximizing entropy subject to the available information is a well-established principle in machine learning and statistics, see, e.g., (103). The usual formal setting reads: Let us, for simplicity, assume that $X$ is a variable that attains values in some alphabet $\mathcal{X}$. Assume the only information available on $P(X)$ is given by the expectations

$$\sum p(x) f_j(X) = c_j, \quad \text{with} \quad c_j \in \mathbb{R}, \tag{A.6}$$

where $f_j$ are measurable functions. One constraint is always $f_0(X) = 1$ and $c_0 = 1$; that is, we constrain that it must be a proper probability distribution and integrate (sum) to 1. According to MEP, we would then choose the unique distribution maximizing the Shannon entropy subject to the constraints A.6, which yields

$$p(x) = \exp\big( - \sum_j \lambda_j f_j(x) - \lambda_0 - 1 \big), \tag{A.7}$$

with appropriate Lagrange multipliers $\lambda_j$. When we have the unconstrained condition, $f_j = 0$ for every $j \neq 0$ we arrive in

$$p(x) = \exp\big( - \lambda_0 - 1 \big), \quad \text{subject to} \quad \sum p(x) = 1, \tag{A.8}$$

giving the uniform distribution as the solution.

## A.3 No monotonicity of conditioning operation

Conditioning can either increase or decrease the mutual information between two variables, so $I(X;Y|Z) \nleq I(X;Y)$, and $I(X;Y|Z) \ngeq I(X;Y)$. To illustrate the last point, consider the following two examples where conditioning has different effects. In both cases, we will make use of the following equation

$$I(X;Y) + I(Z;Y|X) \quad = \quad I(Z;Y) + I(X;Y|Z). \tag{A.9}$$

Increasing example: if we have some $X, Y, Z$ such that $I(Z;Y) = 0$ (which means $X$ and $Y$ are independent variables), then Eq.(A.9) becomes: $I(X;Y|Z) = I(X;Y) + I(Z;Y|X)$, so $I(X;Y|Z) - I(X;Y) = I(Z;Y|X) \geq 0$, which implies $I(X;Y|Z) \geq I(X;Y)$.

Decreasing example: on the other hand, if we have a situation in which $I(Z;Y|X) = 0$, Eq.(A.9) becomes: $I(X;Y) = I(X;Z) + I(X;Y|Z)$, which in implies that $I(X;Y|Z) \leq I(X;Y)$.

# APPENDIX B – A BIT MORE ON CAUSALITY

## B.1 Estimating causal effects

The Back-Door criterion and Front-Door criterion are two methods proposed by Judea Pearl for determining causality from observational data (1). The Back-Door criterion states that to identify the causal effect of a variable $X$ on a variable $Y$, one must control for all variables that affect both $X$ and $Y$, also known as *back-door* variables. The Front-Door criterion states that to identify the causal effect of a variable $X$ on a variable $Y$, one must adjust for an instrumental variable, which is a variable that affects only $X$ and $Y$, but not through any back-door variables.

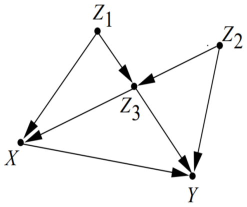
**Definition B.1 (Back-Door Criterion)**

A set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to an ordered pair of variables $(X, Y)$ in a DAG $G$ if:

1. no node in $\mathbf{Z}$ is a descendant of $X$; and
2. $\mathbf{Z}$ blocks every path between $X$ and $Y$ that contains an arrow into $X$.

When $Z$ satisfies the back-door criterion, for instance:

$$P(Y = y|\mathrm{do}(X = x_0)) = \sum_{\mathbf{z}} P(\mathbf{z})P(Y|x_0, \mathbf{z}) \tag{B.1}$$



$$P(y|do(x_0)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

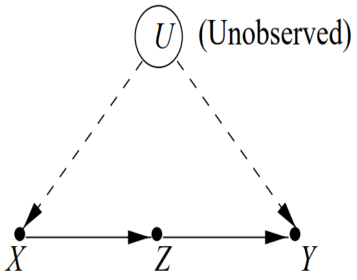
**Definition B.2 (Front-Door Criterion.)**

A set of variables $\mathbf{Z}$ satisfies the front-door criterion relative to an ordered pair of variables $(X, Y)$ in a DAG $G$ if:

1. $\mathbf{Z}$ all directed paths from $X$ to $Y$;
2. there is no unblocked backdoor path from $X$ to $\mathbf{Z}$;
3. All back-door paths from $\mathbf{Z}$ to $Y$ are blocked by $X$.

When $Z$ satisfies the front-door criterion relative to $(X, Y)$ and if $P(x, z) > 0$, then the causal effect of $X$ on $Y$ is identifiable and is given by:

$$P(y|\mathrm{do}(x_0)) = \sum_{\mathbf{z}} P(\mathbf{z}|x_0) \sum_{x'} \mathrm{P}(y|x', \mathbf{z})\mathrm{P}(x') \tag{B.2}$$



## B.2 Equivalence between defs. 2.3 and 2.4

**Proof** (Equivalence of causal Markov condition)**.** *Let's apply (CF) in the DAG of Fig.17. Notice that, it holds for any disjoint set $X$ and $Z$, by choosing $X \equiv X_i$ (a single node) and $Z \equiv \emptyset$ we have that*

$$(X_i \perp\!\!\!\perp_P Y) \implies (X_i \perp\!\!\!\perp_G Y), \tag{B.3}$$

*as i is arbitrary, (CF) holds for any single parent from $PA_Y$ saying that faithfulness ensures that every causal parent presents an observable effect regardless of the information about other causal parents — stable under the background parents (the reason that is also called stability, Def.2.4.1 from (1)). Therefore, the only way to violate (CF) in the DAG of Fig.17 is by having $X = Z^1$. Finally, by noting that $X, Z \subseteq PA_Y$, we have $X \cap PA_Y = X$ and $Z \cap PA_Y = Z$, ending the proof.*

**Proof** (Equivalence of causal minimality)**.** *Consider the random vector $\mathbf{X} = (X_1, \dots, X_n)$ and assume that the joint distribution $P$ has a density with respect to a product measure. Suppose that $P$ is Markovian with respect to $G$. Then $P$ satisfies causal minimality with respect to $G$ if and only if $\forall X_j$ and $\forall X_i \in PA_j$ we have that $\big(X_j \not\perp\!\!\!\perp_P X_i \mid PA_j \setminus X_i\big)$.*

*($\Longrightarrow$) Assume that causal minimality is not satisfied. Then, there are $X_j$ and $X_i \in PA_{X_j}$, such that $P$ is also Markovian with respect to the graph obtained when removing the edge $X_i \to X_j$ from $G$. This implies $\big(X_j \perp\!\!\!\perp_P X_i \mid PA_{X_j} \setminus X_i\big)$, by the causal Markov property.*

*($\Longleftarrow$): If $P$ has a density, then $CMC_{Factorization}$ is equivalent to $CMC_{d\text{-}separation}$. Assume now that $X_i \in PA_{X_j}$ and $\big(X_j \perp\!\!\!\perp_P X_i \mid PA_{X_j} \setminus Y\big)$, which implies $P(X_j \mid PA_{X_j}) = P(X_j \mid PA_{x_j} \setminus X_i)$. Then, $P(x) = P(X_j \mid PA_{X_j}) \prod_{k \neq j} P(X_k \mid PA_{X_k})$, which implies that $P$ is Markovian with respect to $G$ without $Y \to X_j$.*

[1] d-separation here is only possible by cutting the link between source and target or conditioning on an equal source since the data is time-ordered and, therefore, the only set able to d-separate any $X$ from $Y$ is the parents set $\mathrm{PA}_Y$ (1).

# APPENDIX C – CONDITIONING ANTICHAINS

Going to the PID in the multivariate case is not so straightforward. Indeed, adding only one variable more is sufficient to see the failure of elimination of redundancy when conditioning. Consider the three variable case, $\mathbf{X} = \{X_1, X_2, X_3\}$, then

$$\begin{aligned} I(X_1; Y|X_2, X_3) &= I(X_1, X_2, X_3; Y) - I(X_2, X_3; Y) \quad \text{(C.1)} \\ &= I_\partial(X_1; Y) + I_\partial(X_1X_2; Y) + I_\partial(X_1X_3; Y) + I_\partial(X_1X_2 \cdot X_1X_3; Y). \end{aligned}$$

From Eq.C.1, we can see that there is the existence of redundancy in the last term which is not eliminated by the operation of conditioning, see Fig.38. This is because there are new kinds of terms representing combinations of redundancy and synergy which are not included in the down set[1] of $\{X_2, X_3\}$, $\downarrow\{X_2, X_3\}$. On the other hand, we can see that all orders of synergic atoms are included. Prop.C.1 formalizes it.

**Proposition C.1 (PID view of conditioning operation)**

*Given the set $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathbf{Z} \subseteq \mathbf{X}_{\setminus X}$ and $X \in \mathbf{X}$ the PID view of conditioning on $\mathbf{Z}$ the information between $X$ and $Y$ is given by*

$$\begin{aligned} I(X; Y \mid \mathbf{Z}) &:= I(X, \mathbf{Z}; Y) - I(\mathbf{Z}; Y) \\ &= \sum_{\alpha \in \downarrow\{X, \mathbf{Z}\}} I_\partial(\alpha; Y) - \sum_{\alpha \in \downarrow \mathbf{Z}} I_\partial(\alpha; Y) = \sum_{\alpha \in (\downarrow \mathbf{Z})^{\complement}} I_\partial(\alpha; Y) \quad \text{(C.2)} \end{aligned}$$

*where $(\downarrow\mathbf{Z})^{\complement}$ is the complementary set of $\downarrow\mathbf{Z}$ given the particular subset of collections of $\mathbf{X}$ used to build the information lattice, in this case, $\{X, \mathbf{Z}\}$.*

**Proposition C.2 (The CMI and MI antichains)**

*Consider the informational antichain $\mathcal{A}(\mathbf{X})$ with $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$. The subset of atoms computed by the MI, $I(X_i; Y)$, being denoted by $\mathcal{A}_{MI}(X)$, captures all redundant terms of the form $\{\{X_i\}, \{\mathcal{A}(\mathbf{X}_{\setminus X_i})\}\}$. On the other hand, the subset of the CMI, $I(X_i; Y|\mathbf{X}_{\setminus X_i})$, denoted by $\mathcal{A}_{CMI}(\mathbf{X}_{\setminus X_i})$, captures redundant terms of the form $\{\mathcal{A}(X_i!,\ \mathbf{X}_{\setminus X_i})\}$ with the notation $X_i!$ meaning that the term $X_i$ is carried throughout the atoms in the respective antichain. See Fig.39 for a graphical illustration.*

[1] the down set of $\alpha$ means that $\beta \preceq \alpha$ for $\alpha, \beta \in \mathcal{A}(\mathbf{X})$.

## C.1 Graphical illustration of Def.3.2 in informational antichains

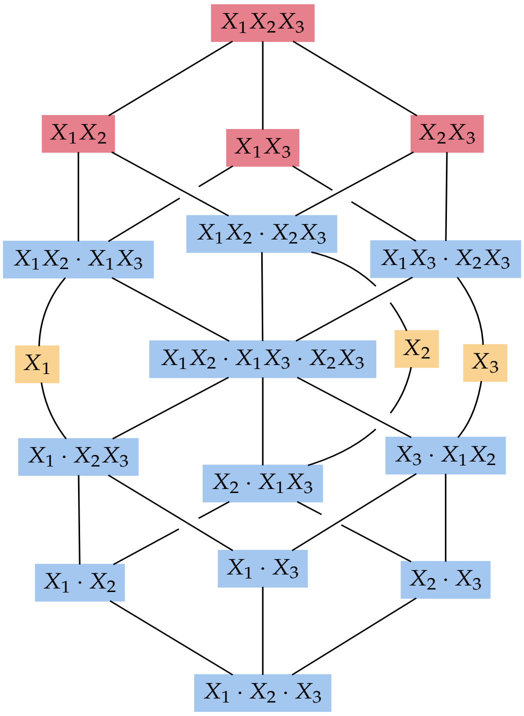


Figure 37 – Illustration of Def.1 for the informational lattice $\mathcal{A}(\mathbf{X})$ with $\mathbf{X} = \{X_1, X_2, X_3\}$. Colored boxes explanation: 🟦 atoms inside $\mathcal{R}(\mathbf{X})$; 🟨 atoms inside $\mathcal{U}(\mathbf{X})$ and 🟥 atoms inside $\mathcal{S}(\mathbf{X})$. Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

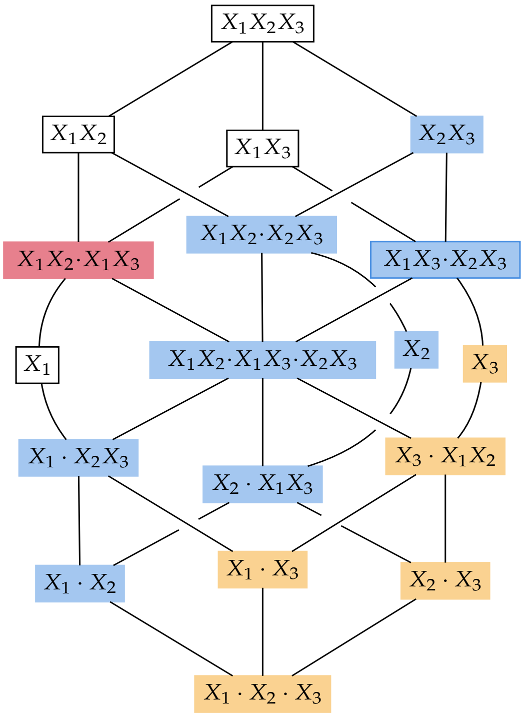


Figure 38 – Illustration of the effect of conditioning the information between $\mathbf{X} = \{X_1, X_2, X_3\}$ and target node $Y$ on the element $\{X_2X_3\} \in \mathbf{X}^-$ against $\{X_3\} \in$ for informational lattice $\mathcal{A}(\mathbf{X})$. Coloured boxes explanation: 🟨 goes away when $X_1 \mid \{X_3\}$; 🟨+🟦 goes away when $X_1 \mid \{X_2, X_3\}$; 🟥 remains when $X_1 \mid \{X_2, X_3\}$. Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

**Proof.** *The proof can be done by noting that the down set* $\downarrow \mathbf{Z} \subseteq \downarrow \mathbf{X}$ *is exactly a sub-antichain of* $\mathcal{A}(\mathbf{X})$ *with elements of the form* $\{\mathcal{A}(\mathbf{Z}!, \ \mathbf{X}_{\setminus \mathbf{Z}})\}$*. By applying the PID in* $I(X_i;Y)$ *and* $I(X_i;Y|\mathbf{X}_{\setminus X_i})$*:*

$$I(X_i;Y) = \sum_{\alpha \in (\downarrow X_i)} I_\partial(\alpha;Y), \tag{C.3}$$

$$I(X_i;Y|\mathbf{X}_{\setminus X_i}) = \sum_{\alpha \in (\downarrow \mathbf{X}_{\setminus X_i})^{\complement}} I_\partial(\alpha;Y), \tag{C.4}$$

*we have that the only synergic element in Eq.C.3 is the singleton* $\{\{X_i\}\}$ *and by the property of being a down set and satisfying the condition of antichain, all the remaining terms* *are redundant of the form $\{\{X_i\}, \{\mathcal{A}(\mathbf{X}_{\setminus X_i})\}\}$. Now, for Eq.C.4, the sinergic elements in Eq.C.3 are the all the singletons inside $2^{\mathbf{X}} \setminus 2^{(\mathbf{X}_{\setminus X_i})}$ where the notation $2^{\mathbf{Z}}$ means the powerset of $\mathbf{Z}$.*

*To illustrate, considering $PA_Y \equiv \mathbf{X} = \{X_1, X_2, X_3\}$, $X_j \equiv X_1$, and $\mathbf{Z} = \{X_2, X_3\}$ we have that $I(X_1; Y|\mathbf{Z})$ preserves the synergic atoms $\{\{X_1\}, \{\{X_1X_2\}\}, \{\{X_1X_3\}\}, \{\{X_1X_2X_3\}\}\}$, see Fig.39.*

*Finally, the redundant terms are the ones presented in $\mathcal{A}(\mathbf{X})$ not belonging into the down set $\downarrow \{\mathbf{X}_{\setminus X_i}\}$. These terms are exactly of the form $\{\mathcal{A}(X_i!,\ \mathbf{X}_{\setminus X_i})\}$.*

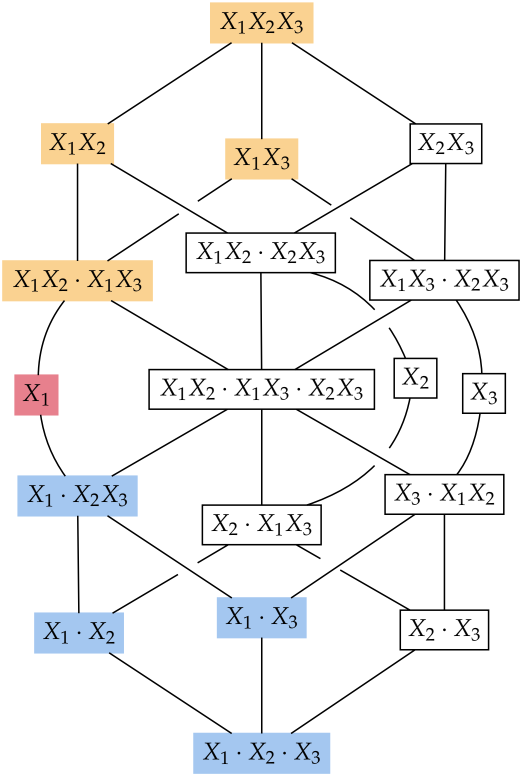


Figure 39 – Illustration of the ordered subsets $\mathcal{A}_{MI}(X_1)$ and $\mathcal{A}_{CMI}(\mathbf{X}_{\setminus X_1})$ constituted by the atoms of partial information in MI and CMI (with the longest conditioned) operations, respectively, for the set $\mathbf{X} = \{X_1, X_2, X_3\}$. Coloured boxes explanation: 🟧 + 🟥 atoms inside $\mathcal{A}_{CMI}(\mathbf{X}_{\setminus X_1})$; 🟦 + 🟥 atoms inside $\mathcal{A}_{MI}(X_1)$.
Source: MARTINELLI; SOARES-PINTO; RODRIGUES. (72)

## C.2 Proofs of Props. 3.1 and 3.2

**Proof** (Proof of Prop.3.1)**.** *The first part was proven in Prop.C.2. Let's show the synergism monotonic increasing. That is, given $I(X_j; Y|Z_1)$ and $I(X_j; Y|Z_2)$, then*

$$|Z_1| \leq |Z_2| \Rightarrow |\mathcal{S}_{CMI}(Z_1)| \leq |\mathcal{S}_{CMI}(Z_2)|, \tag{C.5}$$

*where $\mathcal{S}_{CMI}(Z)$ denotes the set of all synergic atoms captured by CMI with conditioning set $Z$. It says that the CMI with a large conditioned set has more synergic terms. This* *can be seen by noting that* $|\mathcal{S}_{CMI}(Z)| = |2^Z \setminus 2^{(Z \setminus X_j)}| \propto 2^Z$ *grows exponentially according to* $|Z|$.

*Also, the term of order* $S^{(k)}$ *is only captured when the conditioned* $Z$ *set has size* $k-2$. *To clarify, if we choose* $Z_2 = \{X_2, X_3\}$ *as the example above and* $Z_1 = \{X_2\}$ *we have that* $I(X_1;Y|Z_1)$ *preserves* $\{\{X_1\},\{\{X_1X_2\}\}\}$ *and the synergic term of order* $k=3$, $\mathcal{S}^{(3)} \equiv \{\{X_1X_2X_3\}\}$ *(present in* $I(X_1;Y|Z_2)$*) is missing since* $|Z_1| = 1$ *and we must have a conditioned set of size* $k-2=2$ *to capture this term.*

*The proof can be seen by noting that the cardinality of* $\mathbf{Z}$ *— which means conditioning on higher nodes of the lattice — tells the order of synergic terms that* $I(X_j;Y \mid \mathbf{Z})$ *includes. Indeed, suppose that we want to include all synergic influence of orders* $\leq n$ *among* $X_j \in \mathbf{X}$ *and* $n-1$ *elements from* $\mathbf{X}_{\setminus X_j}$ *on node* $Y$. *Then, w.l.o.g., note that* $I(X_j;Y \mid \mathbf{Z})$ *with* $\mathbf{Z} := \mathbf{X}_{\setminus\{X_j,X_1\}}$ *and* $j \neq 1$ *does not include any synergic informational atom with order* $n$ *of the type* $I_\partial(X_1..X_j..X_{n-1};Y)$ *because* $X_1$ *was discarded. The argument is the same for* $\mathbf{Z} := \mathbf{X}_{\setminus\{X_i,\mathbf{W}\}}, \forall \mathbf{W} \subset \mathbf{X}$.

**Proof** (Proof of Prop.3.2)**.** *Consider the joint distribution* $P$ *given by the tuple* $(Y, PA_Y)$ *and the causal link of* $X \in PA_Y$ *on* $Y$.

- *(a) Suppose that* $X$ *is faithful. By the correspondence* $I(X;Y|Z) = 0 \iff (X \perp\!\!\!\perp_P Y|Z)$ *we have that* $I(X;Y) \neq 0$, *where we used* $K \equiv X$ *and* $L \equiv \emptyset$ *in Def.2.4-(CF'). And, using Eq.C.3 from Prop.C.2 we have that* $\exists \alpha \neq 0$ *such that* $\alpha \in \mathcal{U}(X)$ *or* $\alpha \in \mathcal{R}^{(1)}(X)$;

- *(b) Now, suppose that* $X$ *satisfies minimality. Then* $I(X;Y|PA_Y \setminus X) \neq 0$ *by Def.2.4-(CM') which give us the condition of Eq.C.3 from Prop.C.2 having that* $\exists \alpha \neq 0$ *such that* $\alpha \in \{\mathcal{U}(X), \mathcal{S}^{(k>1)}(X), \mathcal{R}^{(k>1)}(X)\}$.

# APPENDIX D – SIMULATION DETAILS

For Figs.33 and 34 the systems were of $n+1$ spins, where $\mathcal{X}_i = \{-1,1\}$ for $i = 1,\ldots,n+1$ whose joint probability distributions can be expressed in the form $p_{\mathbf{X}_{n+1}}(\mathbf{x}_{n+1}) = \exp\{H_k(\mathbf{X}_{n+1})\}/Z$, with $\beta$ the inverse temperature choose as 1, $Z$ the normalization constant to make sure that the $p_{\mathbf{X}_{n+1}}$'s are probabilities, and $H_k(\mathbf{X}_{n+1})$ the Hamiltonian function. In all simulations, all interaction coefficients $J$ in the Hamiltonians were generated i.i.d. from a uniform distribution weighted by the coefficient 0.2. Also, 100 Hamiltonians were sampled at random for each order $k$ in every experiment.

It is worth mentioning that we have generated (holistic) synergy (75) of specific orders using multivariate statistics from high-order terms in the system's Hamiltonian and the corresponding Boltzmann distribution to point out the failure of faithfulness when (holistic) synergism is present due to high-order mechanisms.

We have fixed the system to have holistic synergy because one should have special systems to give raise to non-holistic synergism from low order connections (104, 105). Nonetheless, the faithfulness failure in the main text is due to high-order behaviors — the need to go beyond pairwise statistics — which are present undoubtedly when the data-generating process has higher-order interactions but is not restricted to it.