Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices

Conor Finn [1,2,*] and **Joseph T. Lizier** [1]

[1]   Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering & IT,
     The University of Sydney, NSW 2006, Australia; joseph.lizier@sydney.edu.au
[2]   CSIRO Data61, Marsfield NSW 2122, Australia
[*]   Correspondence: conor.finn@sydney.edu.au

Abstract: What are the distinct ways in which a set of predictor variables can provide information about a target variable? When does a variable provide unique information, when do variables share redundant information, and when do variables combine synergistically to provide complementary information? The redundancy lattice from the partial information decomposition of Williams and Beer provided a promising glimpse at the answer to these questions. However, this structure was constructed using a much criticised measure of redundant information, and despite sustained research, no completely satisfactory replacement measure has been proposed. In this paper, we take a different approach, applying the axiomatic derivation of the redundancy lattice to a single realisation from a set of discrete variables. To overcome the difficulty associated with signed pointwise mutual information, we apply this decomposition separately to the unsigned entropic components of pointwise mutual information which we refer to as the specificity and ambiguity. This yields a separate redundancy lattice for each component. Then based upon an operational interpretation of redundancy, we define measures of redundant specificity and ambiguity enabling us to evaluate the partial information atoms in each lattice. These atoms can be recombined to yield the sought-after multivariate information decomposition. We apply this framework to canonical examples from the literature and discuss the results and the various properties of the decomposition. In particular, the pointwise decomposition using specificity and ambiguity satisfies a chain rule over target variables, which provides new insights into the so-called two-bit-copy example.

1. Introduction

The aim of information decomposition is to divide the total amount of information provided by a set of predictor variables, about a target variable, into atoms of partial information contributed either individually or jointly by the various subsets of the predictors. Suppose that we are trying to predict a target variable T, with discrete state space $\mathcal{T}$, from a pair of predictor variables S_1 and S_2, with discrete state spaces $\mathcal{S}_1$ and $\mathcal{S}_2$. The mutual information $I(S_1; T)$ quantifies the information S_1 individually provides about T. Similarly, the mutual information $I(S_2; T)$ quantifies the information S_2 individually provides about T. Now consider the joint variable $S_{1,2}$ with the state space $\mathcal{S}_1 \times \mathcal{S}_2$. The (joint) mutual information $I(S_{1,2}; T)$ quantifies the total information S_1 and S_2 together provide about T. Although Shannon's information theory provides the prior three measures of information, there are four possible ways S_1 and S_2 could contribute information about T: the predictor S_1 could uniquely provide information about T; or the predictor S_2 could uniquely provide information about T;

both S_1 and S_2 could both individually, yet redundantly, provide the same information about T; or the predictors S_1 and S_2 could synergistically provide information about T which is not available in either predictor individually. Thus we have the following underdetermined set of equations,

$$
\begin{aligned}
I(S_{1,2}; T) &= R(S_1, S_2 \to T) + U(S_1 \backslash S_2 \to T) + U(S_2 \backslash S_1 \to T) + C(S_1, S_2 \to T), \\
I(S_1; T) &= R(S_1, S_2 \to T) + U(S_1 \backslash S_2 \to T), \\
I(S_2; T) &= R(S_1, S_2 \to T) + U(S_2 \backslash S_1 \to T),
\end{aligned}
\tag{1}
$$

where $U(S_1 \backslash S_2 \to T)$ and $U(S_2 \backslash S_1 \to T)$ are the unique information provided by S_1 and S_2 respectively, $R(S_1, S_2 \to T)$ is the redundant information, and $C(S_1, S_2 \to T)$ is the synergistic or complementary information. (The directed notation is utilise here to emphasis the privileged role of the variable T.) Together, the equations in (1) form the bivariate information decomposition. The problem is to define one of the unique, redundant or complementary information—something not provided by Shannon's information theory—in order to uniquely evaluate the decomposition.

Now suppose that we are trying to predict a target variable T from a set of n finite state predictor variables $\mathbf{S} = \{S_1, \ldots, S_n\}$. In this general case, the aim of information decomposition is to divide the total amount of information $I(S_1, \ldots, S_n; T)$ into atoms of partial information contributed either individually or jointly by the various subsets of $\mathbf{S}$. But what are the distinct ways in which these subsets of predictors might contribute information about the target? Multivariate information decomposition is more involved than the bivariate information decomposition because it is not immediately obvious how many atoms of information one needs to consider, nor is it clear how these atoms should relate to each other. Thus the general problem of information decomposition is to provide both a structure for multivariate information which is consistent with the bivariate decomposition, and a way to uniquely evaluate the atoms in this general structure.

In the remainder of Section 1, we will introduce an intriguing framework called partial information decomposition (PID), which aims to address the general problem of information decomposition, and highlight some of the criticisms and weaknesses of this framework. In Section 2, we will consider the underappreciated pointwise nature of information and discuss the relevance of this to the problem of information decomposition. We will then propose a modified pointwise partial information decomposition (PPID), but then quickly repudiate this approach due to complications associated with decomposing the signed pointwise mutual information. In Section 3, we will discuss circumventing this issue by examining information on a more fundamental level, in terms of the unsigned entropic components of pointwise mutual information which we refer to as the specificity and the ambiguity. Then in Section 4—the main section of this paper—we will introduce the PPID using the specificity and ambiguity lattices and the measures of redundancy in Definitions 1 and 2. In Section 5, we will apply this framework to a number of canonical examples from the PID literature, discuss some of the key properties of the decomposition, and compare these to existing approaches to information decomposition. Section 6 will conclude the main body of the paper. Appendix A contains discussions regarding the so-called two-bit-copy problem in terms of Kelly gambling, Appendix B contains many of the technical details and proofs, while Appendix B contains some more examples.

1.1. Notation

The following notational conventions are observed throughout this article:

$T, \mathcal{T}, t, t^c$ denote the *target* variable, event space, event and complementary event respectively;

$S, \mathcal{S}, s, s^c$ denote the *predictor* variable, event space, event and complementary event respectively;

$\mathbf{S}, \mathbf{s}$ represent the *set* of n predictor variables $\{S_1, \ldots, S_n\}$ and events $\{s_1, \ldots, s_n\}$ respectively;

$\mathcal{T}^t, \mathcal{S}^s$ denote the *two-event partition* of the event space, i.e., $\mathcal{T}^t = \{t, t^c\}$ and $\mathcal{S}^s = \{s, s^c\}$;

$H(T), I(S; T)$ uppercase function names be used for *average* information-theoretic measures;

$h(t), i(s, t)$ lowercase function names be used for *pointwise* information-theoretic measures.

When required, the following index conventions are observed:

s^1, s^2, t^1, t^2 superscripts distinguish between different *different events* in a variable;

S_1, S_2, T_1, T_2 subscripts distinguish between *different variables*;

$S_{1,2}, s_{1,2}$ multiple superscripts represent *joint variables* and *joint events*.

Finally, to be discussed in more detail when appropriate, consider the following:

$A_1, \ldots, A_k$ *sources* are sets of predictor variables, i.e., $A_i \in \mathscr{P}_1(S)$ where $\mathscr{P}_1$ is the power set without $\varnothing$;

$a_1, \ldots, a_k$ *source events* are sets of predictor events, i.e., $a_i \in \mathscr{P}_1(s)$.

1.2. Partial Information Decomposition

The *partial information decomposition* (PID) of Williams and Beer [1,2] was introduced to address the problem of multivariate information decomposition. The approach taken is appealing as rather than speculating about the structure of multivariate information, Williams and Beer took a more principled, axiomatic approach. They start by considering potentially overlapping subsets of S called sources, denoted $A_1, \ldots, A_k$. To examine the various ways these sources might contain the same information, they introduce three axioms which "any reasonable measure for redundant information [$I_\cap$] should fulfil" ([3], p. 3502). Note that the axioms appear explicitly in [2] but are discussed in [1] as mere properties; a published version of the axioms can be found in [4].

W&B Axiom 1 (Commutativity). *Redundant information is invariant under any permutation σ of sources,*

$$I_\cap(A_1, \ldots, A_k \to T) = I_\cap(\sigma(A_1), \ldots, \sigma(A_k) \to T).$$

W&B Axiom 2 (Monotonicity). *Redundant information decreases monotonically as more sources are included,*

$$I_\cap(A_1, \ldots, A_{k-1} \to T) \leq I_\cap(A_1, \ldots, A_k \to T)$$

with equality if $A_k \supseteq A_i$ for any $A_i \in \{A_1, \ldots, A_{k-1}\}$.

W&B Axiom 3 (Self-redundancy). *Redundant information for a single source A_i equals the mutual information,*

$$I_\cap(A_i \to T) = I(A_i ; T).$$

These axioms are based upon the intuition that redundancy should be analogous to the set-theoretic notion of intersection (which is commutative, monotonically decreasing and idempotent). Crucially, Axiom 3 ties this notion of redundancy to Shannon's information theory. In addition to these three axioms, there is an (implicit) axiom assumed here known as *local positivity* [5], which is the requirement that all atoms be non-negative. Williams and Beer [1,2] then show how these axioms reduce the number of sources to the collection of sources such that no source is a superset of any other. These remaining sources are called *partial information atoms* (PI atoms). Each PI atom corresponds to a distinct way the set of predictors S can contribute information about the target T. Furthermore, Williams and Beer show that these PI atoms are partially ordered and hence form a lattice which they call the *redundancy lattice*. For the bivariate case, the redundancy lattice recovers the decomposition (1), while in the multivariate case it provides a meaningful structure for decomposition of the total information provided by an arbitrary number of predictor variables.

While the redundancy lattice of PID provides a structure for multivariate information decomposition, it does not uniquely determine the value of the PI atoms in the lattice. To do so requires a definition of a measure of redundant information which satisfies the above axioms. Hence, in order to complete the PID framework, Williams and Beer simultaneously introduced a measure of redundant information called $I_{\min}$ which quantifies redundancy as the minimum information that any source provides about a target event t, averaged over all possible events from T. However, not long after its introduction $I_{\min}$ was heavily criticised. Firstly, $I_{\min}$ does not distinguish between "whether different random variables carry the *same* information or just the *same amount* of information" ([5], p. 269; see also [6,7]). Secondly,

$I_{\min}$ does not possess the target chain rule introduced by Bertschinger et al. [5] (under the name left chain rule). This latter point is problematic as the target chain rule is a natural generalisation of the chain rule of mutual information—i.e., one of the fundamental, and indeed characterising, properties of information in Shannon's theory [8,9].

These issues with $I_{\min}$ prompted much research attempting to find a suitable replacement measure compatible with the PID framework. Using the methods of information geometry, Harder et al. [6] focused on a definition of redundant information called I_{red} (see also [10]). Bertschinger et al. [11] defined a measure of unique information $\widetilde{UI}$ based upon the notion that if one variable contains unique information then there must be some way to exploit that information in a decision problem. Griffith and Koch [12] used an entirely different motivation to define a measure of synergistic information $\mathcal{S}_{\mathrm{VK}}$ whose decomposition transpired to be equivalent to that of $\widetilde{UI}$ [11]. Despite this effort, none of these proposed measures are entirely satisfactory. Firstly, just as for $I_{\min}$, none of these proposed measures possess the target chain rule. Secondly, these measures are not compatible with the PID framework in general, but rather are only compatible with PID for the special case of bivariate predictors, i.e., the decomposition (1). This is because they all simultaneously satisfy the Williams and Beers axioms, local positivity, and the *identity property* introduced by Harder et al. [6]. In particular, Rauh et al. [13] proved that no measure satisfying the identity property and the Williams and Beer Axioms 1–3 can yield a non-negative information decomposition beyond the bivariate case of two predictor variables. In addition to these proposed replacements for $I_{\min}$, there is also a substantial body of literature discussing either PID, similar attempts to decompose multivariate information, or the problem of information decomposition in general [3–5,7,10,13–28]. Furthermore, the current proposals have been applied to various problems in neuroscience [29–34]. Nevertheless (to date), there is no generally accepted measure of redundant information that is entirely compatible with PID framework, nor has any other well-accepted multivariate information decomposition emerged.

To summarise the problem, we are seeking a meaningful decomposition of the information provided an arbitrarily large set of predictor variables about a target variable, into atoms of partial information contributed either individually or jointly by the various subsets of the predictors. Crucially, the redundant information must capture when two predictor variables are carrying the same information about the target, not merely the same amount of information. Finally, any proposed measure of redundant information should satisfy the target chain rule so that net redundant information can be consistently computed for consistently for multiple target events.

2. Pointwise Information Theory

Both the entropy and mutual information can be derived from first principles as fundamentally *pointwise* quantities which measure the information content of individual events rather than entire variables. The pointwise entropy $h(t) = -\log p(t)$ quantifies the information content of a single event t, while the pointwise mutual information

$$i(s;t) = \log \frac{p(t|s)}{p(t)} = \log \frac{p(s,t)}{p(s)p(t)} = \log \frac{p(s|t)}{p(s)}, \tag{2}$$

quantifies the information provided by s about t, or vice versa. To our knowledge, these quantities were first considered by Woodward and Davies [35,36] who noted that the average form of Shannon's entropy "tempts one to enquire into other simpler methods of derivation [of the per state entropy]" ([35], p. 51). Indeed, they went on to show that the pointwise entropy and pointwise mutual information can be derived from two axioms concerning the addition of the information provided by the occurrence of individual events [36]. Fano [9] further formalised this pointwise approach by deriving both quantites from four postulates which "should be satisfied by a useful measure of information" ([9], p. 31). Taking the expectation of these pointwise quantities over all events recovers the average entropy $H(T) = \langle h(t) \rangle$ and average mutual information $I(S;T) = \langle i(s;t) \rangle$ first derived by Shannon [8]. Although both approaches arrive at the same average quantities, Shannon's treatment

obfuscates the pointwise nature of the fundamental quantities. In contrast, the approach of Woodward, Davis and Fano makes this pointwise nature manifestly obvious.

It is important to note that, in contrast to the average mutual information, the pointwise mutual information is not non-negative. Positive pointwise information corresponds to the predictor event s raising the probability $p(t|s)$ relative to the prior probability $p(t)$. Hence when the event t occurs it can be said that the event s was *informative* about the event t. Conversely, negative pointwise information corresponds to the event s lowering the posterior probability $p(t|s)$ relative to the prior probability $p(t)$. Hence when the event t occurs we can say that the event s was *misinformative* about the event t. (Not to be confused with disinformation, i.e., intentionally misleading information.) Although a source event s may be misinformative about a particular target event t, a source event s is never misinformative about the target variable T since the pointwise mutual information averaged over all target realisations is non-negative [9]. The information provided by s is helpful for predicting T on average; however, in certain instances this (typically helpful) information is misleading in that it lowers $p(t|s)$ relative to $p(t)$—typically helpful information which subsequently turns out to be misleading is misinformation.

Finally, before continuing, there are two points to be made about the terminology used to describe pointwise information. Firstly, in certain literature (typically in the context of time-series analysis), the word *local* is used instead of pointwise, e.g., [4,18]. Secondly, in contemporary information theory, the word average is generally omitted while the pointwise quantities are explicitly prefixed; however, this was not always the accepted convention. Woodward [35] and Fano [9] both referred to pointwise mutual information as the *mutual information* and then explicitly prefixed the *average mutual information*. To avoid confusion, we will always prefix both pointwise and average quantities.

2.1. Pointwise Information Decomposition

Now that we are familiar with pointwise nature of information, suppose that we have a discrete realisation from the joint event space $\mathcal{T} \times \mathcal{S}_1 \times \mathcal{S}_2$ consisting of the target event t and predictor events s_1 and s_2. The pointwise mutual information $i(s_1; t)$ quantifies the information provided individually by s_1 about t, while the pointwise mutual information $i(s_2; t)$ quantifies the information provided individually by s_2 about t. The pointwise joint mutual information $i(s_{1,2}; t)$ quantifies the total information provided jointly by s_1 and s_2 about t. In correspondence with the (average) bivariate decomposition (1), consider the pointwise bivariate decomposition, first suggested by Lizier et al. [4],

$$
\begin{aligned}
i(s_{1,2}; t) &= r(s_1, s_2 \rightarrow t) + u(s_1 \backslash s_2 \rightarrow t) + u(s_2 \backslash s_1 \rightarrow t) + c(s_1, s_2 \rightarrow t), \\
i(s_1; t) &= r(s_1, s_2 \rightarrow t) + u(s_1 \backslash s_2 \rightarrow t), \\
i(s_2; t) &= r(s_1, s_2 \rightarrow t) + u(s_2 \backslash s_1 \rightarrow t).
\end{aligned}
\tag{3}
$$

Note that the lower case quantities denote the pointwise equivalent of the corresponding upper case quantities in (1). This decomposition could be considered for every discrete realisation on the support of the joint distribution $P(S_1, S_2, T)$. Hence, consider taking the expectation of these pointwise atoms over all discrete realisations,

$$
\begin{aligned}
U(S_1 \backslash S_2 \rightarrow T) &= \langle u(s_1 \backslash s_2 \rightarrow t) \rangle, & R(S_1, S_2 \rightarrow T) &= \langle r(s_1, s_2 \rightarrow t) \rangle, \\
U(S_2 \backslash S_1 \rightarrow T) &= \langle u(s_2 \backslash s_1 \rightarrow t) \rangle, & C(S_1, S_2 \rightarrow T) &= \langle c(s_1, s_2 \rightarrow t) \rangle.
\end{aligned}
\tag{4}
$$

Since the expectation is a linear operation, this will recover the (average) bivariate decomposition (1). Equation (3) for every discrete realisation, together with (1) and (4) form the bivariate pointwise information decomposition. Just as in (1), these equations are underdetermined requiring a separate definition of either the pointwise unique, redundant or complementary information for uniqueness. (Defining an average atom is sufficient for a unique bivariate decomposition (1), but still leaves the pointwise decomposition (3) within each realisation underdetermined).

2.2. Pointwise Unique

Now consider applying this pointwise information decomposition to the probability distribution *Pointwise Unique* (PwUnq) in Table 1. In PwUnq, observing 0 in either of S_1 or S_2 provides zero information about the target T, while complete information about the outcome of T is obtained by observing 1 or a 2 in either predictor. The probability distribution is structured such that in each of the four realisations, one predictor provides complete information while the other predictor provides zero information—the two predictors never provide the same information about the target which is justified by noting that one of the two predictors always provides zero pointwise information.

Given that redundancy is supposed to capture the same information, it seems reasonable to assume there must be zero pointwise redundant information for each realisation. This assumption is made without any measure of pointwise redundant information; however, no other possibility seems justifiable. This assertion is used to determine the pointwise redundant information terms in Table 1. Then using the pointwise information decomposition (3), we can then evaluate the other pointwise atoms of information in Table 1. Finally using (4), we get that there is zero (average) redundant information, and 1/2 bit of (average) unique information from each predictor. From the pointwise perspective, the only reasonable conclusion seems to be that the predictors in PwUnq must contain only unique information about the target.

Table 1. Example PwUnq. For each realisation, the pointwise mutual information provided by each individual and joint predictor events, about the target event has been evaluated. Note that one predictor event always provides full information about the target while the other provides zero information. Based on the this, it is assumed that there must be zero redundant information. The pointwise partial information (PPI) atoms are then calculated via (3).

p	s_1	s_2	t	$i(s_1;t)$	$i(s_2;t)$	$i(s_{1,2};t)$	$u(s_1 \backslash s_2 \to t)$	$u(s_2 \backslash s_1 \to t)$	$r(s_1, s_2 \to t)$	$c(s_1, s_2 \to t)$
1/4	0	1	1	0	1	1	0	1	0	0
1/4	1	0	1	1	0	1	1	0	0	0
1/4	0	2	2	0	1	1	0	1	0	0
1/4	2	0	2	1	0	1	1	0	0	0
Expected values				1/2	1/2	1	1/2	1/2	0	0

However, in contrast to the above, $I_{\min}$, I_{red}, $\widetilde{UI}$, and $\mathcal{S}_{\text{VK}}$ all say that the predictors in PwUnq contain no unique information, rather only 1/2 bit of redundant information plus 1/2 bit of complementary information. This problem, which will be referred to as the *pointwise unique problem*, is a consequence of the fact that these measures all satisfy Assumption (∗) of Bertschinger et al. [11], which (in effect) states that the unique and redundant information should only depend on the marginal distributions $P(S_1, T)$ and $P(S_2, T)$. In particular, any measure which satisfies Assumption (∗) will yield zero unique information when $P(S_1, T)$ is isomorphic to $P(S_2, T)$, as is the case for PwUnq. (Here, isomorphic should be taken to mean isomorphic probability spaces, e.g., [37], p. 27 or [38], p. 4.) It arises because Assumption (∗) (and indeed the operational interpretation the led to its introduction) does not respect the pointwise nature of information. This operational view does not take into account the fact that individual events s_1 and s_2 may provide different information about the event t, even if the probability distributions $P(S_1, T)$ and $P(S_2, T)$ are the same. Hence, we contend that for any measure to capture the same information (not merely the same amount), it must respect the pointwise nature of information.

2.3. Pointwise Partial Information Decomposition

With the pointwise unique problem in mind, consider constructing an information decomposition with the pointwise nature of information as an inherent property. Let $a_1, \ldots, a_k$ be potentially intersecting subsets of the predictor events $s = \{s_1, \ldots, s_n\}$, called source events. Now consider rewriting the Williams and Beer axioms in terms of a measure of pointwise redundant information $i_\cap$ where the aim is to deriving a *pointwise partial information decomposition* (PPID).

PPID Axiom 1 (Symmetry). *Pointwise redundant information is invariant under any permutation σ of source events,*

$$i_\cap(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k \to t) = i_\cap(\sigma(\boldsymbol{a}_1), \ldots, \sigma(\boldsymbol{a}_k) \to T).$$

PPID Axiom 2 (Monotonicity). *Pointwise redundant information decreases monotonically as more source events are included,*

$$i_\cap(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k-1} \to t) \leq i_\cap(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k \to t)$$

with equality if $\boldsymbol{a}_k \supseteq \boldsymbol{a}_i$ for any $\boldsymbol{a}_i \in \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k-1}\}$.

PPID Axiom 3 (Self-redundancy). *Pointwise redundant information for a single source event $\boldsymbol{a}_i$ equals the pointwise mutual information,*

$$i_\cap(\boldsymbol{a}_i \to t) = i(\boldsymbol{a}_i ; t).$$

It seems that the next step should be to define some measure of pointwise redundant information which is compatible with these PPID axioms; however, there is a problem—the pointwise mutual information is not non-negative. While this would not be an issue for the examples like PwUnq, where none of the source events provide negative pointwise information, it is an issue in general (e.g., see RdnErr in Section 5.4). The problem is that set-theoretic intuition behind Axiom 2 (monotonicity) makes little sense when considering signed measures like the pointwise mutual information.

Given the desire to address the pointwise unique problem, there is a need to overcome this issue. Ince [18] suggested that the set-theoretic intuition is only valid when all source events provide either positive or negative pointwise information. Ince contends that information and misinformation are "fundamentally different" ([18], p. 11) and that the set-theoretic intuition should be admitted in the difficult to interpret situations where both are present. We however, will take a different approach—one which aims to deal with these difficult to interpret situations whilst preserving the set-theoretic intuition that redundancy corresponds to overlapping information.

By way of a preview, we first consider precisely how an event s_1 provides information about an event t by the means of two distinct types of probability mass exclusion. We show how considering the process in this way naturally splits the pointwise mutual information into particular entropic components, and how one can consider redundancy on each of these components separately. Splitting the signed pointwise mutual information into these unsigned entropic components circumvents the above issue with Axiom 2 (monotonicity). Crucially, however, by deriving these entropic components from the probability mass exclusions, we retain the set-theoretic intuition of redundancy—redundant information will correspond to overlapping probability mass exclusions in the two-event partition $\mathcal{T}^t = \{t, t^c\}$.

3. Probability Mass Exclusions and the Directed Components of Pointwise Mutual Information

By definition, the pointwise information provided by s about t is associated with a change from the prior $p(t)$ to the posterior $p(t|s)$. As we explored from first principles in Finn and Lizier [39], this change is a consequence of the *exclusion* of probability mass in the target distribution $P(T)$ induced by the occurrence of the event s and inferred via the joint distribution $P(S, T)$. To be specific, when the event s occurs, one knows that the complementary event $s^c = \{S \backslash s\}$ did not occur. Hence one can *exclude* the probability mass in the joint distribution $P(S, T)$ associated with the complementary event, i.e., exclude $P(s^c, T)$, leaving just the probability mass $P(s, T)$ remaining. The new target distribution $P(T|s)$ is evaluated by normalising this remaining probability mass. In [39] we introduced *probability mass diagrams* in order to visually explore the exclusion process. Figure 1 provides an example of such a diagram. Clearly, this process is merely a description of the definition of conditional probability. Nevertheless, we content that by viewing the change from the prior to the posterior in this way—by focusing explicitly on the exclusions rather than the resultant conditional probability—the vague intuition that redundancy corresponds to overlapping information becomes more apparent. This point will elaborated upon in Section 3.3. However, in order to do so, we need to first discuss the two distinct types of probability mass exclusion (which we do in Section 3.1) and then relate these to information-theoretic quantities (which we do in Section 3.2).
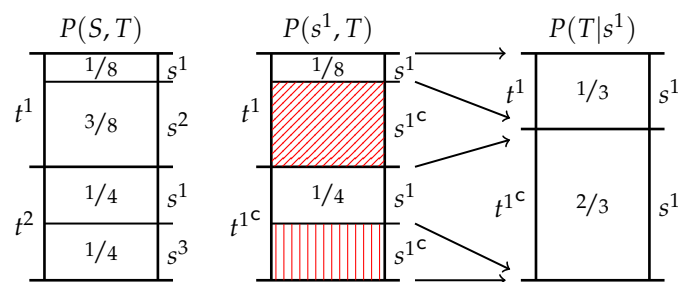
Figure 1. Sample probability mass diagrams, which use length to represent the probability mass of each joint event from $\mathcal{T} \times \mathcal{S}$. (**Left**) the joint distribution $P(S, T)$; (**Middle**) The occurrence of the event s^1 leads to exclusions of the complementary event s^{1^c} which consists of two elementary event, i.e., $s^{1^c} = \{s^2, s^3\}$. This leaves the probability mass $P(s^1, T)$ remaining. The exclusion of the probability mass $p(s^{1^c}, t^1)$ was misinformative since the event t^1 did occur. By convention, misinformative exclusions will be indicated with diagonal hatching. On the other hand, the exclusion of the probability mass $p(t^{1^c}, s^{1^c})$ was informative since the complementary event t^{1^c} did not occur. By convention, informative exclusions will be indicated with horizontal or vertical hatching; (**Right**) this remaining probability mass can be normalised yielding the conditional distribution $P(T|s^1)$.

3.1. Two Distinct Types of Probability Mass Exclusions

In [39] we examined the two distinct types of probability mass exclusions. The difference between the two depends on where the exclusion occurs in the target distribution $P(T)$ and the particular target event t which occurred. *Informative exclusions* are those which are confined to the probability mass associated with the set of elementary events in the target distribution which *did not occur*, i.e., exclusions confined to the probability mass of the complementary event $p(t^c)$. They are called such because the pointwise mutual information $i(s; t)$ is a monotonically increasing function of the total size of these exclusions $p(t^c)$. By convention, informative exclusions are represented on the probability mass diagrams by horizontal or vertical lines. On the other hand, the *misinformative exclusion* is confined to the probability mass associated with the elementary event in the target distribution which *did occur*, i.e., an exclusion confined to $p(t)$. It is referred to as such because the pointwise mutual information $i(s; t)$ is a monotonically decreasing function of the size of this type of exclusion $p(t)$. By convention, misinformative exclusions are represented on the probability mass diagrams by diagonal lines.

Although an event s may exclusively induce either type of exclusion, in general both types of exclusion are present simultaneously. The distinction between the two types of exclusions leads naturally to the following question—can one decompose the pointwise mutual information $i(s; t)$ into a positive informational component associated with the informative exclusions, and a negative informational component associated with the misinformative exclusions? This question is considered in detail in Section 3.2. However, before moving on, there is a crucial observation to be made about the pointwise mutual information which will have important implications for the measure of redundant information to be introduced later.

Remark 1. *The pointwise mutual information $i(s; t)$ depends only on the size of informative and misinformative exclusions. In particular, it does not depend on the apportionment of the informative exclusions across the set of elementary events contained in the complementary event t^c.*

In other words, whether the event s turns out to be net informative or misinformative about the event t—whether $i(s; t)$ is positive or negative—depends on the size of the two types of exclusions; but, to be explicit, does *not* depend on the distribution of the informative exclusion across the set of target events which did not occur. This remark will be crucially important when it comes to providing the operational interpretation of redundant information in Section 3.3. (It is also further discussed in terms of Kelly gambling [40] in Appendix A).

3.2. The Directed Components of Pointwise Information: Specificity and Ambiguity

We return now to the idea that one might be able to decompose the pointwise mutual information into a positive and negative component associated with the informative amd misinformative exclusions respectively. In [39] we proposed four postulates for such a decomposition. Before stating the postulates, it is important to note that although there is a "surprising symmetry" ([41], p. 23) between the information provided by *s* about *t* and the information provided by *t* about *s*, there is nothing to suggest that the components of the decomposition should be symmetric—indeed the intuition behind the decomposition only makes sense when considering the information is considered in a directed sense. As such, directed notation will be used to explicitly denote the information provided by *s* about *t*.

Postulate 1 (Decomposition). *The pointwise information provided by s about t can be decomposed into two non-negative components, such that $i(s;t) = i_+(s \to t) - i_-(s \to t)$.*

Postulate 2 (Monotonicity). *For all fixed $p(s,t)$ and $p(s^c,t)$, the function $i_+(s \to t)$ is a monotonically increasing, continuous function of $p(t^c,s^c)$. For all fixed $p(t^c,s)$ and $p(t^c,s^c)$, the function $i_-(s \to t)$ is a monotonically increasing continuous function of $p(s^c,t)$. For all fixed $p(s,t)$ and $p(t^c,s)$, the functions $i_+(s \to t)$ and $i_-(s \to t)$ are monotonically increasing and decreasing functions of $p(t^c,s^c)$, respectively.*

Postulate 3 (Self-Information). *An event cannot misinform about itself, $i_+(s \to s) = i(s;s) = -\log p(s)$.*

Postulate 4 (Chain Rule). *The functions $i_+(s_{1,2} \to t)$ and $i_-(s_{1,2} \to t)$ satisfy a chain rule, i.e.,*

$$
\begin{aligned}
i_+(s_{1,2} \to t) &= i_+(s_1 \to t) + i_+(s_2 \to t|s_1) \\
&= i_+(s_2 \to t) + i_+(s_1 \to t|s_2), \\
i_-(s_{1,2} \to t) &= i_-(s_1 \to t) + i_-(s_2 \to t|s_1) \\
&= i_-(s_2 \to t) + i_-(s_1 \to t|s_2)
\end{aligned}
$$

In Finn and Lizier [39], we proved that these postulates lead to the following forms which are unique up to the choice of the base of the logarithm in the mutual information in Postulates 1 and 3,

$$
\begin{aligned}
i^+(s_1 \to t) &= h(s_1) &&= -\log p(s_1), & (5) \\
i^+(s_1 \to t|s_2) &= h(s_1|s_2) &&= -\log p(s_1|s_2), & (6) \\
i^+(s_{1,2} \to t) &= h(s_{1,2}) &&= -\log p(s_{1,2}), & (7) \\
i^-(s_1 \to t) &= h(s_1|t) &&= -\log p(s_1|t), & (8) \\
i^-(s_1 \to t|s_2) &= h(s_1|t,s_2) &&= -\log p(s_1|t,s_2), & (9) \\
i^-(s_{1,2} \to t) &= h(s_{1,2}|t) &&= -\log p(s_{1,2}|t). & (10)
\end{aligned}
$$

That is, the Postulates 1–4 uniquely decompose the pointwise information provided by *s* about *t* into the following entropic components,

$$
\begin{aligned}
i(s;t) &= i^+(s \to t) - i^-(s \to t) \\
&= h(s) - h(s|t).
\end{aligned}
\tag{11}
$$

Although the decomposition of mutual information into entropic components is well-known, it is non-trivial that Postulates 1 and 3, based on the size of the two distinct types of probability mass exclusions, lead to this particular form, but not $i(s;t) = h(t) - h(t|s)$ or $i(s;t) = h(s) + h(t) - h(s,t)$.

It is important to note that although the original motivation was to decompose the pointwise mutual information into separate components associated with informative and misinformative exclusion, the decomposition (11) does not quite possess this direct correspondence:

- The positive informational component $i^+(s \to t)$ does not depend on *t* but rather only on *s*. This can be interpreted as follows: the less likely *s* is to occur, the more specific it is when it does occur,

the greater the total amount of probability mass excluded $p(s^c)$, and the greater the potential for s to inform about t (or indeed any other target realisation).
- The negative informational component $i^-(s \rightarrow t)$ depends on both s and t, and can be interpreted as follows: the less likely s is to coincide with the event t, the more uncertainty in s given t, the greater size of the misinformative probability mass exclusion $p(s^c, t)$, and therefore the greater the potential for s to misinform about t.

In other words, although the negative informational component $i^-(s \rightarrow t)$ does correspond directly to the size of the misinformative exclusion $p(s^c, t)$, the positive informational component $i^+(s \rightarrow t)$ does not correspond directly to the size of the informative exclusion $p(t^c, s^c)$. Rather, the positive informational component $i^+(s \rightarrow t)$ corresponds to the *total* size of the probability mass exclusions $p(s^c)$, which is the sum of the sum of the informative and misinformative exclusions. For the sake of brevity, the positive informational component $i^+(s \rightarrow t)$ will be referred to as the *specificity*, while the negative informational component $i^-(s \rightarrow t)$ will be referred to as the *ambiguity*. The term ambiguity is due to Shannon: "[equivocation] measures the average ambiguity of the received signal" ([42], p. 67). Specificity is an antonym of ambiguity and the usage here is inline with the definition since the more specific an event s, the more information it could provide about t after the ambiguity is taken into account.

3.3. Operational Interpretation of Redundant Information

Arguing about whether one piece of information differs from another piece of information is nonsensical without some kind of unambiguous definition of what it means for two pieces of information to be the same. As such, Bertschinger et al. [11] advocate the need to provide an operational interpretation of what it means for information to be unique or redundant. This section provides our operational definition of what it means for information to be the same. This definition provides a concrete interpretation of what it means for information to be redundant in terms of overlapping probability mass exclusions.

The operational interpretation of redundancy adopted here is based upon the following idea: since the pointwise information is ultimately derived from probability mass exclusions, the *same information* must induce the *same exclusions*. More formally, the information provided by a set of predictor events $s_1, \ldots, s_k$ about a target event t must be the same information if each source event induces the same exclusions with respect to the two-event partition $\mathcal{T}^t = \{t, t^c\}$. While this statement makes the motivational intuition clear, it is not yet sufficient to serve as an operational interpretation of redundancy: there is no reference to the two distinct types of probability mass exclusions, the specific reference to the pointwise event space $\mathcal{T}^t$ has not been explained, and there is no reference to the fact the exclusions from each source may differ in size.

Informative exclusions are fundamentally different from misinformative exclusions and hence each type of exclusion should be compared separately: informative exclusions can overlap with informative exclusions, and misinformative exclusions can overlap with misinformative exclusions. In information-theoretic terms, this means comparing the specificity and the ambiguity of the sources separately—i.e., considering a measure of redundant specificity and a separate measure of redundant ambiguity. Crucially, these quantities (being pointwise entropies) are unsigned meaning that the difficulties associated with Axiom 2 (Monotonicity) and signed pointwise mutual information in Section 2.3 will not be an issue here.

The specific reference to the two-event partition $\mathcal{T}^t$ in the above statement is based upon Remark 1 and is crucially important. The pointwise mutual information does not depend on the apportionment of the informative exclusions across the set of events which did not occur, hence the pointwise redundant information should not depend on this apportionment either. In other words, it is immaterial if two predictor events s_1 and s_2 exclude different elementary events within the target complementary event t^c (assuming the probability mass excluded is equal) since with respect to the realised target event t the difference between the exclusions is only semantic. This has important implications for

the comparison of exclusions from different predictor events. As the pointwise mutual information depends on, and only depends on, the size of the exclusions, then the only sensible comparison is a comparison of size. Hence, the common or overlapping exclusion must be the smallest exclusion. Thus, consider the following operational interpretation of redundancy:

Operational Interpretation (Redundant Specificity). *The redundant specificity between a set of predictor events $s_1, \ldots, s_n$ is the specificity associated with the source event which induces the smallest total exclusions.*

Operational Interpretation (Redundant Ambiguity). *The redundant ambiguity between a set of predictor events $s_1, \ldots, s_n$ is the ambiguity associated with the source event which induces the smallest misinformative exclusion.*

3.4. Motivational Example

To motivate the above operational interpretation, and in particular the need to treat the specificity separately to the ambiguity, consider Figure 2. In this pointwise example, two different predictor events provide the same amount of pointwise information since $P(T|s_1^1) = P(T|s_2^1)$, and yet the information provided by each event is in some way different since each excludes different sections of the target distribution $P(T)$. In particular, s_1^1 and s_2^1 both preclude the target event t^2, while s_2^1 additionally excludes probability mass associated with target events t^1 and t^3. From the perspective of the pointwise mutual information the events s_1^1 and s_2^1 seem to be providing the same information as

$$i(s_1^1 \rightarrow t^1) = i(s_2^1 \rightarrow t^1) = \log 4/3 \text{ bit.} \tag{12}$$

However, from the perspective of the specificity and the ambiguity it can be seen that information is being provided in different ways since

$$\begin{aligned} i^+(s_1^1 \rightarrow t^1) &= \log 4/3 \text{ bit}, & i^-(s_1^1 \rightarrow t^1) &= 0 \text{ bit}, \\ i^+(s_2^1 \rightarrow t^1) &= \log 8/3 \text{ bit}, & i^-(s_2^1 \rightarrow t^1) &= 1 \text{ bit}. \end{aligned} \tag{13}$$

Now consider the problem of decomposing information into its unique, redundant and complementary components. Figure 2 shows where exclusions induced by s_1^1 and s_2^1 overlap where they both exclude the target event t^2 which is an informative exclusion. This is the only exclusion induced by s_1^1 and hence all of the information associated with this exclusion must be redundantly provided by the event s_2^1. Without any formal framework, consider taking the redundant specificity and redundant ambiguity,

$$r^+(s_1^1, s_2^1 \rightarrow t^1) = i^+(s_1^1 \rightarrow t^1) = \log 4/3 \text{ bit}, \tag{14}$$

$$r^-(s_1^1, s_2^1 \rightarrow t_1) = i^-(s_1^1 \rightarrow t^1) = \quad 0 \text{ bit}. \tag{15}$$

This would mean that the event s_2^1 provides the following unique specificity and unique ambiguity,

$$u^+(s_1^1 \backslash s_2^1 \rightarrow t^1) = i^+(s_1^1 \rightarrow t^1) - r^+(s_1^1, s_2^1 \rightarrow t^1) = 1 \text{ bit}, \tag{16}$$

$$u^-(s_1^1 \backslash s_2^1 \rightarrow t^1) = i^-(s_1^1 \rightarrow t^1) - r^-(s_1^1, s_2^1 \rightarrow t^1) = 1 \text{ bit}. \tag{17}$$

The redundant specificity $\log 4/3$ bit accounts for the overlapping informative exclusion of the event t^2. The unique specificity and unique ambiguity from s_2^1 are associated with its non-overlapping informative and misinformative exclusions; however, both of these 1 bit and hence, on net, s_2^1 is no more informative than s_1^1. Although obtained without a formal framework, this example highlights a need to consider the specificity and ambiguity rather than merely the pointwise mutual information.
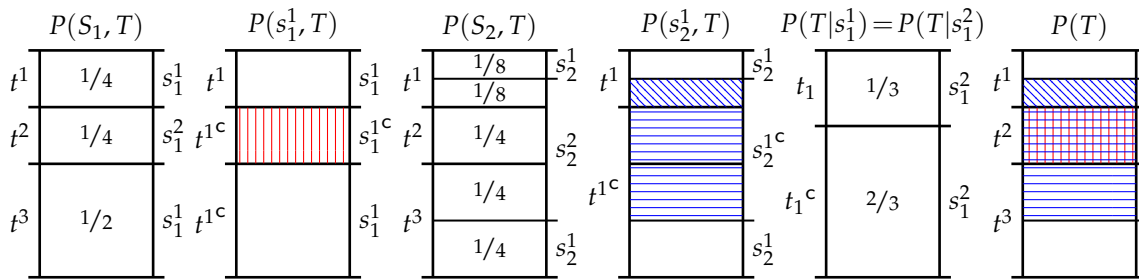
Figure 2. Sample probability mass diagrams for two predictors S_1 and S_2 to a given target T. Here events in the two different predictor spaces provide the same amount of pointwise information about the target event, $\log_2 {}^4\!/_3$ bits, since $P(T|s_1^1) = P(T|s_2^1)$, although each excludes different sections of the target distribution $P(T)$. Since they both provide the same amount of information, is there a way to characterise what information the additional unique exclusions from the event s_2^1 are providing?

4. Pointwise Partial Information Decomposition Using Specificity and Ambiguity

Based upon the argumentation of Section 3, consider the following axioms:

Axiom 1 (Symmetry). *Pointwise redundant specificity $i_\cap^+$ and pointwise redundant ambiguity $i_\cap^-$ are invariant under any permutation σ of source events,*

$$i_\cap^+(\mathbf{a}_1, \ldots, \mathbf{a}_k \to t) = i_\cap^+(\sigma(\mathbf{a}_1), \ldots, \sigma(\mathbf{a}_k) \to t),$$
$$i_\cap^-(\mathbf{a}_1, \ldots, \mathbf{a}_k \to t) = i_\cap^-(\sigma(\mathbf{a}_1), \ldots, \sigma(\mathbf{a}_k) \to t).$$

Axiom 2 (Monotonicity). *Pointwise redundant specificity $i_\cap^+$ and pointwise redundant ambiguity $i_\cap^-$ decreases monotonically as more source events are included,*

$$i_\cap^+(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}, \mathbf{a}_k \to t) \le i_\cap^+(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1} \to t),$$
$$i_\cap^-(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}, \mathbf{a}_k \to t) \le i_\cap^-(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1} \to t).$$

with equality if $\mathbf{a}_k \supseteq \mathbf{a}_i$ for any $\mathbf{a}_i \in \{\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}\}$.

Axiom 3 (Self-redundancy). *Pointwise redundant specificity $i_\cap^+$ and pointwise redundant ambiguity $i_\cap^-$ for a single source event $\mathbf{a}_i$ equals the specificity and ambiguity respectively,*

$$i_\cap^+(\mathbf{a}_i \to t) = i^+(\mathbf{a}_i \to t) = h(\mathbf{a}_i),$$
$$i_\cap^-(\mathbf{a}_i \to t) = i^-(\mathbf{a}_i \to t) = h(\mathbf{a}_i|t).$$

As shown in Appendix B.1, Axioms 1–3 induce two lattices—namely the *specificity lattice* and *ambiguity lattice*—which are depicted in Figure 3. Furthermore, each lattice is defined for every discrete realisation from $P(S_1, \ldots, S_n, T)$. The redundancy measures $i_\cap^+$ or $i_\cap^-$ can be thought of as a cumulative information functions which integrate the specificity or ambiguity uniquely contributed by each node as one moves up each lattice. Finally, just as in PID, performing a Möbius inversion over each lattice yielding the unique contributions of specificity and ambiguity from each sources event.

Similarly to PID, the specificity and ambiguity lattices provide a structure for information decomposition, but unique evaluation requires a separate definition of redundancy. However, unlike PID (or even PPID), this evaluation requires both a definition of pointwise redundant specificity and pointwise redundant ambiguity. Before providing these definitions, it is helpful to first see how the specificity and ambiguity lattices can be used to decompose multivariate information in the now familiar bivariate case.
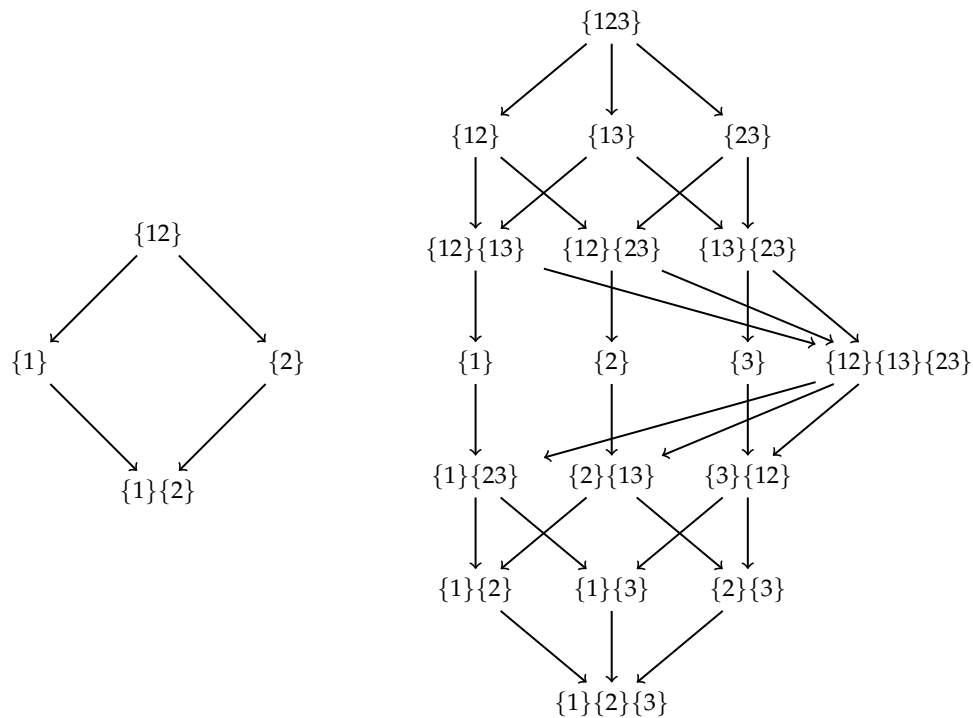
Figure 3. The lattice induced by the partial order $\preceq$ (A15) over the sources $\mathscr{A}(s)$ (A14). (**Left**) the lattice for $s = \{s_1, s_2\}$; (**Right**) the lattice for $s = \{s_1, s_2, s_3\}$. See Appendix B for further details. Each node corresponds to the self-redundancy (Axiom 3) of a source event, e.g., $\{1\}$ corresponds to the source event $\{\{s_1\}\}$, while $\{12, 13\}$ corresponds to the source event $\{\{s_{1,2}\}, \{s_{1,3}\}\}$. Note that the specificity and ambiguity lattices share the same structure as the redundancy lattice of partial information decomposition (PID) (cf. Figure 2 in [1]).

4.1. Bivariate PPID Using the Specificity and Ambiguity

Consider again the bivariate case where the aim is to decompose the information provided by s_1 and s_2 about t. The specificity lattice can be used to decompose the pointwise specificity,

$$
\begin{aligned}
i^+(s_{1,2} \to t) &= r^+(s_1, s_2 \to t) + u^+(s_1 \backslash s_2 \to t) + u^+(s_2 \backslash s_1 \to t) + c^+(s_1, s_2 \to t), \\
i^+(s_1 \to t) &= r^+(s_1, s_2 \to t) + u^+(s_1 \backslash s_2 \to t), \\
i^+(s_2 \to t) &= r^+(s_1, s_2 \to t) + u^+(s_2 \backslash s_1 \to t);
\end{aligned}
\tag{18}
$$

while the ambiguity lattice can be used to decompose the pointwise ambiguity,

$$
\begin{aligned}
i^-(s_{1,2} \to t) &= r^-(s_1, s_2 \to t) + u^-(s_1 \backslash s_2 \to t) + u^-(s_2 \backslash s_1 \to t) + c^-(s_1, s_2 \to t), \\
i^-(s_1 \to t) &= r^-(s_1, s_2 \to t) + u^-(s_1 \backslash s_2 \to t), \\
i^-(s_2 \to t) &= r^-(s_1, s_2 \to t) + u^-(s_2 \backslash s_1 \to t).
\end{aligned}
\tag{19}
$$

These equations share the same structural form as (3) only now decompose the specificity and the ambiguity rather than the pointwise mutual information, e.g., $r^+(s_1, s_2 \to t)$ denotes the redundant specificity while $u^-(s_1 \backslash s_2 \to t)$ denoted the unique ambiguity from s_1. Just as in for (3), this decomposition could be considered for every discrete realisation on the support of the joint distribution $P(S_1, S_2, T)$.

There are two ways one can be combine these values. Firstly, in a similar manner to (4), one could take the expectation of the atoms of specificity, or the atoms of ambiguity, over all discrete realisations yielding the average PI atoms of specificity and ambiguity,

$$
\begin{aligned}
U^+(S_1\backslash S_2\to T) &= \langle u^+(s_1\backslash s_2\to t)\rangle, & U^-(S_1\backslash S_2\to T) &= \langle u^-(s_1\backslash s_2\to t)\rangle, \\
U^+(S_2\backslash S_1\to T) &= \langle u^+(s_2\backslash s_1\to t)\rangle, & U^-(S_2\backslash S_1\to T) &= \langle u^-(s_2\backslash s_1\to t)\rangle, \\
R^+(S_1,S_2\to T) &= \langle r^+(s_1,s_2\to t)\rangle, & R^-(S_1,S_2\to T) &= \langle r^-(s_1,s_2\to t)\rangle, \\
C^+(S_1,S_2\to T) &= \langle c^+(s_1,s_2\to t)\rangle. & C^-(S_1,S_2\to T) &= \langle c^-(s_1,s_2\to t)\rangle.
\end{aligned}
\tag{20}
$$

Alternatively, one could subtract the pointwise unique, redundant and complementary ambiguity from the pointwise unique, redundant and complementary specificity yielding the pointwise unique, pointwise redundant and pointwise complementary information, i.e., recover the atoms from PPID,

$$
\begin{aligned}
r(s_1,s_2\to t) &= r^+(s_1,s_2\to t) - r^-(s_1,s_2\to t), \\
u(s_1\backslash s_2\to t) &= u^+(s_1\backslash s_2\to t) - u^-(s_1\backslash s_2\to t), \\
u(s_2\backslash s_1\to t) &= u^+(s_2\backslash s_1\to t) - u^-(s_2\backslash s_1\to t), \\
c(s_1,s_2\to t) &= c^+(s_1,s_2\to t) - c^-(s_1,s_2\to t).
\end{aligned}
\tag{21}
$$

Both (20) and (21) are linear operations, hence one could perform both of these operations (in either order) to obtain the average unique, average redundant and average complementary information, i.e., recover the atoms from PID,

$$
\begin{aligned}
R(S_1,S_2\to T) &= R^+(S_1,S_2\to T) - R^-(S_1,S_2\to T), \\
U(S_1\backslash S_2\to T) &= U^+(S_1\backslash S_2\to T) - U^-(S_1\backslash S_2\to T), \\
U(S_2\backslash S_1\to T) &= U^+(S_2\backslash S_1\to T) - U^-(S_2\backslash S_1\to T), \\
C(S_1,S_2\to T) &= C^+(S_1,S_2\to T) - C^-(S_1,S_2\to T).
\end{aligned}
\tag{22}
$$

4.2. Redundancy Measures on the Specificity and Ambiguity Lattices

Now that we have a structure for our information decomposition, there is a need to provide a definition of the pointwise redundant specificity and pointwise redundant ambiguity. However, before attempting to provide such a definition, there is a need to consider Remark 1 and the operational interpretation of in Section 3.3. In particular, the pointwise redundant specificity $i_\cap^+$ and pointwise redundant ambiguity $i_\cap^-$ should only depend on the size of informative and misinformative exclusions. They should not depend on the apportionment of the informative exclusions across the set of elementary events contained in the complementary event t^c. Formally, this requirement will be enshrined via the following axiom.

Axiom 4 (Two-event Partition). *The pointwise redundant specificity $i_\cap^+$ and pointwise redundant ambiguity $i_\cap^-$ are functions of the probability measures on the two-event partitions $\mathcal{A}_1^{a_1}\times\mathcal{T}^t,\dots,\mathcal{A}_k^{a_k}\times\mathcal{T}^t$.*

Since the pointwise redundant specificity $i_\cap^+$ is specificity associated with the source event which induces the smallest total exclusions, and pointwise redundant ambiguity $i_\cap^-$ is the ambiguity associated with the source event which induces the smallest misinformative exclusion, consider the following definitions.

Definition 1. *The pointwise redundant specificity is given by*

$$r^+_{min}(a_1, \ldots, a_k \rightarrow t) = \min_{a_i} i^+(a_i \rightarrow t) = \min_{a_i} h(a_i). \tag{23}$$

Definition 2. *The pointwise redundant ambiguity is given by*

$$r^-_{min}(a_1, \ldots, a_k \rightarrow t) = \min_{a_i} i^-(a_i \rightarrow t) = \min_{a_j} h(a_j|t). \tag{24}$$

Theorem 1. *The definitions of r^+_{min} and r^-_{min} satisfy Axioms 1–4.*

Theorem 2. *The redundancy measures r^+_{min} and r^-_{min} increase monotonically on the $\langle \mathscr{A}(s), \preceq \rangle$.*

Theorem 3. *The atoms of partial specificity π^+ and partial ambiguity π^- evaluated using the measures r^+_{min} and r^-_{min} on the specificity and ambiguity lattices (respectively), are non-negative.*

Appendix B.2 contains the proof of Theorems 1–3 and further relevant consideration of Definitions 1 and 2. As in (20), one can take the expectation of the either the pointwise redundant specificity r^+_{min} or the pointwise redundant ambiguity r^-_{min} to get the average redundant specificity R^+_{min} or the average redundant ambiguity R^-_{min}. Alternatively, just as in (21), one can recombine the pointwise redundant specificity r^+_{min} and the pointwise redundant ambiguity r^-_{min} to get the pointwise redundant information r_{min}. Finally, as per (22), one could perform both of these (linear) operations in either order to obtain the average redundant information R_{min}. Note that while Theorem 3 proves that the atoms of partial specificity π^+ and partial ambiguity π^- are non-negative, it is trivial to see that r_{min} could be negative since when source events can redundantly provide misinformation about a target event. As shown in the following theorem, R_{min} can also be negative.

Theorem 4. *The atoms of partial average information Π evaluated by recombining and averaging $\pi^{\pm}$ are not non-negative.*

This means that the measure R_{min} does not satisfy local positivity. Nonetheless the negativity of R_{min} is readily explainable in terms of the operational interpretation of Section 3.3, as will be discussed further in Section 5.4. However, failing to satisfy local positivity does mean that r_{min} and R_{min} do not satisfy the *target monotonicity* property first discussed in Bertschinger et al. [5]. Despite this, as the following theorem shows, the measures do satisfy the target chain rule.

Theorem 5 (Pointwise Target Chain Rule)**.** *Given the joint target realisation $t_{1,2}$, the pointwise redundant information r_{min} satisfies the following chain rule,*

$$\begin{aligned} r_{min}(a_1, \ldots, a_k \rightarrow t_{1,2}) &= r_{min}(a_1, \ldots, a_k \rightarrow t_1) + r_{min}(a_1, \ldots, a_k \rightarrow t_2|t_1), \\ &= r_{min}(a_1, \ldots, a_k \rightarrow t_2) + r_{min}(a_1, \ldots, a_k \rightarrow t_1|t_2). \end{aligned} \tag{25}$$

The proof of the last theorem is deferred to Appendix B.3. Note that since the expectation is a linear operation, Theorem 5 also holds for the average redundant information R_{min}. Furthermore, as these results apply to any of the source events, the target chain rule will hold for any of the PPI atoms, e.g., (21), and any of the PI atoms, e.g., (22). However, no such rule holds for the pointwise redundant specificity or ambiguity. The specificity depends only on the predictor event, i.e., does not depend on the target events. As such, when an increasing number of target events are considered, the specificity remains unchanged. Hence, a target chain rule cannot hold for the specificity, or the ambiguity alone.

5. Discussion

PPID using the specificity and ambiguity takes the ideas underpinning PID and applies them on a pointwise scale while circumventing the monotonicity issue associated with the signed pointwise mutual information. This section will explore the various properties of the decomposition in an example driven manner and compare the results to the most widely-used measures from the existing PID literature. (Further examples can be found in Appendix C.) The following shorthand notation will be utilised in the figures throughout this section:

$$i_1^+ = i^+(s_1 \rightarrow t), \qquad i_2^+ = i^+(s_2 \rightarrow t), \qquad i_{1,2}^+ = i^+(s_{1,2} \rightarrow t),$$
$$i_1^- = i^-(s_1 \rightarrow t), \qquad i_2^- = i^-(s_2 \rightarrow t), \qquad i_{1,2}^- = i^-(s_{1,2} \rightarrow t),$$

$$u_1^+ = u^+(s_1 \backslash s_2 \rightarrow t), \quad u_2^+ = u^+(s_2 \backslash s_1 \rightarrow t), \quad r^+ = r^+(s_1, s_2 \rightarrow t), \quad c^+ = c^+(s_1, s_2 \rightarrow t),$$
$$u_1^- = u^-(s_1 \backslash s_2 \rightarrow t), \quad u_2^- = u^-(s_2 \backslash s_1 \rightarrow t), \quad r^- = r^-(s_1, s_2 \rightarrow t), \quad c^- = c^-(s_1, s_2 \rightarrow t).$$

5.1. Comparison to Existing Measures

A similar approach to the decomposition presented in this paper is due to Ince [18], who also sought to define a pointwise information decomposition. Despite the similarity in this regard, the redundancy measure I_{CCS} presented in [18] approaches the pointwise monotonicity problem of Section 2.3 in a different way to the decomposition presented in this paper. Specifically, I_{CCS} aims to utilise the pointwise co-information as a measure of pointwise redundant information since it "quantifies the set-theoretic overlap of the two univariate [pointwise] information values" ([18], p. 14). There are, however, difficulties with this approach. Firstly (unlike the average mutual information and the Shannon inequalities), there are no inequalities which support this interpretation of pointwise co-information as the set-theoretic overlap of the univariate pointwise information terms—indeed, both the univariate pointwise information and the pointwise co-information are signed measures. Secondly, the pointwise co-information conflates the pointwise redundant information with the pointwise complementary information, since by (3) we have that

$$co\text{-}i(s_1; s_2; t) := i(s_1; t) + i(s_2; t) - i(s_{1,2}, t) = r(s_1, s_2 \rightarrow t) - c(s_1, s_2 \rightarrow t). \qquad (26)$$

Aware of these difficulties, Ince defines I_{CCS} such that it only interprets the pointwise co-information as a measure of set-theoretic overlap in the case where all three pointwise information terms have the same sign, arguing that these are the only situations which admit a clear interpretation in terms of a common change in surprisal. In the other difficult to interpret situations, I_{CCS} defines the pointwise redundant information to be zero. This approach effectively assumes that $c(s_1, s_2 \rightarrow t) = 0$ in (26) when $i(s_1; t)$, $i(s_2; t)$ and $co\text{-}i(s_1; s_2; t)$ all have the same sign.

In a subsequent paper, Ince [19] also presented a partial entropy decomposition which aims to decompose multivariate entropy rather than multivariate information. As such, this decomposition is more similar to PPID using specificity and ambiguity than Ince's aforementioned decomposition. Although similar in this regard, the measure of pointwise redundant entropy H_{cs} presented in [19] takes a different approach to the one presented in this paper. Specifically, H_{cs} also uses the pointwise co-information as a measure of set-theoretic overlap and hence as a measure of pointwise redundant entropy. As the pointwise entropy is unsigned, the difficulties are reduced but remain present due to the signed pointwise co-information. In a manner similar to I_{CCS}, Ince defines H_{cs} such that it only interprets the pointwise co-information as a measure of set-theoretic overlap when it is positive. As per I_{CCS}, this effectively assumes that $c(s_1, s_2 \rightarrow t) = 0$ in (26) when all information terms have the same sign. When the pointwise co-information is negative, H_{cs} simply ignores the co-information by defining the pointwise redundant information to be zero. In contrast to both of Ince's approaches, PPID using specificity and ambiguity does not dispose of the set-theoretic intuition in these difficult to interpret situations. Rather, our approach considers the notion of redundancy in terms of overlapping exclusions—i.e., in terms of the underlying, unsigned measures which are amenable to a set-theoretic interpretation.
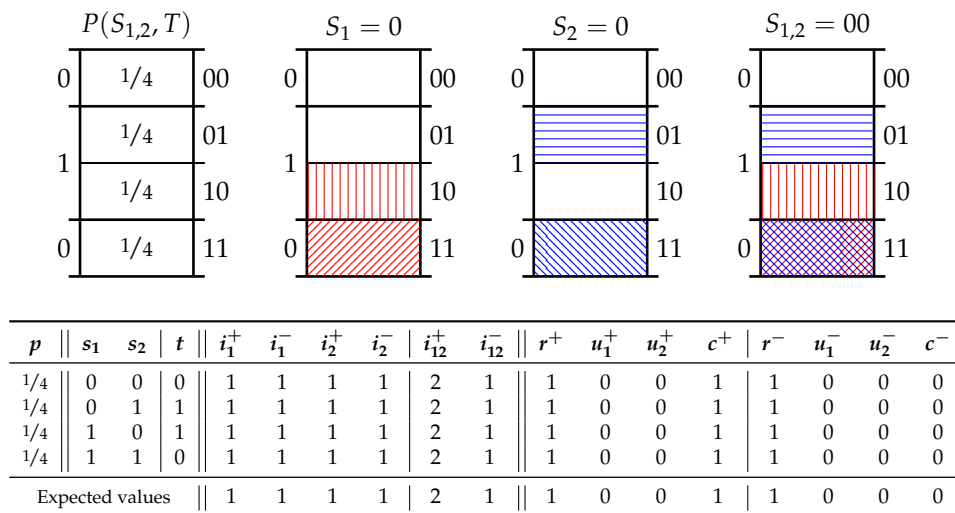
The measures of pointwise redundant specificity $r_{\min}^+$ and pointwise redundant ambiguity $r_{\min}^-$, from Definitions 1 and 2 are also similar to both the minimum mutual information I_{MMI} [17] and the original PID redundancy measure $I_{\min}$ [1]. Specifically, all three of these approaches consider the redundant information to be the minimum information provided about a target event t. The difference is that $I_{\min}$ applies this idea to the sources $A_1, \ldots, A_k$, i.e., to collections of entire predictor variables from S, while $r_{\min}^\pm$ apply this notion to the source events $a_1, \ldots, a_k$, i.e., to collections of predictor events from s. In other words, while the measure $I_{\min}$ can be regarded as being semi-pointwise (since it considers the information provided by the variables $S_1, \ldots, S_n$ about an event t), the measures $r_{\min}^\pm$ are fully pointwise (since they consider the information provided by the events $s_1, \ldots, s_n$ about an event t). This difference in approach is most apparent in the probability distribution PWUNQ—unlike PID, PPID using the specificity and ambiguity respects the pointwise nature of information, as we will see in Section 5.3.

PPID using specificity and ambiguity also share certain similarities with the bivariate PID induced by the measure $\widetilde{UI}$ of Bertschinger et al. [11]. Firstly, Axiom 4 can be considered to be a pointwise adaptation of their Assumption (∗), i.e., the measures $r_{\min}^\pm$ depend only on the marginal distributions $P(S_1, T)$ and $P(S_2, T)$ with respect to the two-event partitions $\mathcal{S}_1^{s_1} \times \mathcal{T}^t$ and $\mathcal{S}_2^{s_2} \times \mathcal{T}^t$. Secondly, in PPID using specificity and ambiguity, the only way one can only decide if there is complementary information $c(s_1, s_2 \to t)$ is by knowing the joint distribution $P(S_1, S_2, T)$ with respect to the joint two-event partitions $\mathcal{S}_1^{s_1} \times \mathcal{S}_2^{s_2} \times \mathcal{T}^t$. This is (in effect) a pointwise form of their Assumption (∗∗). Thirdly, by definition $r_{\min}^\pm$ are given by the minimum value that any one source event provides. This is the largest possible value that one could take for these quantities whilst still requiring that the unique specificity and ambiguity be non-negative. Hence, within each discrete realisation, $r_{\min}^\pm$ minimise the unique specificity and ambiguity whilst maximising the redundant specificity and ambiguity. This is similar to $\widetilde{UI}$ which minimises the (average) unique information while still satisfying Assumption (∗). Finally, note that since the measure $\mathcal{S}_{\text{VK}}$ produces a bivariate decomposition which is equivalent to that of $\widetilde{UI}$ [11], the same similarities apply between PPID using specificity and ambiguity and the decomposition induced by $\mathcal{S}_{\text{VK}}$ from Griffith and Koch [12].

5.2. Probability Distribution XOR

Figure 4 shows the canonical example of synergy, *exclusive-or* (XOR) which considers two independently distributed binary predictor variables S_1 and S_2 and a target variable $T = S_1$ XOR S_2. There are several important points to note about the decomposition of XOR. Firstly, despite providing zero pointwise information, an individual predictor event does indeed induce exclusions. However, the informative and misinformative exclusions are perfectly balanced such that the posterior (conditional) distribution is equal to the prior distribution, e.g., see the red coloured exclusions induced by $S_1 = 0$ in Figure 4. In information-theoretic terms, for each realisation, the pointwise specificity equals 1 bit since half of the total probability mass remains while the pointwise ambiguity also equals 1 bit since half of the probability mass associated with the event which subsequently occurs (i.e., $T = 0$), remains. These are perfectly balanced such that when recombined, as per (11), the pointwise mutual information is equal to 0 bit, as one would expect.

Secondly, $S_1 = 0$ and $S_2 = 0$ both induce the same exclusions with respect to the target pointwise event space $\mathcal{T}^{T=0}$. Hence, as per the operational interpretation of redundancy adopted in Section 3.3, there is 1 bit of pointwise redundant specificity and 1 bit of pointwise redundant ambiguity in each realisation. The presence of (a form of) redundancy in XOR is novel amongst the existing measures in the PID literature. (Ince [19] also identifies a form of redundancy in XOR.) Thirdly, despite the presence of this redundancy, recombining the atoms of pointwise specificity and ambiguity for each realisation, as per (21), leaves only one non-zero PPI atom: namely the pointwise complementary information $c(s_1, s_2 \to t) = 1$ bit. Furthermore, this is true for every pointwise realisation and hence, by (22), the only non-zero PI atom is the average complementary information $C(S_1, S_2 \to T) = 1$ bit.
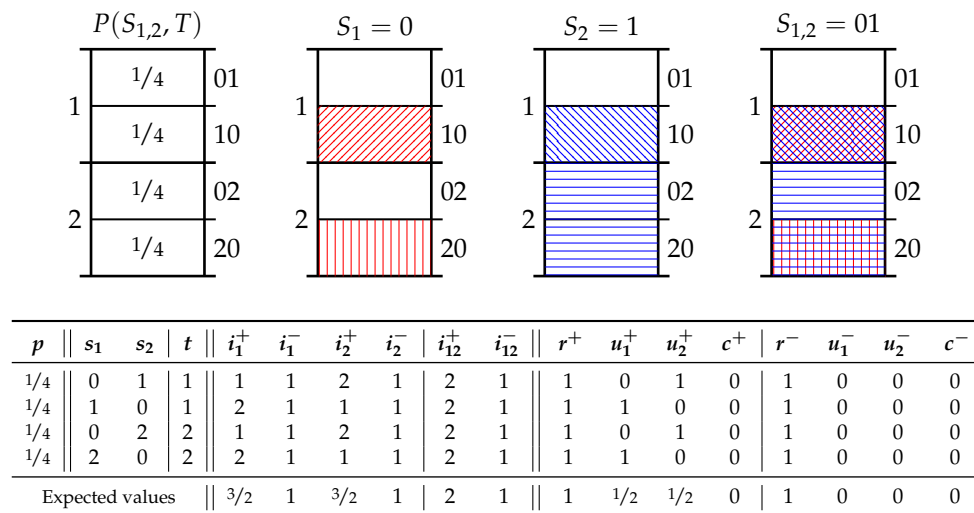
p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	0	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	0	1	1	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	1	0	1	1	1	1	1	2	1	1	0	0	1	1	0	0	0
1/4	1	1	0	1	1	1	1	2	1	1	0	0	1	1	0	0	0
Expected values				1	1	1	1	2	1	1	0	0	1	1	0	0	0

$$R(S_1, S_2 \to T) = 0 \text{ bit} \qquad U(S_1 \setminus S_2 \to T) = 0 \text{ bit} \qquad U(S_2 \setminus S_1 \to T) = 0 \text{ bit} \qquad C(S_1, S_2 \to T) = 1 \text{ bit}$$

Figure 4. Example XOR. (**Top**) probability mass diagrams for the realisation $(S_1 = 0, S_2 = 0, T = 0)$; (**Middle**) For each realisation, the pointwise specificity and pointwise ambiguity has been evaluated using (5) and (8) respectively. The pointwise redundant specificity and pointwise redundant ambiguity are then determined using (23) and (24). The decomposition is calculated using (18) and (19). The expected specificity and ambiguity are calculated with (20); (**Bottom**) The average information is given by (22). As expected, XOR yields 1 bit of complementary information.

5.3. Probability Distribution PwUNQ

Figure 5 shows the probability distribution PwUNQ introduced in Section 2.2. Recombining the decomposition via (21) yields the pointwise information decomposition proposed in Table 1—unsurprisingly, the explicitly pointwise approach results in a decomposition which does not suffer from the pointwise unique problem of Section 2.2.

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	1	1	1	1	2	1	2	1	1	0	1	0	1	0	0	0
1/4	1	0	1	2	1	1	1	2	1	1	1	0	0	1	0	0	0
1/4	0	2	2	1	1	2	1	2	1	1	0	1	0	1	0	0	0
1/4	2	0	2	2	1	1	1	2	1	1	1	0	0	1	0	0	0
Expected values				3/2	1	3/2	1	2	1	1	1/2	1/2	0	1	0	0	0

$$R(S_1, S_2 \to T) = 0 \text{ bit} \qquad U(S_1 \setminus S_2 \to T) = 1/2 \text{ bit} \qquad U(S_2 \setminus S_1 \to T) = 1/2 \text{ bit} \qquad C(S_1, S_2 \to T) = 0 \text{ bit}$$

Figure 5. Example PwUNQ. (**Top**) probability mass diagrams for the realisation $(S_1 = 0, S_2 = 1, T = 1)$; (**Middle**) For each realisation, the pointwise partial information decomposition (PPID) using specificity and ambiguity is evaluated (see Figure 4 for details). Upon recombination as per (21), the PPI decomposition from Table 1 is attained; (**Bottom**) as does the average information—the decomposition does not have the pointwise unique problem.

In each realisation, observing a 0 in either source provides the same balanced informative and misinformative exclusions as in XOR. Observing either a 1 or 2 provides the same misinformative exclusion as observing the 0, but provides a larger informative exclusion than 0. This leaves only the probability mass associated with the event which subsequently occurs remaining (hence why observing a 1 and 2 is fully informative about the target). Information theoretically, in each realisation the predictor events provide 1 bit of redundant pointwise specificity and 1 bit of redundant pointwise ambiguity while the fully informative event additionally provides 1 bit of unique specificity.

5.4. Probability Distribution RDNERR

Figure 6 shows the probability distribution *redundant-error* (RDNERR) which considers two predictors which are nominally redundant and fully informative about the target, but where one predictor occasionally makes an erroneous prediction. Specifically, Figure 6 shows the decomposition of RDNERR where S_2 makes an error with a probability $\varepsilon = 1/4$. The important feature to note about this probability distribution is that upon recombining the specificity and ambiguity and taking the expectation over every realisation, the resultant average unique information from S_2 is $U(S_2 \setminus S_1 \to T) = -0.811$ bit.

On first inspection, the result that the average unique information can be negative may seem problematic; however, it is readily explainable in terms of the operational interpretation of Section 3.3. In RDNERR, a source event always excludes exactly $1/2$ of the total probability mass, thus every realisation contains 1 bit of redundant pointwise specificity. The events of the error-free S_1 induce only informative exclusions and as such provide 0 bit of pointwise ambiguity in each realisation. In contrast, the events in the error-prone S_2 always induce a misinformative exclusion, meaning that S_2 provides unique pointwise ambiguity in every realisation. Since S_2 never provides unique specificity, the average unique information is negative on average.

Despite the negativity of the average unique information, in is important to observe that S_2 provides 0.189 bit of information since S_2 also provides 1 bit of average redundant information. It is not that S_2 provides negative information on average (as this is not possible); rather it is that not all of the information provided by S_2 (i.e., the specificity) is "useful" ([42], p. 21). This is in contrast to S_1 which only provides useful specificity. To summarise, it is the unique ambiguity which distinguishes the information provided by variable S_2 from S_1, and hence why S_2 is deemed to provide negative average unique information. This form of uniqueness can only be distinguished by allowing the average unique information to be negative. This of course, requires abandoning the local positivity as a required property, as per Theorem 4. Few of the existing measures in the PID literature consider dropping this requirement as negative information quantities are typically regarded as being "unfortunate" ([43], p. 49). However, in the context of the pointwise mutual information, negative information values are readily interpretable as being misinformative values. Despite this, the average information from each predictor must be non-negative; however, it may be that what distinguishes one predictor from another are precisely the misinformative predictor events, meaning that the unique information is in actual fact, unique misinformation. Forgoing local positivity makes the PPID using specificity and ambiguity novel (the other exception in this regard is Ince [18] who was first to consider allowing negative average unique information.)
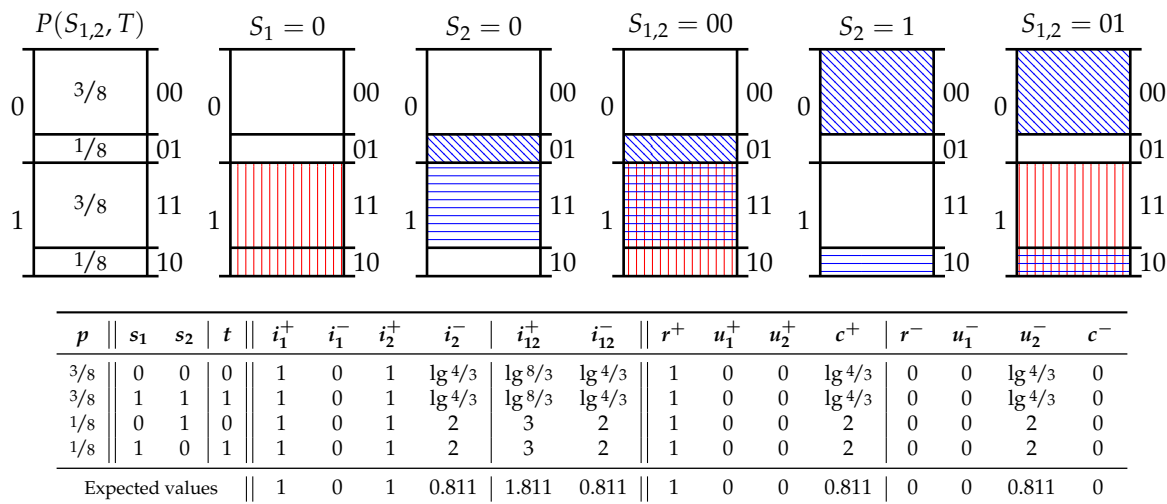
p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
3/8	0	0	0	1	0	1	lg 4/3	lg 8/3	lg 4/3	1	0	0	lg 4/3	0	0	lg 4/3	0
3/8	1	1	1	1	0	1	lg 4/3	lg 8/3	lg 4/3	1	0	0	lg 4/3	0	0	lg 4/3	0
1/8	0	1	0	1	0	1	2	3	2	1	0	0	2	0	0	2	0
1/8	1	0	1	1	0	1	2	3	2	1	0	0	2	0	0	2	0
Expected values				1	0	1	0.811	1.811	0.811	1	0	0	0.811	0	0	0.811	0

$$R(S_1, S_2 \to T) = 1 \text{ bit} \quad U(S_1 \backslash S_2 \to T) = 0 \text{ bit} \quad U(S_2 \backslash S_1 \to T) = -0.811 \text{ bit} \quad C(S_1, S_2 \to T) = 0.811 \text{ bit}$$

Figure 6. Example RDNERR. (**Top**) probability mass diagrams for the realisations $(S_1 = 0, S_2 = 0, T = 0)$ and $(S_1 = 0, S_2 = 1, T = 0)$; (**Middle**) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4 for details); (**Bottom**) the average PI atoms may be negative as the decomposition does not satisfy local positivity.

5.5. Probability Distribution TBC

Figure 7 shows the probability distribution *two-bit-copy* (TBC) which considers two independently distributed binary predictor variables S_1 and S_2, and a target variable T consisting of a separate elementary event for each joint event $S_{1,2}$. There are several important points to note about the decomposition of TBC. Firstly, due to the symmetry in the probability distribution, each realisation will have the same pointwise decomposition. Secondly, due to the construction of the target, there is an isomorphism (Again, isomorphism should be taken to mean isomorphic probability spaces, e.g., [37], p. 27 or [38], p. 4) between $P(T)$ and $P(S_1, S_2)$, and hence the pointwise ambiguity provided by any (individual or joint) predictor event is 0 bit (since given t, one knows s_1 and s_2). Thirdly, the individual predictor events s_1 and s_2 each exclude 1/2 of the total probability mass in $P(T)$ and so each provide 1 bit of pointwise specificity; thus, by (23), there is 1 bit of redundant pointwise specificity in each realisation. Fourthly, the joint predictor event $s_{1,2}$ excludes 3/4 of the total probability mass, providing 2 bit of pointwise specificity; hence, by (18), each joint realisation provides 1 bit of pointwise complementary specificity in addition to the 1 bit of redundant pointwise specificity. Finally, putting this together via (22), TBC consists of 1 bit of average redundant information and 1 bit of average complementary information.

Although "surprising" ([5], p. 268), according to the operational interpretation adopted in Section 3.3, two independently distributed predictor variables can share redundant information. That is, since the exclusions induced by s_1 and s_2 are the same with respect to the two-event partition $\mathcal{T}^t$, the information associated with these exclusions is regarded as being the same. Indeed, this probability distribution highlights the significance of specific reference to the two-event partition in Section 3.3 and Axiom 4. (This can be seen in the probability mass diagram in Figure 7, where the events $S_1 = 0$ and $S_2 = 0$ exclude different elementary target events within the complementary event 0^c and yet are considered to be the same exclusion with respect to the two-event partition $\mathcal{T}^0$.) That these exclusions should be regarded as being the same is discussed further in Appendix A. Now however, there is a need to discuss TBC in terms of Theorem 5 (Target Chain Rule).
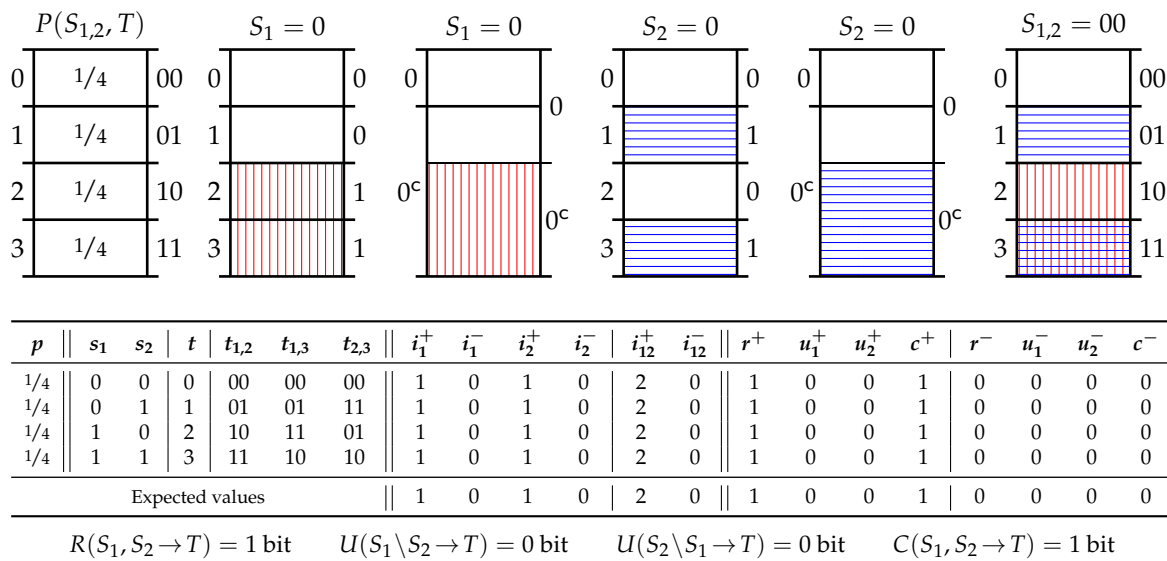
p	s_1	s_2	t	$t_{1,2}$	$t_{1,3}$	$t_{2,3}$	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	0	00	00	00	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	0	1	1	01	01	11	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	1	0	2	10	11	01	1	0	1	0	2	0	1	0	0	1	0	0	0	0
1/4	1	1	3	11	10	10	1	0	1	0	2	0	1	0	0	1	0	0	0	0
	Expected values						1	0	1	0	2	0	1	0	0	1	0	0	0	0

$$R(S_1, S_2 \to T) = 1 \text{ bit} \qquad U(S_1 \backslash S_2 \to T) = 0 \text{ bit} \qquad U(S_2 \backslash S_1 \to T) = 0 \text{ bit} \qquad C(S_1, S_2 \to T) = 1 \text{ bit}$$

Figure 7. Example TBC. (**Top**) the probability mass diagrams for the realisation $(S_1 = 0, S_2 = 0, T = 00)$; (**Middle**) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4); (**Bottom**) the decomposition of XOR yields the same result as $I_{\min}$.

TBC was first considered as a "mechanism" ([6], p. 3) where "the wires don't even touch" ([12], p. 167), which merely copies or concatenates S_1 and S_2 into a composite target variable $T_{1,2} = (T_1, T_2)$ where $T_1 = S_1$ and $T_2 = S_2$. However, using causal mechanisms as a guiding intuition is dubious since different mechanisms can yield isomorphic probability distributions ([44], and references therein). In particular, consider two mechanisms which generate the composite target variables $T_{1,3} = (T_1, T_3)$ and $T_{2,3} = (T_2, T_3)$ where $T_3 = S_1$ XOR S_2. As can be seen in Figure 7, both of these mechanisms generate the same (isomorphic) probability distribution $P(S_1, S_2, T)$ as the mechanism generating $T_{1,2}$. If an information decomposition is to depend only on the probability distribution $P(S_1, S_2, T)$, and no other semantic details such as labelling, then all three mechanisms must yield the same information decomposition—this is not clear from the mechanistic intuition.

Although the decomposition of the various composite target variables must be the same, there is no requirement that the three systems must yield the same decomposition when analysed in terms of the individual components of the composite target variables. Nonetheless, there ought to be a consistency between the decomposition of the composite target variables and the decomposition of the component target variables—i.e., there should be a target chain rule. As shown in Theorem 5, the measures $r_{\min}$ and $R_{\min}$ satisfy the target chain rule, whereas $I_{\min}$, $\widetilde{UI}$, I_{red} and $\mathcal{S}_{\text{VK}}$ do not [5,7]. Failing to satisfy the target chain rule can lead to inconsistencies between the composite and component decompositions, depending on the order in which one considers decomposing the information (this is discussed further in Appendix A.3). In particular, Table 2 shows how $\widetilde{UI}$, I_{red} and $\mathcal{S}_{\text{VK}}$ all provide the same inconsistent decomposition for TBC when considered in terms of the composite target variable $T_{1,3}$. In contrast, $R_{\min}$ produces a consistent decomposition of $T_{1,3}$. Finally, based on the above isomorphism, consider the following (the proof is deferred to Appendix B.3).

Theorem 6. *The target chain rule, identity property and local positivity, cannot be simultaneously satisfied.*

Table 2. Shows the decomposition of the quantities in the first row induced by the measures in the first column. For consistency, the decomposition of $I(S_{1,2}; T_{1,3})$ should equal both the sum of the decomposition of $I(S_{1,2}; T_1)$ and $I(S_{1,2}; T_3|T_1)$, and the sum of the decomposition of $I(S_{1,2}; T_3)$ and $I(S_{1,2}; T_1|3)$. Note that the decomposition induced by $\widetilde{UI}$, I_{red} and $\mathcal{S}_{\text{VK}}$ are not consistent. In contrast, $R_{\min}$ is consistent due to Theorem 5.

	$I(S_{1,2}; T_{1,3})$		$I(S_{1,2}; T_1)$	$I(S_{1,2}; T_3|T_1)$		$I(S_{1,2}; T_3)$	$I(S_{1,2}; T_1|T_3)$
$\widetilde{UI}$, I_{red}, $\mathcal{S}_{\text{VK}}$	$U(S_1 \backslash S_2 \rightarrow T_{1,3}) = 1$ $U(S_2 \backslash S_1 \rightarrow T_{1,3}) = 1$		$U(S_1 \backslash S_2 \rightarrow T_1) = 1$	$U(S_2 \backslash S_1 \rightarrow T_3|T_1) = 1$		$C(S_1, S_2 \rightarrow T_3) = 1$	$R(S_1, S_2 \rightarrow T_1|T_3) = 1$
$R_{\min}$	$R(S_1, S_2 \rightarrow T_{1,3}) = 1$ $C(S_1, S_2 \rightarrow T_{1,3}) = 1$		$U(S_2 \backslash S_1 \rightarrow T_1) = -1$ $R(S_1, S_2 \rightarrow T_1) = 1$ $C(S_1, S_2 \rightarrow T_1) = 1$	$U(S_2 \backslash S_1 \rightarrow T_3|T_1) = 1$		$C(S_1, S_2 \rightarrow T_3) = 1$	$R(S_1, S_2 \rightarrow T_1|T_3) = 1$

5.6. Summary of Key Properties

The following are the key properties of the PPID using the specificity and ambiguity. Property 1 follows directly from the Definitions 1 and 2. Property 2 follows from Theorems 3 and 4. Property 3 follows from the probability distribution TBC in Section 5.5. Property 4 was discussed in Section 4.2. Property 5 is proved in Theorem 5.

Property 1. *When considering the redundancy between the source events* $a_1, \ldots, a_k$, *at least one source event* a_i *will provide zero unique specificity, and at least one source event* a_j *will provide zero unique ambiguity. The events* a_i *and* a_j *are not necessarily the same source event.*

Property 2. *The atoms of partial specificity and partial ambiguity satisfy local positivity,* $\pi^{\pm} \geq 0$. *However, upon recombination and averaging, the atoms of partial information do not satisfy local positivity,* $\Pi \geq 0$.

Property 3. *The decomposition does not satisfy the identity property.*

Property 4. *The decomposition does not satisfy the target monotonicity property.*

Property 5. *The decomposition satisfies the target chain rule.*

6. Conclusions

The partial information decomposition of Williams and Beer [12] provided an intriguing framework for the decomposition of multivariate information. However, it was not long before "serious flaws" ([11], p. 2163) were identified. Firstly, the measure of redundant information $I_{\min}$ failed to distinguish between whether predictor variables provide the same information or merely the same amount of information. Secondly, $I_{\min}$ fails to satisfy the target chain rule, despite this addativity being one of the defining characteristics of information. Notwithstanding the problems, the axiomatic derivation of the redundancy lattice was too elegant to be abandoned and hence several alternate measures were proposed, i.e., I_{red}, $\widetilde{UI}$ and $\mathcal{S}_{\text{VK}}$ [6,11,12]. Nevertheless, as these measures all satisfy the identity property, they cannot produce a non-negative decomposition for an arbitrary number of variables [13]. Furthermore, none of these measures satisfy the target chain rule meaning they produce inconsistent decompositions for multiple target variables. Finally, in spite of satisfying the identity property (which many consider to be desirable), these measures still fail to identify when variables provide the same information, as exemplified by the pointwise unique problem presented in Section 2.

This paper took the axiomatic derivation of the redundancy lattice from PID and applied it to the unsigned entropic components of the pointwise mutual information. This yielded two separate redundancy lattices—the specificity and the ambiguity lattices. Then based upon an operational interpretation of redundancy, measures of pointwise redundant specificity $r_{\min}^{+}$ and pointwise redundant

ambiguity $r_{\min}^-$ were defined. Together with specificity and ambiguity lattices, these measures were used to decompose multivariate information for an arbitrary number of variables. Crucially, upon recombination, the measure $r_{\min}$ satisfies the target chain rule. Furthermore, when applied to PwUnq, these measures do not result in the pointwise unique problem. In our opinion, this demonstrates that the decomposition is indeed correctly identifying redundant information. However, others will likely disagree with this point given that the measure of redundancy does not satisfy the identity property. According to the identity property, independent variables can never provide the same information. In contrast, according to the operational interpretation adopted in this paper, independent variables can provide the same information if they happen to provide the same exclusions with respect to the two-event target distribution. In any case, the proof of Theorem 6 and the subsequent discussion in Appendix B.3, highlights the difficulties that the identity property introduces when considering the information provided about events in separate target variables. (See further discussion in Appendix A.3).

Our future work with this decomposition will be both theoretical and empirical. Regarding future theoretical work, given that the aim of information decomposition is to derive measures pertaining to sets of random variables, it would be worthwhile to derive the information decomposition from first principles in terms of measure theory. Indeed, such an approach would surely eliminate the semantic arguments (about what it means for information to unique, redundant or complementary), which currently plague the problem domain. Furthermore, this would certainly be a worthwhile exercise before attempting to generalise the information decomposition to continuous random variables. Regarding future empirical work, there are many rich data sets which could be decomposed using this decomposition including financial time-series and neural recordings, e.g., [28,33,34].

Author Contributions: C.F. and J.L. conceived the idea; C.F. designed, wrote and analyzed the computational examples; C.F. and J.L. wrote the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A. Kelly Gambling, Axiom 4, and TBC

In Section 3.3, it was argued that the information provided by a set of predictor events $s_1, \ldots, s_k$ about a target event t is the same information if each source event induces the same exclusions with respect to the two-event partition $\mathcal{T}^t = \{t, t^c\}$. This was based on the fact that pointwise mutual information does not depend on the apportionment of the exclusions across the set of events which did not occur t^c. It was argued that since the pointwise mutual information is independent of these differences, the redundant mutual information should also be independent of these differences. This requirement was then integrated into the operational interpretation of Section 3.3 and was later enshrined in the form of Axiom 4. This appendix aims to justify this operational interpretation and argue why redundant information in TBC is not "unreasonably large" ([5], p. 269).

Appendix A.1. Pointwise Side Information and the Kelly Criterion

Consider a set of horses $\mathcal{T}$ running in a race which can be considered a random variable T with distribution $P(T)$. Say that for each $t \in \mathcal{T}$ a bookmaker offers odds of $o(t)$-for-1, i.e., the bookmaker will pay out $o(t)$ dollars on a \$1 bet if the horse t wins. Furthermore, say that there is no track take as $\sum_{t \in \mathcal{T}} 1/o(t) = 1$, and these odds are *fair*, i.e., $o(t) = 1/p(t)$ for all $t \in \mathcal{T}$ [40]. Let $b(T)$ be the fraction of a gambler's capital bet on each horse $t \in \mathcal{T}$ and assume that the gambler stakes all of their capital on the race, i.e., $\sum_{t \in \mathcal{T}} b(t) = 1$.

Now consider an i.i.d. series of these races $T_1, T_2, \ldots$ such that $P(T_k) = P(T)$ for all $k \in \mathbb{N}$ and let $t_k \in \mathcal{T}$ represent the winner of the k-th race. Say that the bookmaker offers the same odds on each race and the gambler bets their entire capital on each race. The gambler's capital after m races D_m is a random variable which depends on two factors per race: the amount the gambler staked on each race winner t_k, and the odds offered on each winner t_k. That is,

$$D_m = \prod_{k=1}^{m} b(t_k)\, o(t_k), \tag{A1}$$

where monetary units $ have been chosen such that $D_0 = \$1$. The gambler's wealth grows (or shrinks) exponentially, i.e.,

$$D_m = 2^{m\, W(\boldsymbol{b}, T)} \tag{A2}$$

where

$$W(\boldsymbol{b}, T) = \frac{1}{m} \log D_m = \frac{1}{m} \sum_{k=1}^{m} \log b(t_k)\, o(t_k) = \mathrm{E}\big[\log b(t_k)\, o(t_k)\big] \tag{A3}$$

is the doubling rate of the gambler's wealth using a betting strategy $\boldsymbol{b}(T)$. Here, the last equality is by the weak law of large numbers for large m.

Any reasonable gambler would aim to use an optimal strategy $\boldsymbol{b}^*(T)$ which maximises the doubling rate $W(\boldsymbol{b}, T)$. Kelly [40,43] proved that the optimal doubling rate is given by

$$W^*(T) = \max_{\boldsymbol{b}} W(\boldsymbol{b}, T) = \mathrm{E}\big[\log b^*(t_k)\, o(t_k)\big] \tag{A4}$$

and is achieved by using the proportional gambling scheme $\boldsymbol{b}^*(T) = P(T)$. When the race T_k occurs and the horse t_k wins, the gambler will receive a payout of $b^*(t^k)\, o(t^k) = \$1$, i.e., the gambler receives their stake back regardless of the outcome. In the face of fair odds, the proportional Kelly betting scheme is the optimal strategy—non-terminating repeated betting with any other strategy will result in losses.

Now consider a gambler with access to a private wire S which provides (potentially useful) side information about the upcoming race. Say that these messages are selected from the set $\mathcal{S}$, and that the gambler receives the message s_k before the race T_k. Kelly [40,43] showed that the optimal doubling rate in the presence of this side information is given by

$$W^*(T|S) = \max_{\boldsymbol{b}} W(\boldsymbol{b}, T|S) = \mathrm{E}\big[\log b^*(t_k|s_k)\, o(t_k)\big], \tag{A5}$$

and is achieved by using the conditional proportional gambling scheme $\boldsymbol{b}^*(T|s_k) = P(T|s_k)$. Both the proportional gambling scheme $b^*(T)$ and the conditional proportional gambling scheme $b^*(T|S)$ are based upon the *Kelly criterion* whereby bets are apportioned according to the best estimation of the outcome available. The financial value of the private wire to a gambler can be ascertained by comparing their doubling rate of the gambler with access to the side wire to that of a gambler with no side information, i.e.,

$$\begin{aligned} \Delta W = W^*(T|S) - W^*(T) &= \mathrm{E}\big[\log b^*(t_k|s_k)\, o(t_k)\big] - \mathrm{E}\big[\log b^*(t_k)\, o(t_k)\big] \\ &= \mathrm{E}\big[i(s_k; t_k)\big] = I(S; T). \end{aligned} \tag{A6}$$

This important result due to Kelly [40] equates the increase in the doubling rate ΔW due to the presence of side information, with the mutual information between the private wire S and the horse race T. If on average, the gambler receives 1 bit of information from their private wire, then on average the gambler can expect to double their money per race. Furthermore, as one would expect, independent side information does not increase the doubling rate.

With no side information, the Kelly gambler always received their original stake back from the bookmaker. However, this is not true for the Kelly gambler with side information. Although their doubling rate is greater than or equal to that of the gambler with no side information, this is only true *on average*. Before the race T_k, the gambler receives the private wire message s_k and then, the horse t_k wins the race. From (A6), one can see that the return Δw_k for the k-th race is given by the pointwise mutual information,

$$\Delta w = i(s_k; t_k). \tag{A7}$$

Hence, just like the pointwise mutual information, the per race return can be positive or negative: if it is positive, the gambler will make a profit; if it is negative, the gambler will sustain a loss. Despite the potential for pointwise loses, the average return (i.e., the doubling rate) is, just like the average mutual information, non-negative—and indeed, is optimal. Furthermore, while a Kelly gambler with side information can lose money on any single race, they can never actually go bust. The Kelly gambler with side information s still hedges their risk by placing bets on all horses with a non-zero probability of winning according to their side information, i.e., according to $P(T|s_k)$. The only reason they would fail to place a bet on a horse is if their side information completely precludes any possibility of that horse winning. That is, a Kelly gambler with side information will never fall foul of gambler's ruin.

Appendix A.2. Justification of Axiom 4 and Redundant Information in TBC

Consider TBC semantically described in terms of a horse race. That is, consider a four horse race T where each horse has an equiprobable chance of winning, and consider the binary variables T_1, T_2, and T_3 which represent the following, respectively: the colour of the horse, black 0 or white 1; the sex of the jockey, female 0 or male 1; and the colour of the jockey's jersey, red 0 or green 1. Say that the four horses have the following attributes:

Horse 0　is a black horse $T_1 = 0$, ridden by a female jockey $T_2 = 0$, who is wearing a red jersey $T_3 = 0$.
Horse 1　is a black horse $T_1 = 0$, ridden by a male jockey $T_2 = 1$, who is wearing a green jersey $T_3 = 1$.
Horse 2　is a white horse $T_1 = 1$, ridden by a female jockey $T_2 = 0$, who is wearing a green jersey $T_3 = 1$.
Horse 3　is a white horse $T_1 = 1$, ridden by a male jockey $T_2 = 1$, who is wearing a red jersey $T_3 = 0$.

There are two important points to note. Firstly, the horses in the race T could also be uniquely described in terms of the composite binary variables $T_{1,2}$, $T_{1,3}$ or $T_{2,3}$. Secondly, if one knows T_1 and T_2 then one knows T_3 (which can be represented by the relationship $T_3 = T_1$ XOR T_2). Finally, consider private wires S_1 and S_2 which *independently* provide the colour of the horse and the colour of the jockey's jersey (respectively) before the upcoming race, i.e., $S_1 = T_1$ and $S_2 = T_2$.

Now say a bookmaker offers fair odds of 4-for-1 on each horse in the race T. Consider two gamblers who each have access to one of S_1 and S_2. Before each race, the two gamblers receive their respective private wire messages and place their bets according to the Kelly strategy. This means that each gambler lays half of their, say \$1, stake on each of their two respective non-excluded horses: unknowingly, both of the gamblers have placed a bet on the soon-to-be race winner, and each gambler has placed a distinct bet on one of the two soon-to-be losers. The only horse neither has bet upon is also a soon-to-be loser. (See [5] for a related description of TBC in term of the game-theoretic notions of shared and common knowledge). After the race, the bookmaker pays out \$2 dollars to each gambler: both have doubled their money. This happens because both of the gamblers had one bit of 1 bit of information about the race, i.e., pointwise mutual information. In particular, both gamblers improved their probability of predicting the eventual race winner. It did not matter, in any way, that the gamblers had each laid distinct bets on one of the three eventual race losers. The fact that they laid different bets on the horses which did not win, made no difference to their winnings. The apportionment of the exclusions across the set of events which did not occur, makes no difference to the pointwise mutual information. With respect to what occurred (i.e., with respect to which horse won), the fact the that they excluded different losers is only semantic. When it came to predicting the would-be-winner,

both gamblers had the same predictive power; they both had the same freedom of choice with regards to selecting what would turn out to be the eventual race winner—they had the same information. It is for this reason that this information should be regarded as redundant information, regardless of the independence of the information sources. Hence, the introduction of both the operational interpretation of redundancy in Section 3.3 and Axiom 4 in Section 4.2.

Now consider a third gambler who has access to both private wires S_1 and S_2, i.e., $S_{1,2}$. Before the race, this gambler receives both private wire messages which, in total, precludes three of the horses from winning. This gambler then places the entirety of their $1 stake on the remaining horse which is sure to win. After the race, the bookmaker pays out $4: this gambler has quadrupled their money as they had 2 bit of pointwise mutual information about the race. Having both private wire messages simultaneously gave this gambler a 1 bit informational edge over the two gamblers with access to a single side wire. While each of the singleton gamblers had 1 bit of independent information, the only way one could profit from the independence of this information is by having both pieces of information simultaneously—this makes this 1 bit of information complementary. Although this may seem "palpably strange" ([12], p. 167), it is not so strange when from the following perspective: the only way to exploit two pieces of independent information is by having both pieces together simultaneously.

Appendix A.3. Accumulator Betting and the Target Chain Rule

Say that in addition to the 4-for-1 odds offered on the race T, the bookmaker also offers fair odds of 2-for-1 on each of the binary variables T_1, T_2 and T_3. Now, in addition to being able to directly gamble on the race T, one could indirectly gamble on T by placing a so-called *accumulator* bet on any pair of T_1, T_2 and T_3. An accumulator is a series of chained bets whereby any return from one bet is automatically staked on the next bet; if any bet in the chain is lost then the entire chain is lost. For example, a gambler could place 4-for-1 bet on horse 0 by placing the following accumulator bet: a 2-for1 bet on a black horse winning that chains into a 2-for-1 bet on the winning jockey being female (or equivalently, vice versa). In effect, these accumulators enable a gambler to bet on T by instead placing a chained bet on the independent component variables within the (equivalent) joint variables $T_{1,2}$, $T_{1,3}$ and $T_{2,3}$. Now consider again the three gamblers from the prior section, i.e., the two gamblers who each have a private wire S_1 and S_2, and the third gamble who has access to $S_{1,2}$. Say that they must each place a, say $1, accumulator bet on $T_{1,3}$—what should each gambler do according to the Kelly criterion?

For the sake of clarity, consider only the realisation where the horse $T = 0$ subsequently wins (due to the symmetry, the analysis is equivalent for all realisations). First consider the accumulator whereby the gamblers first bet on the colour of the winning horse T_1, which chains into a bet on the colour of the winning jockey's jersey T_3. Suppose that the private wire S_1 communicates that the winning horse will be black, while the private wire S_2 communicates that the winning horse will be ridden by a female jockey, i.e., $S_1 = 0$ and $S_2 = 0$. Following to the Kelly strategy, the gambler with access to $S_1 = 0$ takes out two $0.5 accumulator bets. Both of these accumulators feature the same initial bet on the winning horse being black since $T_1 = S_1 = 0$. Hence both bets return $1 each which become the stake on the next bet in each accumulator. This gambler knows nothing about the colour of the jockey's jersey T_3. As such, one accumulator chains into a bet on the winning jersey being red $T_3 = 0$, while the other chains into a bet on it being green $T_3 = 1$. When the horse $T = 0$ wins, the stake bet on the green jersey is lost while bet on red jersey pays out $2. This gambler had 1 bit of side information and so doubled their money. Now consider the gambler with private wire S_2, who knows nothing about T_1 or T_3 individually. Nonetheless, this gambler knows that the winner must be a female jockey $T_2 = 0$. As such, this gambler knows that if a black horse $T_1 = 0$ wins then its jockey must be wearing a red jersey $T_3 = 0$, or if a white horse $T_1 = 0$ wins then its jockey must be wearing a green jersey $T_3 = 1$ (since $T_3 = T_1$ XOR T_2). Thus this gambler can also utilise the Kelly strategy to place the following two $0.5 accumulator bets: the first accumulator bets on the winning horse being black $T_1 = 0$ and then chains into a bet on the winner's jersey being red $T_3 = 0$, while the second accumulator bets on the winning horse being white $T_1 = 1$ and then chains into a bet on the winner's jersey being green

$T_3 = 1$. When the horse $T = 0$ wins, the first accumulator pays out \$2, while the second accumulator is be lost. Hence, this gambler also doubles their money and so also had 1 bit of side information. Finally, consider the gambler with access to both private wires $S_{1,3}$, who can place an accumulator on the black horse $T_1 = 0$ winning chaining into a bet on the winning jockey wearing red $T_3 = 0$. This gambler can quadruple their stake, and so must possess 2 bit of side information.

Each of the three gamblers have the same final return regardless of whether the gamblers are betting on the variable T, or placing accumulator bets on the variables $T_{1,2}$, $T_{1,3}$ or $T_{2,3}$. However, the paths to the final result differs between the gamblers, reflecting the difference between the information the each gambler had about the sub-variables T_1, T_2 or T_3. Given the result of Kelly [40], the proposed information decomposition should reflect these differences, but yet still arrive at the same result—in other words, the information decomposition should satisfy a target chain rule. This is clear if the Kelly interpretation of information is to remain as a "duality" ([43], p. 159) in information theory.

Appendix B. Supporting Proofs and Further Details

This appendix contains many of the important theorems and proofs relating to PPID using specificity and ambiguity.

Appendix B.1. Deriving the Specificity and Ambiguity Lattices from Axioms 1–4

The following section is based directly on the original work of Williams and Beer [1,2]. The difference is that we now consider sources events a_i rather than sources A_i.

Proposition A1. *Both $i_\cap^+$ and $i_\cap^-$ are non-negative.*

Proof. Since $\varnothing \subseteq a_i$ for any a_i, Axioms 2 and 3 imply

$$i_\cap^+ (a_1,\ldots,a_k \to t) \geq i_\cap^+ (a_1,\ldots,a_k,\varnothing \to t) = i_\cap^+ (\varnothing \to t) = h(\varnothing) \quad = 0, \tag{A8}$$
$$i_\cap^- (a_1,\ldots,a_k \to t) \geq i_\cap^- (a_1,\ldots,a_k,\varnothing \to t) = i_\cap^- (\varnothing \to t) = h(\varnothing|t) = 0. \tag{A9}$$

Hence, both $i_\cap^+ (a_1,\ldots,a_k \to t)$ and $i_\cap^- (a_1,\ldots,a_k \to t)$ are non-negative. □

Proposition A2. *Both $i_\cap^+$ and $i_\cap^-$ are bounded from above by the specificity and the ambiguity from any single source event, respectively.*

Proof. For any single source a_i, Axioms 2 and 3 yield

$$h(a_i) = i_\cap^+ (a_i \to t) = i_\cap^+ (a_i, a_i \to t) \geq i_\cap^+ (a_i,\ldots \to t), \tag{A10}$$
$$h(a_i|t) = i_\cap^- (a_i \to t) = i_\cap^- (a_i, a_i \to t) \geq i_\cap^+ (a_i,\ldots \to t), \tag{A11}$$

as required. □

In keeping with Williams and Beer's approach [1,2], consider all of the distinct ways in which a collection of source events $a = \{a_1,\ldots,a_k\}$ could contribute redundant information. Thus far we have assumed that the redundancy measure can be applied to any collection of source events, i.e., $\mathscr{P}_1(a)$ where $\mathscr{P}_1$ denotes the power set with the empty set removed. Recall that the sources events are themselves collections of predictor events, i.e., $\mathscr{P}_1(s)$. That is, we can apply both $i_\cap^+$ and $i_\cap^-$ to elements of $\mathscr{P}_1(\mathscr{P}_1(s))$. However, this can be greatly reduced using Axiom 2 which states that if $a_i \subseteq a_j$, then

$$i_\cap^+ (a_j, a_i,\ldots \to t) = i_\cap^+ (a_i,\ldots \to t), \tag{A12}$$
$$i_\cap^- (a_j, a_i,\ldots \to t) = i_\cap^- (a_i,\ldots \to t). \tag{A13}$$

Hence, one need only consider the collection of source events such that no source event is a superset of any other in order,

$$\mathscr{A}(s) = \left\{ \alpha \in \mathscr{P}_1\big(\mathscr{P}_1(s)\big) \mid \forall\, a_i,\, a_j \in \alpha,\, a_i \not\subset a_j \right\}. \tag{A14}$$

This collection $\mathscr{A}(s)$ captures all the distinct ways in the source events could provide redundant information.

As per Williams and Beer's PID, this set of source events $\mathscr{A}(s)$ is structured. Consider two sets of source events $\alpha, \beta \in \mathscr{A}(s)$. If for every source event $b \in \beta$ there exists a source event $a \in \alpha$ such that $a \subseteq b$, then all of the redundant specificity and ambiguity shared by $b \in \beta$ must include any redundant specificity and ambiguity shared by $a \in \alpha$. Hence, a partial order $\preceq$ can be defined over the elements of the domain $\mathscr{A}(s)$ such that any collection of predictors event coalitions precedes another if and only if the latter provides any information the former provides,

$$\forall \alpha, \beta \in \mathscr{A}(s), \big(\alpha \preceq \beta \iff \forall\, b \in \beta,\, \exists\, a \in \alpha \mid a \subseteq b\big). \tag{A15}$$

Applying this partial ordering to the elements of the domain $\mathscr{A}(s)$ produces a lattice which has the same structure as the redundancy lattice from PID, i.e., the structure of the sources events here is the same as the structure of the sources in PID. (Figure 3 depicts this structure for the case of 2 and 3 predictor variables.) Applying $i_{\cap}^{+}$ to these sources events yields a *specificity lattice* while applying $i_{\cap}^{-}$ yields an *ambiguity lattice*.

Similar to $I_{\cap}$ in PID, the redundancy measures $i_{\cap}^{+}$ or $i_{\cap}^{-}$ can be thought of as a cumulative information functions which integrate the specificity or ambiguity uniquely contributed by each node as one moves up each lattice. In order in evaluate the unique contribution of specificity and ambiguity from each node in the lattice, consider the Möbius inverse [45,46] of $i_{\cap}^{+}$ and $i_{\cap}^{-}$. That is, the specificity and ambiguity of a node α is given by

$$i_{\cap}^{\pm}(\alpha \to t) = \sum_{\beta \preceq \alpha} i_{\cap}^{\pm}(\beta \to t) \qquad \forall\, \alpha,\, \beta \in \mathscr{A}(s). \tag{A16}$$

Thus the unique contributions of *partial specificity* i_{∂}^{+} and *partial ambiguity* i_{∂}^{-} from each node can be calculated recursively from the bottom-up, i.e.,

$$i_{\partial}^{\pm}(\alpha \to t) = i_{\cap}^{\pm}(\alpha \to t) - \sum_{\beta \prec \alpha} i_{\partial}^{\pm}(\beta \to t). \tag{A17}$$

Theorem A1. *Based on the principle of inclusion-exclusion, we have the following closed-from expression for the partial specificity and partial ambiguity,*

$$i_{\partial}^{\pm}(\alpha \to t) = i_{\cap}^{\pm}(\alpha \to t) - \sum_{\varnothing \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1}\, i_{\cap}^{\pm}\big(\bigwedge \gamma \to t\big) \tag{A18}$$

Proof. For $\mathscr{B} \subseteq \mathscr{A}(s)$, define the sub-addative function $f^{\pm}(\mathscr{B}) = \sum_{\beta \in \mathscr{B}} = i^{\pm}(\beta \to t)$. From (A16), we get that $i_{\cap}^{\pm}(\alpha \to t) = f^{\pm}(\downarrow \alpha)$ and

$$i_{\partial}^{\pm}(\alpha \to t) = f^{\pm}(\downarrow \alpha) - f^{\pm}(\dot{\downarrow} \alpha) = f^{\pm}(\downarrow \alpha) - f^{\pm}\big(\bigcup_{\beta \in \alpha^{-}} \downarrow \beta\big). \tag{A19}$$

By the principle of inclusion-exclusion (e.g., see [46], p. 195) we get that

$$i_{\partial}^{\pm}(\alpha \to t) = f^{\pm}(\downarrow \alpha) - \sum_{\varnothing \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1}\, f^{\pm}\big(\bigcap_{\beta \in \gamma} \downarrow \beta\big) \tag{A20}$$

For any lattice L and $A \subseteq L$, we have that $\cap_{a \in A} \downarrow a = \downarrow(\bigwedge A)$ (see [47], p. 57) thus

$$
\begin{aligned}
i_{\partial}^{\pm}(\alpha \to t) &= f^{\pm}(\downarrow \alpha) - \sum_{\varnothing \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1} f^{\pm}\left(\bigwedge \gamma\right) \\
&= f^{\pm}(\downarrow \alpha) - \sum_{\varnothing \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1} i^{\pm}\left(\bigwedge \gamma \to t\right)
\end{aligned}
\tag{A21}
$$

as required.　□

Similarly to PID, the specificity and ambiguity lattices provide a structure for information decomposition—unique evaluation requires a separate definition of redundancy. However, unlike PID (or even PPID), this evaluation requires both a definition of pointwise redundant specificity and pointwise redundant ambiguity.

Appendix B.2. Redundancy Measures on the Lattices

In Section 4.2, Definitions 1 and 2 provided the require measures. This section will prove some of the key properties of these measures when they are applies to the lattices derived in the previous section. The correspondence with the approach taken by Williams and Beer [1,2] continues in this section. However, sources events a_i are used in place of sources A_i and the measures $r_{\min}^{\pm}$ are used in place of $I_{\min}$. Note that the basic concepts from lattice theory and the notion used here are the same as found in ([1], Appendix B).

Theorem 1. *The definitions of r_{min}^{+} and r_{min}^{-} satisfy Axioms 1–4.*

Proof. Axioms 1, 3 and 4 follow trivially from the basic properties of the minimum. The main statement of Axiom 2 also immediately follows from the properties of the minimum; however, there is a need to verify the equality condition. As such, consider a_k such that $a_k \supseteq a_i$ for some $a_i \in \{a_1, \ldots, a_{k-1}\}$. From Postulate 4, we have that $h(a_k) \geq h(a_i)$ and hence that $\min_{a_j \in \{a_1,\ldots,a_k\}} h(a_j) = \min_{a_j \in \{a_1,\ldots,a_{k-1}\}} h(a_j)$, as required for $r_{\min}^{+}$. *Mutatis mutandis*, similar follows for $r_{\min}^{-}$.　□

Theorem 2. *The redundancy measures r_{min}^{+} and r_{min}^{-} increase monotonically on the $\langle \mathscr{A}(s), \preceq \rangle$.*

The proof of this theorem will require the following lemma.

Lemma A1. *The specificity and ambiguity $i^{\pm}(a \to t)$ are increasing functions on the lattice $\langle \mathscr{P}_1(s), \subseteq \rangle$*

Proof. Follows trivially from Postulate 4.　□

Proof of Theorem 2. Assume there exists $\alpha, \beta \in \mathscr{A}(s)$ such that $\alpha \prec \beta$ and $r_{\min}^{\pm}(\beta \to t) < r_{\min}^{\pm}(\alpha \to t)$. By definition, i.e., (23) and (24), there exists $b \in \beta$ such that $i^{\pm}(b \to t) < i^{\pm}(a \to t)$ for all $a \in \alpha$. Hence, by Lemma A1, there does not exist $a \in \alpha$ such that $a \subseteq b$. However, by assumption $\alpha \prec \beta$ and hence there exists $a \in \alpha$ such that $a \subseteq b$, which is a contradiction.　□

Theorem A2. *When using $r_{min}^{\pm}$ in place of the general redundancy measures $i_{\cap}^{\pm}$, we have the following closed-from expression for the partial specificity π^{+} and partial ambiguity π^{-},*

$$
\pi^{\pm}(\alpha \to t) = r_{min}^{\pm}(\alpha \to t) - \max_{\beta \in \alpha^{-}} \min_{b \in \beta} i^{\pm}(b \to t).
\tag{A22}
$$

Proof. Let $i_{\cap}^{+} = r_{\min}^{+}$ and $i_{\cap}^{-} = r_{\min}^{-}$ in the general closed form expression for $i_{\partial}^{\pm}$ in Theorem A1,

$$
\pi^{\pm}(\alpha \to t) = r_{\min}^{\pm}(\alpha \to t) - \sum_{\varnothing \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1} \min_{b \in \bigwedge \gamma} i^{\pm}(b \to t).
\tag{A23}
$$

Since $\alpha \wedge \beta = \underline{\alpha \cup \beta}$ (see [1], Equation (23)), and by Postulate 4, we have that

$$\pi^{\pm}(\alpha \to t) = r_{\min}^{\pm}(\alpha \to t) - \sum_{\emptyset \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1} \min_{\beta \in \gamma} \min_{b \in \beta} i^{\pm}(b \to t). \tag{A24}$$

By the maximum-minimums identity (see [48]), we have that, $\max \alpha^{-} = \sum_{\emptyset \neq \gamma \subseteq \alpha^{-}} (-1)^{|\gamma|-1} \min \gamma$, and hence

$$\pi^{\pm}(\alpha \to t) = r_{\min}^{\pm}(\alpha \to t) - \max_{\beta \in \alpha^{-}} \min_{b \in \beta} i^{\pm}(\alpha \to t). \tag{A25}$$

as required. $\square$

Theorem 3. *The atoms of partial specificity π^{+} and partial ambiguity π^{-} evaluated using the measures r_{min}^{+} and r_{min}^{-} on the specificity and ambiguity lattices (respectively), are non-negative.*

Proof. It $\alpha = \perp$, the $\pi^{\pm}(\alpha \to t) = r_{\min}^{\pm} \geq 0$ by the non-negativity of entropy. If $\alpha \neq \perp$, assume there exists $\alpha \in \mathscr{A}(s) \backslash \{\perp\}$ such that $\pi^{\pm}(\alpha \to t) < 0$. By Theorem A2,

$$\pi^{\pm}(\alpha \to t) = \min_{a \in \alpha} i^{\pm}(a \to t) - \max_{\beta \in \alpha^{-}} \min_{b \in \beta} i^{\pm}(b \to t). \tag{A26}$$

From this it can be seen that there must exist $\beta \in \alpha^{-}$ such that for all $b \in \beta$, we have that $i^{\pm}(a \to t) < i^{\pm}(b \to t)$ for some $a \in \alpha$. By Postulate 4 there does not exist $b \in \beta$ such that $b \subset a$. However, since by definition, $\beta \prec \alpha$ there exists $b \in \beta$ such that $b \subset a$, which is a contradiction. $\square$

Theorem 4. *The atoms of partial average information Π evaluated by recombining and averaging $\pi^{\pm}$ are not non-negative.*

Proof. The proof is by the counter-example using RDNERR. $\square$

Appendix B.3. Target Chain Rule

By using the appropriate conditional probabilities in Definitions 1 and 2, one can easily obtain the conditional pointwise redundant specificity,

$$r_{\min}^{+}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}) = \min_{a_{i}} h(a_{i}|t_{2}), \tag{A27}$$

or the conditional pointwise redundant ambiguity,

$$r_{\min}^{-}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}) = \min_{a_{j}} h(a_{j}|t_{1,2}). \tag{A28}$$

As per (21) these could be recombined, e.g., via (21), to obtain the conditional redundant information,

$$r_{\min}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}) = r_{\min}^{+}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}) - r_{\min}^{-}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}). \tag{A29}$$

The relationship between the regular forms and the conditional forms of the redundant specificity and redundant ambiguity has some important consequences.

Proposition A3. *The conditional pointwise redundant specificity provided by $a_{1}, \ldots, a_{k}$ about t_{1} given t_{2} is equal to pointwise redundant ambiguity provided by $a_{1}, \ldots, a_{k}$ about t_{2} with the conditioned variable,*

$$r_{min}^{+}(a_{1}, \ldots, a_{k} \to t_{1}|t_{2}) = r_{min}^{-}(a_{1}, \ldots, a_{k} \to t_{2}). \tag{A30}$$

Proof. By (24) and (A27). $\square$

Proposition A4. *The pointwise redundant specificity provided by $a_1, \ldots, a_k$ is independent of the target event and even the target variable itself,*

$$r^+_{min}(a_1, \ldots, a_k \to t_1) = r^+_{min}(a_1, \ldots, a_k \to t_2) \qquad \forall\, t_1, t_2, T_1, T_2. \tag{A31}$$

Proof. By inspection of (23). □

Proposition A5. *The conditional pointwise redundant ambiguity provided by $a_1, \ldots, a_k$ about t_1 given t_2 is equal to the pointwise redundant ambiguity provided by $a_1, \ldots, a_k$ about $t_{1,2}$,*

$$r^-_{min}(a_1, \ldots, a_k \to t_1 | t_2) = r^-_{min}(a_1, \ldots, a_k \to t_{1,2}). \tag{A32}$$

Proof. By (24) and (A28). □

Note that specificity itself is not a function of the target event or variable. Hence, all of the target dependency is bound up in the ambiguity. Now consider the following.

Theorem 5 (Pointwise Target Chain Rule). *Given the joint target realisation $t_{1,2}$, the pointwise redundant information r_{min} satisfies the following chain rule,*

$$\begin{aligned} r_{min}(a_1, \ldots, a_k \to t_{1,2}) &= r_{min}(a_1, \ldots, a_k \to t_1) + r_{min}(a_1, \ldots, a_k \to t_2 | t_1), \\ &= r_{min}(a_1, \ldots, a_k \to t_2) + r_{min}(a_1, \ldots, a_k \to t_1 | t_2). \end{aligned} \tag{25}$$

Proof. Starting from r_{min}, by Corollary A4 and Corollary A5 we get that

$$\begin{aligned} r_{min}(a_1, \ldots, a_k \to t_{1,2}) &= r^+_{min}(a_1, \ldots, a_k \to t_{1,2}) - r^-_{min}(a_1, \ldots, a_k \to t_{1,2}), \\ &= r^+_{min}(a_1, \ldots, a_k \to t_1) - r^-_{min}(a_1, \ldots, a_k \to t_2 | t_1), \end{aligned} \tag{A33}$$

Then, by Corollary A3 we get that

$$\begin{aligned} r_{min}(a_1, \ldots, a_k \to t_{1,2}) &= r^+_{min}(a_1, \ldots, a_k \to t_1) - r^-_{min}(a_1, \ldots, a_k \to t_1) \\ &\quad + r^-_{min}(a_1, \ldots, a_k \to t_1) - r^-_{min}(a_1, \ldots, a_k \to t_2 | t_1), \\ &= r^+_{min}(a_1, \ldots, a_k \to t_1) - r^-_{min}(a_1, \ldots, a_k \to t_1) \\ &\quad + r^+_{min}(a_1, \ldots, a_k \to t_2 | t_1) - r^-_{min}(a_1, \ldots, a_k \to t_2 | t_1), \\ &= r_{min}(a_1, \ldots, a_k \to t_1) + r_{min}(a_1, \ldots, a_k \to t_2 | t_1), \end{aligned} \tag{A34}$$

as required for the first equality in (25). *Mutatis mutandis*, we obtain the second equality in (25). □

Theorem 6. *The target chain rule, identity property and local positivity, cannot be simultaneously satisfied.*

Proof. Consider the probability distribution Tʙᴄ, and in particular, the isomorphic probability distributions $P(T_{1,2})$ and $P(T_{1,3})$. By the identity property,

$$U(S_1 \backslash S_2 \to T_{1,2}) = 1 \text{ bit}, \qquad\qquad U(S_2 \backslash S_1 \to T_{1,2}) = 1 \text{ bit}, \tag{A35}$$

and hence, $R(S_1, S_2 \to T_{1,2}) = 0$ bit. On the other hand, by local positivity,

$$C(S_1, S_2 \to T_3) = 1 \text{ bit}, \qquad\qquad R(S_1, S_2 \to T_1 | T_3) = 1 \text{ bit} \tag{A36}$$

Then by the target chain rule,

$$C(S_1, S_2 \to T_{1,3}) = 1 \text{ bit} \qquad\qquad R(S_1, S_2 \to T_{1,3}) = 1 \text{ bit}, \tag{A37}$$

Finally, since $P(T_{1,2})$ is isomorphic to $P(T_{1,3})$ we have that, $R(S_1, S_2 \rightarrow T_{1,3}) = R(S_1, S_2 \rightarrow T_{1,2})$, which is a contradiction.　□

Theorem 6 can be informally generalised as follows: it is not possible to simultaneously satisfy the target chain rule, the identity property, and have only $C(S_1, S_2 \rightarrow T) = 1$ bit in the probability distribution XOR without having negative (average) PI atoms in probability distributions where there is no ambiguity from any source. To see this, again consider decomposing the isomorphic probability distributions $P(T_{1,2})$ and $P(T_{1,3})$. In line with (A35), decomposing $T_{1,2}$ via the identity property yields $C(S_1, S_2 \rightarrow T_{1,2}) = 0$ bit. On the other hand, decomposing $T_{1,3}$ yields $C(S_1, S_2 \rightarrow T_3) = 1$ bit. Since $P(T_{1,2})$ is isomorphic to $P(T_{1,3})$, the target chain rule requires that,

$$C(S_1, S_2 \rightarrow T_1 | T_3) = -1 \text{ bit}, \quad U(S_1 \backslash S_2 \rightarrow T_1 | T_3) = 1 \text{ bit}, \quad U(S_2 \backslash S_1 \rightarrow T_1 | T_3) = 1 \text{ bit}. \quad \text{(A38)}$$

That is, one would have to accept the negative (average) PI atom $C(S_1, S_2 \rightarrow T_1 | T_3) = -1$ bit despite the fact that there are no non-zero pointwise ambiguity terms upon splitting any of $i(s_1; t_1 | t_3)$, $i(s_2; t_1 | t_3)$ and $i(s_{1,2}; t_1 | t_3)$ into specificity and ambiguity. Although this does not constitute a formal proof that the identity property is incompatible with the target chain rule, one would have to accept and find a way to justify $C(S_1, S_2 \rightarrow T_1 | T_3) = -1$ bit. Since there is no ambiguity in $i(s_1; t_1 | t_3)$, $i(s_2; t_1 | t_3)$ and $i(s_{1,2}; t_1 | t_3)$, this result is not reconcilable within the framework of specificity and ambiguity.

Appendix C. Additional Example Probability Distributions

Appendix C.1. Probability Distribution TBEP

Figure A1 shows the probability distribution three bit–even parity (TBEP) which considers binary predictors variables S_1, S_2 and S_3 which are constrained such that together their parity is even. The target variable T is simply a copy of the predictors, i.e., $T = T_{1,2,3} = (T_1, T_2, T_3)$ where $T_1 = S_1$, $T_2 = S_2$ and $T_3 = S_3$. (Equivalently, the target can be represented by any four state variable T.) It was introduced by Bertschinger et al. [5] and revisited by Rauh et al. [13] who (as mentioned in Section 5.5) used it to prove the following by counter-example: there is no measure of redundant average information for more than two predictor variables which simultaneously satisfies the Williams and Beer Axioms, the identity property, and local positivity. The measures I_{red}, $\widetilde{UI}$ and $\mathcal{S}_{\text{VK}}$ these properties. Hence, this probability distribution which has been used to demonstrate that these measures are not consistent with the PID framework in the general case of an arbitrary number of predictor variables.

This example is similar to TBC in the several ways. Firstly, due to the symmetry in the probability distribution, each realisation will have the same pointwise decomposition. Secondly, there is an isomorphism between the probability distributions $P(T)$ and $P(S_1, S_2, S_3)$, and hence the pointwise ambiguity provided by any (individual or joint) predictor event is 0 bit (since given t, one knows s_1, s_2 and s_3). Thirdly, the individual predictor events s_1, s_2 and s_3 each exclude 1/2 of the total probability mass in $P(T)$ and so each provide 1 bit of pointwise specificity. Thus, there is 1 bit of three-way redundant, pointwise specificity in each realisation. Fourthly, the joint predictor event $s_{1,2,3}$ excludes 3/4 of the total probability mass, providing 2 bit of pointwise specificity (which is similar to TBC). However, unlike TBC, one could consider the three joint predictor events $s_{1,2}$, $s_{1,3}$ and $s_{2,3}$. These joint pairs also exclude 3/4 of the total probability mass each, and hence also each provide 2 bit of pointwise specificity. As such, there is 1 bit of pointwise, three-way redundant, pairwise complementary specificity between these three joint pairs of source events, in addition to the 1 bit of three-way redundant, pointwise specificity. Finally, putting this together and averaging over all realisations, TBEP consists of 1 bit of three-way redundant information and 1 bit of three-way redundant, pairwise complementary information. The resultant average decomposition is the same as the decomposition induced by I_{min} [5].
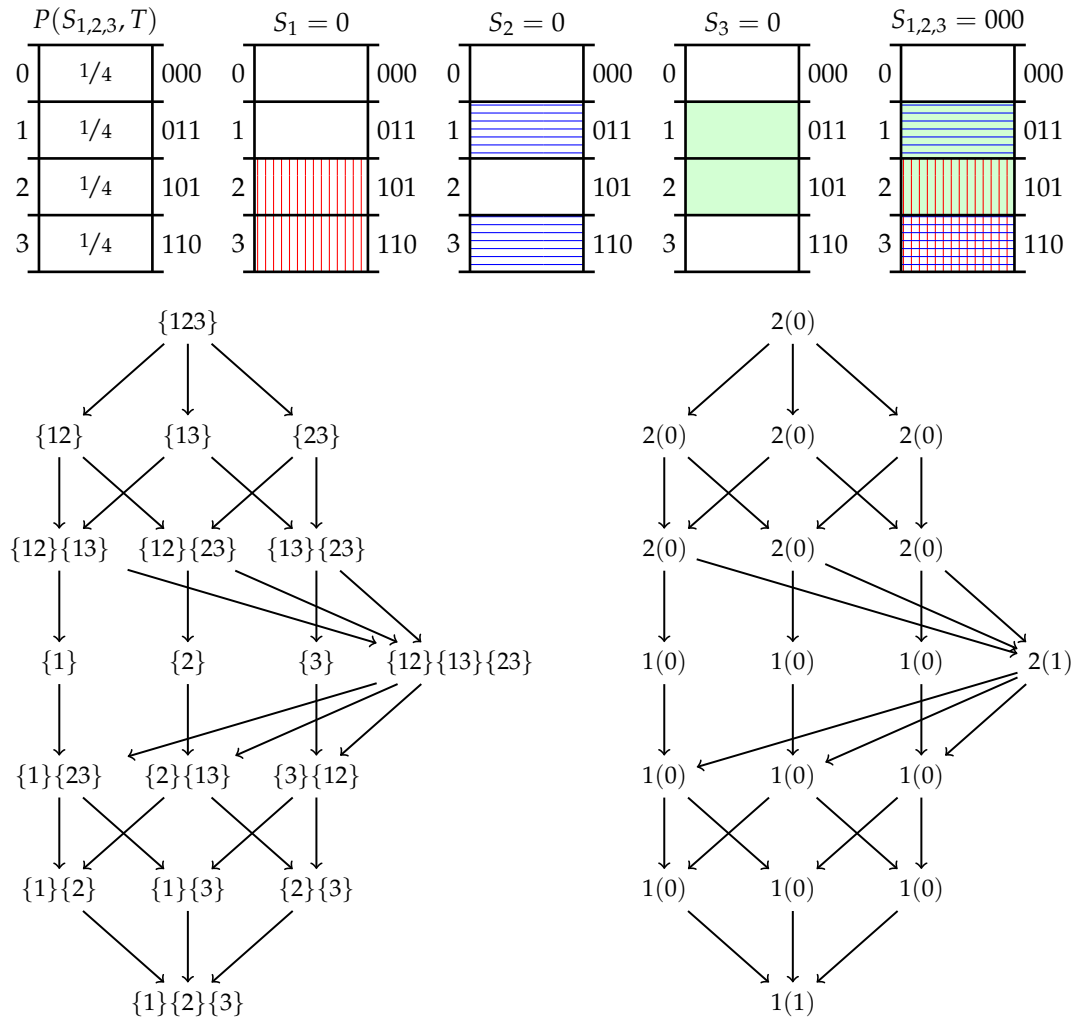
Figure A1. Example TBEP. (**Top**) probability mass diagram for realisation $(S_1=0, S_2=0, S_3=0, T=000)$; (**Bottom left**) With three predictors, it is convenient to represent to decomposition diagrammatically. This is especially true TBEP as one only needs to consider the specificity lattice for one realisation; (**Bottom right**) The specificity lattice for the realisation $(S_1=0, S_2=0, S_3=0, T=000)$. For each source event the left value corresponds to the value of $i_\cap^+$, evaluated using r_{min}^+, while the right value (surrounded by parenthesis) corresponds to the partial information π^+.

Appendix C.2. Probability Distribution UNQ

Figure A2 shows the decomposition of the probability distribution *unique* (UNQ). Note that this probability distribution corresponds to RDNERR where the error probability $\varepsilon = 1/2$, and hence the similarity in the resultant distributions. The results may initially seem unusual, that the predictor S_1 is not uniquely informative since $U(S_1 \backslash S_2 \rightarrow T) = 0$ bit as one might intuitively expect. Rather it is deemed to be redundantly informative $RI = 1$ bit with the predictor S_2 which is also uniquely misinformative $U(S_2 \backslash S_1 \rightarrow T) = -1$ bit. This is because both S_1 and S_2 provide $I^+(S_1 \rightarrow T) = I^+(S_2 \rightarrow T) = 1$ bit of specificity; however the information provided by S_2 is unique in that the 1 bit provided is not "useful" ([42], p. 21) and hence $I(S_2 \rightarrow T) = 1$ bit while $I(S_2 \rightarrow T) = 1$ bit. Finally, the complementary information $C(S_1, S_2 \rightarrow T) = 1$ bit is required by the decomposition in order to balance this 1 bit of unique ambiguity. The results in this example partly explain our preference for term *complementary information* as opposed to *synergistic information*—while $C(S_1, S_2 \rightarrow T) = 1$ bit is readily explainable, it would be dubious to refer to this as *synergy* given that S_1 enables perfect predictions of T without any knowledge of S_2.
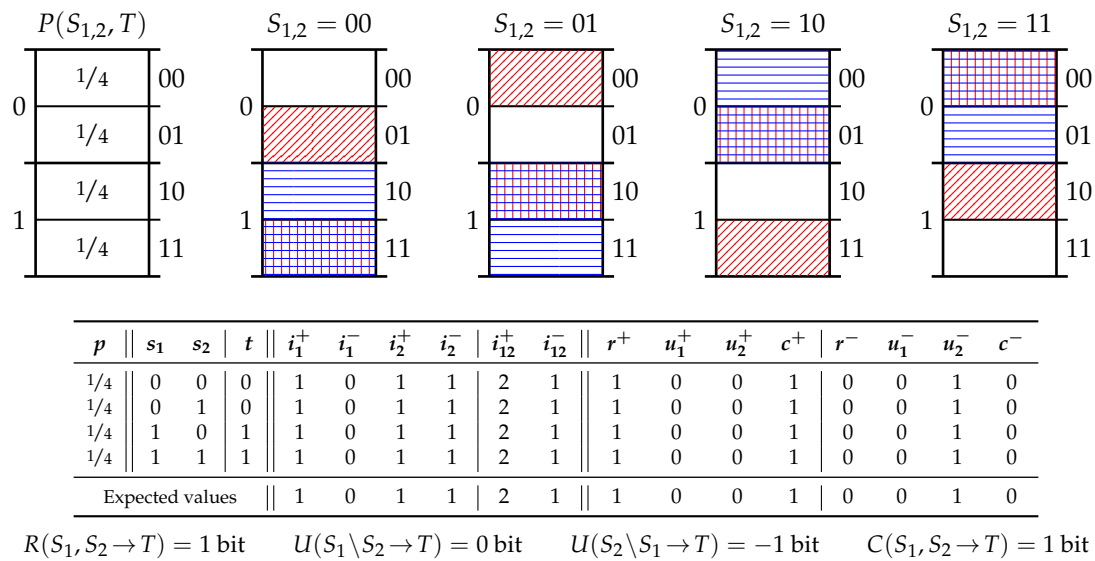
p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	0	1	0	1	1	2	1	1	0	0	1	0	0	1	0
1/4	0	1	0	1	0	1	1	2	1	1	0	0	1	0	0	1	0
1/4	1	0	1	1	0	1	1	2	1	1	0	0	1	0	0	1	0
1/4	1	1	1	1	0	1	1	2	1	1	0	0	1	0	0	1	0
Expected values				1	0	1	1	2	1	1	0	0	1	0	0	1	0

$R(S_1, S_2 \to T) = 1$ bit $\quad U(S_1 \backslash S_2 \to T) = 0$ bit $\quad U(S_2 \backslash S_1 \to T) = -1$ bit $\quad C(S_1, S_2 \to T) = 1$ bit

Figure A2. Example UNQ. (**Top**) the probability mass diagrams for every single possible realisation; (**Middle**) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4); (**Bottom**) the atoms of (average) partial infromation obtained through recombination of the averages.

Appendix C.3. Probability Distribution AND

Figure A3 shows the decomposition of the probability distribution *and* (AND). Note that the probability distribution *or* (OR) has the same decomposition as the target distributions are isomorphic.
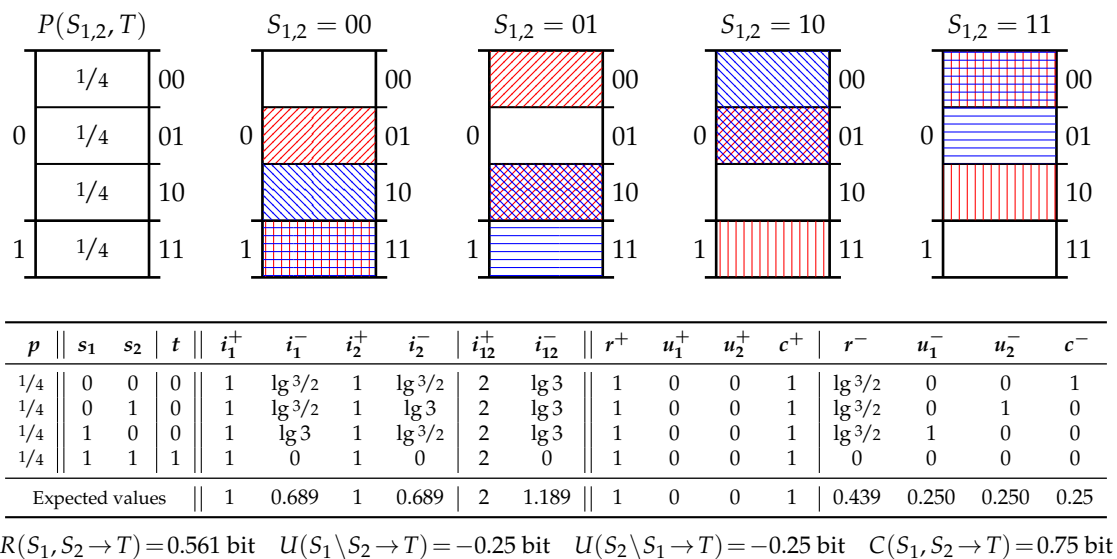

p	s_1	s_2	t	i_1^+	i_1^-	i_2^+	i_2^-	i_{12}^+	i_{12}^-	r^+	u_1^+	u_2^+	c^+	r^-	u_1^-	u_2^-	c^-
1/4	0	0	0	1	$\lg 3/2$	1	$\lg 3/2$	2	$\lg 3$	1	0	0	1	$\lg 3/2$	0	0	1
1/4	0	1	0	1	$\lg 3/2$	1	$\lg 3$	2	$\lg 3$	1	0	0	1	$\lg 3/2$	0	1	0
1/4	1	0	0	1	$\lg 3$	1	$\lg 3/2$	2	$\lg 3$	1	0	0	1	$\lg 3/2$	1	0	0
1/4	1	1	1	1	0	1	0	2	0	1	0	0	1	0	0	0	0
Expected values				1	0.689	1	0.689	2	1.189	1	0	0	1	0.439	0.250	0.250	0.25

$R(S_1, S_2 \to T) = 0.561$ bit $\quad U(S_1 \backslash S_2 \to T) = -0.25$ bit $\quad U(S_2 \backslash S_1 \to T) = -0.25$ bit $\quad C(S_1, S_2 \to T) = 0.75$ bit

Figure A3. Example AND. (**Top**) the probability mass diagrams for every single possible realisation; (**Middle**) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4); (**Bottom**) the atoms of (average) partial infromation obtained through recombination of the averages.

References and Notes

1. Williams, P.L.; Beer, R.D. Information decomposition and synergy. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.

2. Williams, P.L.; Beer, R.D. Indiana University. Decomposing Multivariate Information. Privately communicated, 2010. This unpublished paper is highly similar to [1]. Crucially, however, this paper derives the redundancy lattice from the W&B Axioms 1–3 of Section 1. In contrast, [1] derives the redundancy lattice as a property of the particular measure $I_{\min}$.

3. Olbrich, E.; Bertschinger, N.; Rauh, J. Information decomposition and synergy. *Entropy* **2015**, *17*, 3501–3517.

4. Lizier, J.T.; Flecker, B.; Williams, P.L. Towards a synergy-based approach to measuring information modification. In Proceedings of the IEEE Symposium on Artificial Life (ALife), Singapore, 16–19 April 2013; pp. 43–51.

5. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information—New insights and problems in decomposing information in complex systems. In Proceedings of the European Conference on Complex Systems, Brussels, Belgium, 3–7 September 2012; Springer: Cham, The Netherland, 2013; pp. 251–269.

6. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130.

7. Griffith, V.; Chong, E.K.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection information based on common randomness. *Entropy* **2014**, *16*, 1985–2000.

8. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

9. Fano, R. *Transmission of Information*; The MIT Press: Cambridge, MA, USA, 1961.

10. Harder, M. Information driven self-organization of agents and agent collectives. Ph.D. Thesis, University of Hertfordshire, Hertfordshire, UK, 2013.

11. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.

12. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 9, pp. 159–190.

13. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2232–2236.

14. Perrone, P.; Ay, N. Hierarchical Quantification of Synergy in Channels. *Front. Robot. AI* **2016**, *2*, 35.

15. Griffith, V.; Ho, T. Quantifying redundant information in predicting a target random variable. *Entropy* **2015**, *17*, 4644–4653.

16. Rosas, F.; Ntranos, V.; Ellison, C.J.; Pollin, S.; Verhelst, M. Understanding interdependency through complex information sharing. *Entropy* **2016**, *18*, 38.

17. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, 052802.

18. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **2017**, *19*, 318.

19. Ince, R.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv* **2017**, arXiv:1702.01591.

20. Chicharro, D.; Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. *Entropy* **2017**, *19*, 71.

21. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N. On Extractable Shared Information. *Entropy* **2017**, *19*, 328.

22. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N.; Wolpert, D. Coarse-Graining and the Blackwell Order. *Entropy* **2017**, *19*, 527.

23. Rauh, J. Secret sharing and shared information. *Entropy* **2017**, *19*, 601.

24. Faes, L.; Marinazzo, D.; Stramaglia, S. Multiscale information decomposition: Exact computation for multivariate Gaussian processes. *Entropy* **2017**, *19*, 408.

25. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant components of synergy, redundancy, and unique information among three variables. *Entropy* **2017**, *19*, 451.

26. James, R.G.; Crutchfield, J.P. Multivariate dependence beyond shannon information. *Entropy* **2017**, *19*, 531.

27. Makkeh, A.; Theis, D.O.; Vicente, R. Bivariate Partial Information Decomposition: The Optimization Perspective. *Entropy* **2017**, *19*, 530.

28. Kay, J.W.; Ince, R.A.; Dering, B.; Phillips, W.A. Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction. *Entropy* **2017**, *19*, 560.

29. Angelini, L.; de Tommaso, M.; Marinazzo, D.; Nitti, L.; Pellicoro, M.; Stramaglia, S. Redundant variables and Granger causality. *Phys. Rev. E* **2010**, *81*, 037201.

30. Stramaglia, S.; Angelini, L.; Wu, G.; Cortes, J.M.; Faes, L.; Marinazzo, D. Synergetic and redundant information flow detected by unnormalized Granger causality: Application to resting state fMRI. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 2518–2524.

31. Ghazi-Zahedi, K.; Langer, C.; Ay, N. Morphological computation: Synergy of body and brain. *Entropy* **2017**, *19*, 456.

32. Maity, A.K.; Chaudhury, P.; Banik, S.K. Information theoretical study of cross-talk mediated signal transduction in MAPK pathways. *Entropy* **2017**, *19*, 469.

33. Tax, T.; Mediano, P.A.; Shanahan, M. The partial information decomposition of generative neural network models. *Entropy* **2017**, *19*, 474.

34. Wibral, M.; Finn, C.; Wollstadt, P.; Lizier, J.T.; Priesemann, V. Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition. *Entropy* **2017**, *19*, 494.

35. Woodward, P.M. *Probability and Information Theory: With Applications to Radar*; Pergamon Press: Oxford, UK, 1953.

36. Woodward, P.M.; Davies, I.L. Information theory and inverse probability in telecommunication. *Proc. IEE-Part III Radio Commun. Eng.* **1952**, *99*, 37–44.

37. Gray, R.M. *Probability, Random Processes, and Ergodic Properties*; Springer: New York, NY, USA, 1988.

38. Martin, N.F.; England, J.W. *Mathematical Theory of Entropy*; Cambridge University Press: Cambridge, UK, 1984.

39. Finn, C.; Lizier, J.T. Probability Mass Exclusions and the Directed Components of Pointwise Mutual Information. *arXiv* **2018**, arXiv:1801.09223.

40. Kelly, J.L. A new interpretation of information rate. *Bell Labs Tech. J.* **1956**, *35*, 917–926.

41. Ash, R. *Information Theory*; Interscience tracts in pure and applied mathematics; Interscience Publishers: Geneva, Switzerland, 1965.

42. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Champaign, IL, USA, 1998.

43. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.

44. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1988.

45. Rota, G.C. On the foundations of combinatorial theory I. Theory of Möbius functions. *Probab. Theory Relat. Field* **1964**, *2*, 340–368.

46. Stanley, R.P. Enumerative Combinatorics. In *Cambridge Studies in Advanced Mathematics*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2012; Volume 1.

47. Davey, B.A.; Priestley, H.A. *Introduction to Lattices and Order*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002.

48. Ross, S.M. *A First Course in Probability*, 8th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2009.