# Information Flows? A Critique of Transfer Entropies

R. G. James
N. Barnett
J. P. Crutchfield

SFI WORKING PAPER: 2016-01-001



SANTA FE INSTITUTE

Santa Fe Institute Working Paper 15-12-XXX
arXiv:1512.XXXXX

# Information Flows? A Critique of Transfer Entropies

Ryan G. James,[1,2,*] Nix Barnett,[1,3,†] and James P. Crutchfield[1,2,3,‡]

[1] *Complexity Sciences Center*
[2] *Physics Department*
[3] *Mathematics Department,*
*University of California at Davis, One Shields Avenue, Davis, CA 95616*

(Dated: December 20, 2015)

A central task in analyzing complex dynamics is to determine the loci of information storage and the communication topology of information flows within a system. Over the last decade and a half, diagnostics for the latter have come to be dominated by the *transfer entropy*. Via straightforward examples, we show that it and a derivative quantity, the *causation entropy*, do not, in fact, quantify the flow of information. At one and the same time they can overestimate flow or underestimate influence. We isolate why this is the case and propose alternate measures for information flow. An auxiliary consequence reveals that the proliferation of networks as a now-common theoretical model for large-scale systems in concert with the use of transfer-like entropies has shoehorned dyadic relationships into our structural interpretation of the organization and behavior of complex systems, despite the occurrence of polyadic dependencies. The net result is that much of the sophisticated organization of complex systems goes undetected.

**Keywords**: stochastic process, transfer entropy, causation entropy, partial information decomposition, network science

PACS numbers: 05.45.-a 89.75.Kd 89.70.+c 05.45.Tp 02.50.Ey

## I. INTRODUCTION

An important task in understanding a complex system is determining its information dynamics and information architecture—what mechanisms generate information, where is that information stored, and how is it transmitted within a system? While this pursuit goes back perhaps as far as Shannon's foundational work on communication [1], in many ways it was Kolmogorov [2–6] who highlighted information transmitted from the micro- to the macroscales as a statistic central to monitoring the behavior of complex systems. Later, Lin showed that it is key to understanding network controllability [7] and Shaw speculated that "information flow" between information sources and sinks is a necessary descriptive framework for spatially extended chaotic systems—an alternative to narratives based on tracking energy flows [8, Sec. 14].

A common thread in all these works is quantifying the flow of information. To facilitate our discussion, let's first consider an intuitive framing: Information flow between two processes, from say $X$ to $Y$, means that information is present in random variable $Y_t$ at time $t$ the cause of which can be solely attributed to information in $X$'s past. If information can be solely attributed in such a manner, we refer to it as *localized*.

Ostensibly to measure information flow—and notably much later than the above efforts—Schreiber introduced the transfer entropy [9] as the information shared between $X$'s past and the present $Y_t$, conditioning on information from $Y$'s past. Perhaps not surprisingly, given the broad and pressing need to probe the organization of modern life's increasingly complex systems, its impact has been substantial—over the last decade and a half, its introduction alone garnered an average of 100 citations per year.

The primary goal of the following is to show that the transfer entropy does not, in fact, measure information flow, specifically in that it attributes an information source to influences that are not localizable and so not flows. We draw out the interpretational errors, some quite subtle, that result, including overestimating flow, underestimating influence and, more generally, misidentifying structure when modeling complex systems as networks.

Identifying shortcomings in the transfer entropy is not new. Smirnov [10] pointed out three: Two relate to how it responds to using undersampled empirical distributions and are therefore not conceptual issues with the measure. The third, however, was its inability to differentiate indirect influences from direct influences. This weakness motivated Sun and Bollt to propose the causation entropy [11]. While their measure does allow differentiating

* rgjames@ucdavis.edu
† nix@math.ucdavis.edu
‡ chaos@ucdavis.edu

between direct and indirect effects, via the addition of a third hidden variable, it too ascribes an information source to unlocalizable influences.

Our exposition develops as follows. Section II reviews notation and the information theory needed. Section III considers two rather similar examples—one involving the influences between two processes and the other influences between three. In Section IV we discuss why the transfer entropy fails to capture information flow. Finally, Section V closes by discussing a distinctive philosophy underlying our critique and then turns to speculate about possible resolutions and to concerns about modeling practice in network science.

## II. BACKGROUND

Following standard notation [12], we denote random variables with capital letters and their associated outcomes using lower case. For example, the observation of a coin flip might be denoted $X$, while the coin actually landing on Heads or Tails would be $x$. Concerned with temporal processes, we subscript a random variable with a time index; *e.g.*, the random variable representing a coin flip at time $t$ is denoted $X_t$. We denote a temporally contiguous block of random variables (a time series) using a Python slice-like notation $X_{i:j} = X_i X_{i+1} \ldots X_{j-1}$, where the final index is exclusive. We denote $X_t$'s being distributed according to $\Pr(X_t)$ as $X_t \sim \Pr(X_t)$. We also assume familiarity with basic information measures, specifically the Shannon entropy $\mathrm{H}[X]$, mutual information $\mathrm{I}[X:Y]$, and their conditional forms $\mathrm{H}[X \mid Z]$ and $\mathrm{I}[X:Y \mid Z]$ [12].

The *transfer entropy* $\mathrm{T}_{X\to Y}$ from time series $X$ to time series $Y$ is the information shared between $X$'s past and $Y$'s present, given knowledge of $Y$'s past [9]:

$$\mathrm{T}_{X\to Y} = \mathrm{I}[Y_t : X_{0:t} \mid Y_{0:t}] \ . \tag{1}$$

Intuitively, this quantifies how much better one predicts $Y_t$ using both $X_{0:t}$ and $Y_{0:t}$ over using $Y_{0:t}$ alone. Assuming the distributions used are not undersampled, a nonzero value of the transfer entropy certainly implies a kind of influence of $X$ on $Y$. Our questions are: Is this influence necessarily via information flow? Is it necessarily direct?

Addressing the last question, the *causation entropy* $\mathcal{C}_{X\to Y|(Y,Z)}$ is similar to the transfer entropy, but conditions on the past of a third (or more) time series [11]:

$$\mathcal{C}_{X\to Y|(Y,Z)} = \mathrm{I}[Y_t : X_{0:t} \mid Y_{0:t}, Z_{0:t}] \ . \tag{2}$$

The primary improvement over $\mathrm{T}_{X\to Y}$ is its ability to determine if a dependency is indirect (*i.e.*, mediated by the third process $Z$) or not. Consider, for example, the following system $X \to Z \to Y$: variable $X$ influences $Z$ and $Z$ in turn influences $Y$. Here, any influence that $X$ has on $Y$ must pass through $Z$. In this case, the transfer entropy $\mathrm{T}_{X\to Y} > 0$ bit even though $X$ does not directly influence $Y$. The causation entropy $\mathcal{C}_{X\to Y|(Y,Z)} = 0$ bit, however, due to conditioning on $Z$.

## III. EXAMPLES

Many concerns and pitfalls in applying information measures comes not in their definition, estimation, or derivation of associated properties. Rather, many arise in *interpreting* results. These can even be the most subtle kind of problem encountered, as we now demonstrate.

### A. Two Time Series

Consider two time series, say $X$ and $Y$, given by the probability laws:

$$X_t \sim \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} ,$$

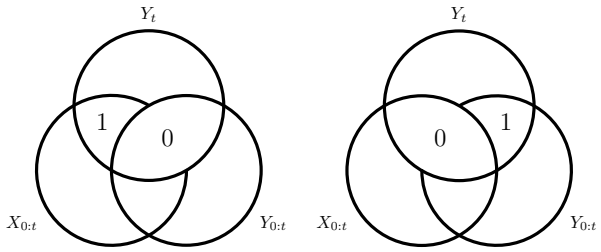
$$Y_0 \sim \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} , \text{ and}$$

$$Y_t = X_{t-1} \oplus Y_{t-1} \ ;$$

that is, $X_t$ and $Y_0$ are independent and take values 0 and 1 with equal probability, and $y_t$ is the *Exclusive Or* of the prior values $x_{t-1}$ and $y_{t-1}$. By a straightforward calculation we find that $\mathrm{T}_{X\to Y} = 1$ bit. Does this mean that one bit of information is being *transferred* from $X$ to $Y$ at each time step? Let's take a closer look.

We first observe that the amount of information in $Y_t$ is $\mathrm{H}[Y_t] = 1$ bit. Therefore, the uncertainty in $Y_t$ can be reduced by at most 1 bit. Furthermore, the information shared by $Y_t$ and the prior behavior of the two time series is $\mathrm{I}[Y_t : (X_{0:t}, Y_{0:t})] = 1$ bit. And so, the 1 bit of $Y_t$'s uncertainty in fact can be reduced by the prior observations of both time series.

How much does $Y_{0:t}$ alone help us predict $Y_t$? We quantify this using mutual information. Since $\mathrm{I}[Y_t : Y_{0:t}] = 0$ bit, the variables are independent: $Y_{0:t}$ alone does not help in predicting $Y_t$. However, knowing $Y_{0:t}$, how much does $X_{0:t}$ help in predicting $Y_t$? The conditional mutual information $\mathrm{I}[Y_t : X_{0:t} \mid Y_{0:t}] = 1$ bit—the transfer entropy we just computed—quantifies this. This situation is graphically analyzed via the *I-Diagram* [13] in Fig. 1a.




(a) $Y_{0:t}$ alone does not predict $Y_t$. (The ⬭ -shaped region $\mathrm{I}[Y_{0:t} : Y_t] = 0$ bit.) However, when used in conjunction with $X_{0:t}$, they completely predict its value. (The ▽ -shaped region $\mathrm{I}[X_{0:t} : Y_t \mid Y_{0:t}] = 1$ bit.)

(b) $X$'s past $X_{0:t}$ alone does not aid in predicting $Y_t$. (The ⬭ -shaped region $\mathrm{I}[X_{0:t} : Y_t] = 0$ bit.) However, given knowledge of $X_{0:t}$, then $Y_{0:t}$ can predict $Y_t$. (The ▽ -shaped region $\mathrm{I}[Y_{0:t} : Y_t \mid X_{0:t}] = 1$ bit.)

FIG. 1. Two complementary ways to view the information shared between $X_{0:t}, Y_{0:t}$, and $Y_t$. In each I-Diagram, a circle represents a random variable whose area represents the random variable's entropy. Overlapping regions are information that is shared. The transfer entropy is a conditional mutual information; a region where two random variables overlap, but falling outside the random variable being conditioned on.

To obtain a more complete picture of the information dynamics under consideration, let's reverse the order in which the time series are queried. The mutual information $\mathrm{I}[Y_t : X_{0:t}] = 0$ bit tells us that the $X$ time series alone does not help predict $Y_t$. However, the conditional mutual information $\mathrm{I}[Y_t : Y_{0:t} \mid X_{0:t}] = 1$ bit. And so, from this point of view it is $Y$'s past that helps predict $Y_t$, contradicting the preceding analysis. This complementary situation is presented diagrammatically in Fig. 1b.

How can we rectify the seemingly inconsistent conclusions drawn by these two lines of reasoning[1]? The answer is quite straightforward: the 1 bit of information about $Y_t$ does not come from *either* time series individually, but rather from *both* of them simultaneously. In short, the 1 bit of reduction in uncertainty $\mathrm{H}[Y_t]$ should not be *localized* to either time series. The transfer entropy, however, erroneously localizes this information to $X_{0:t}$. In light of this, the transfer entropy *overestimates* information flow.

### B. Three Time Series

Our second example parallels the first. Before we considered the case where *one* of two time series is determined by the past of *both*, we now consider the case where *two* time series determine a *third*, again via an Exclusive Or operation. The probability laws governing them are:

$$X_t \sim \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} ,$$

$$Y_t \sim \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} , \text{ and}$$

$$Z_t = X_{t-1} \oplus Y_{t-1} ,$$

with $z_0$'s value being irrelevant. Unlike the prior example (Section III A), the transfer entropy from either $X$ or $Y$ to $Z$ is zero: $\mathrm{T}_{X \to Z} = \mathrm{T}_{Y \to Z} = 0$ bit, and it therefore *underestimates* influence. Furthermore, all of the relevant pairwise mutual informations are zero: $\mathrm{I}[Z_t : X_{0:t}] = \mathrm{I}[Z_t : Y_{0:t}] = \mathrm{I}[Z_t : Z_{0:t}] = 0$ bit. The time series are pairwise independent.

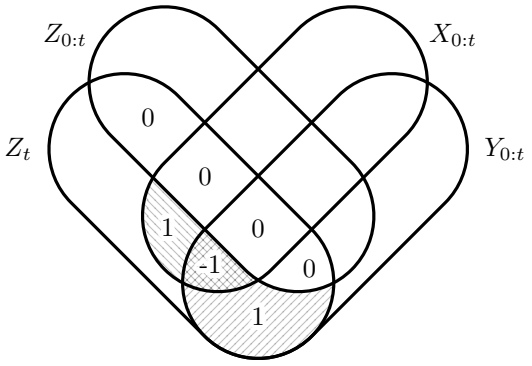



FIG. 2. I-Diagram depicting both transfer entropies and causation entropies for three times series $X$, $Y$, and $Z$. $\mathrm{T}_{X \to Z} = 0$ bit corresponds to the two regions shaded with south-east sloping lines, and $\mathrm{T}_{Y \to Z} = 0$ bit the two regions shaded with north-east sloping lines. $\mathcal{C}_{X \to Z|(Y,Z)} = 1$ bit is the region containing only south-east sloping lines and, similarly, $\mathcal{C}_{Y \to Z|(X,Z)} = 1$ bit is the region containing only north-east sloping lines.

Given that we are probing the influences between three time series, it is natural now to consider the behavior of the causation entropy. In this case, we have $\mathcal{C}_{X \to Z|(Y,Z)} = \mathcal{C}_{Y \to Z|(X,Z)} = 1$ bit, indicating that given the past behavior of $Z$ and one of $X$ or $Y$, the past of the other can predict the behavior of $Z_t$. Like before, this 1 bit of information cannot be localized to either $X$ or $Y$. Due to this, it is inaccurate to ascribe the 1 bit of information in $Z_t$ to either $X$ or $Y$ alone, and therefore the causation entropy also erroneously localizes the 1 bit of joint influence. These information quantities are displayed in the I-Diagram in Fig. 2.

[1] The I-Diagrams are naturally consistent, however, once one recognizes that the *co-information* [14], the inner-most information atom, is $\mathrm{I}[Y_t : X_{0:t} : Y_{0:t}] = -1$ bit.

## IV. DISCUSSION

We see that transfer-like entropies can both overestimate information flow (Section III A) and underestimate influence (Section III B). The primary misunderstanding of these quantities stems from a mischaracterization of the conditional mutual information. Most basically, probabilistic conditioning is not a "subtractive" operation: $\mathrm{I}[X : Y \mid Z]$ is not the information shared by $X$ and $Y$ once the influences of $Z$ have been removed. Rather, it is the information shared by $X$ and $Y$ *taking into account* $Z$. This is not a game of mere semantics. For example, it is possible that $\mathrm{I}[X : Y] < \mathrm{I}[X : Y \mid Z]$. This cannot happen if conditioning merely removed influence: conditional dependence is *additional* dependence that occurs in the presence of a third variable [15].

Another way to understand this phenomenon is through the *partial information decomposition* [16]. Within this framework, the mutual information between two random variables $X_1$ and $X_2$ (call them *inputs*) and a third random variable $Y$ (the *output*) is decomposed into four mutually exclusive components: $\mathrm{I}[(X_1, X_2) : Y] = R + U_1 + U_2 + S$. $R$ quantifies how the inputs may *redundantly* inform the output, $U_1$ and $U_2$ quantify how each may provide *unique* information, and finally $S$ quantifies how the inputs together may *synergistically* inform the output. In this decomposition, the mutual information between one input and the output is equal to what uniquely comes from that input plus what is redundantly provided by both inputs; $\mathrm{I}[X_1 : Y] = R + U_1$, for example. However, the mutual information between that input and the output conditioned on the other input is equal to what uniquely comes from that one input, plus what is synergistically provided by both inputs: $\mathrm{I}[X_1 : Y \mid X_2] = U_1 + S$. In other words, conditioning removes the redundant information, but adds the synergistic information. Here, conditional dependencies manifest themselves as synergy.

Treating $X_{0:t}$ and $Y_{0:t}$ as inputs and $Y_t$ as output, the partial information decomposition identifies the transfer entropy $\mathrm{T}_{X \to Y}$ as the sum of the unique information from $X_{0:t}$ plus the synergistic information from both $X_{0:t}$ and $Y_{0:t}$ together. It seems natural, and was previously proposed [17], to associate only this unique information with information flow. The transfer entropy, however, conflates unique information and synergistic information leading to inconsistencies, such as analyzed in Section III A.[2] If one were able to accurately measure the unique information [18] relating one input to the output, it may prove to be a viable measure of information flow.

[2] Similar conclusions follow for the causation entropy however, due to the additional variable, the analysis is more involved.

## V. CONCLUSIONS AND CONSEQUENCES

Although the examples were intentionally straightforward, the consequences appear far-reaching. Let's consider network science [19] which, over the same decade and a half period since the introduction of the transfer entropy, has developed into a vast and vibrant field, with significant successes in many application areas. Networks are composed of *nodes*, representing system observables, and *edges*, representing relationships between them. As commonly practiced, networks represent dyadic (binary) relationships between nodes[3]—article co-authorship, power transmission between substations, and the like. We speculate that much of the popularity of the transfer entropy is due to it too representing only dyadic relationships. This certainly facilitates its use in constructing and analyzing networks.

As we emphasized here, though, observables may be related by polyadic relationships that cannot naturally be represented on a network as commonly practiced. For example, all three variables in our second example are pairwise independent. A standard network representing dependence between them therefore consists of three disconnected nodes, thus failing to capture the dependence between variables and *pairs* of variables that is, in fact, present. As a start to repair this deficit, it would be more appropriate to represent such a system as a *hypergraph* [20].

Continuing this line of thought, if one believes that a network is an accurate model of a complex system, then one is implicitly assuming that polyadic relationships are either not important or do not exist. Said this way, it is clear that when modeling a complex system, one must test for this lack of polyadic relationship first. With this assumption generally unspoken, though, it is not surprising that a nonzero value of the transfer entropy leads analysts to interpret it as information flow. Within that narrow view, indeed, how else could one time series influence another if all interactions are dyadic? In other words, when a system is modeled as a network, all relationships look dyadic and so that is how one attempts to explain observed dependencies. The cost, of course, is either a

[3] Higher-order dependencies can be represented using networks, but depend on the use of additional latent variables. For example, one can represent polyadic relationships by building a new bipartite network consisting of the original nodes (type A) plus additional nodes representing polyadic relationships (type B). Here, an edge exists between a node of type A and a node of type B if that node is involved in that polyadic relationship. In any case, directly measuring information flow between nodes becomes a much more subtle issue in such augmented networks.

greatly impoverished or a spuriously embellished view of organization in the world.

Many of the preceding issues are difficult to analyze since our notions of "influence" are not sufficiently precise and, even when they are as with the use of information diagrams and measures and the partial information decomposition, there is a combinatorial explosion in possible kinds of informational relationship. Said differently, what one needs is a more explicit, even more elementary, structural view of how one process can be transformed to another. Paralleling the canonical $\epsilon$-machine minimal sufficient statistic representation of stationary processes, two of us (NB and JPC) recently introduced a minimal optimal transformation of one process into another, the $\epsilon$-transducer [21]. This provides a structural analysis for the minimal optimal predictor of one process about another. The corresponding informational analysis, hinted at above in Figs. 1 and 2 is forthcoming [22].

In short, the transfer entropy can both overestimate information flow (Section III A) and underestimate influence (Section III B). These are compounded when viewing complex systems as networks since the latter further conflates dyadic and polyadic relationships. In light of these interpretational concerns, it seems that several recent works that rely heavily on transfer-like entropies—ranging from cellular automata [23] and information thermodynamics [24] to cell regulatory networks [25] and consciousness [26]—will benefit from a close reexamination.

## ACKNOWLEDGMENTS

RGJ thanks Asher Mullokandov and Nicholas Timme for stimulating discussions. We thank Alec Boyd, Korana Burke, Jeffrey Emenheiser, Ben Johnson, John Mahoney, Pierre-André Noël, Paul Riechers, Dowman P. Varn, and Gregory Wimsatt for helpful feedback. JPC is a member of the Santa Fe Institute External Faculty. This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-13-1-0340.

[1] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948. 1

[2] A. N. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Info. Th.*, 2(4):102–108, 1956. 1

[3] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[4] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[5] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[6] D. S. Ornstein. Ergodic theory, randomness, and chaos. *Science*, 243:182, 1989. 1

[7] C. T. Lin. Structural controllability. *Automatic Control, IEEE Transactions on*, 19(3):201–208, 1974. 1

[8] R. Shaw. Strange attractors, chaotic behavior, and information flow. *Z. Naturforsh.*, 36a:80, 1981. 1

[9] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461, 2000. 1, 2

[10] D. A. Smirnov. Spurious causalities with transfer entropy. *Phys. Rev. E*, 87(4):042917, 2013. 1

[11] J. Sun and E. M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014. 1, 2

[12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 2

[13] R. W. Yeung. A new outlook on shannon's information measures. *IEEE Trans. Info. Th.*, 37(3):466–474, 1991. 2

[14] A. J. Bell. The co-information lattice. In S. Makino S. Amari, A. Cichocki and N. Murata, editors, *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*, volume ICA 2003, pages 921–926, New York, 2003. Springer. 3

[15] I. Nemenman. Information theory, multivariate dependence, and genetic network inference. *arXiv preprint q-bio/0406015*, 2004. 4

[16] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. 4

[17] P. L. Williams and R. D. Beer. Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*, 2011. 4

[18] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 4

[19] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. 4

[20] E. Estrada and J. A. Rodriguez-Velazquez. Complex networks as hypergraphs. 2005. arXiv:physics/0505137 [physics.soc-ph]. 4

[21] N. Barnett and J. P. Crutchfield. Computational mechanics of input-output processes: Structured transformations and the $\epsilon$-transducer. *J. Stat. Phys.*, 161(2):404–451, 2015. 5

[22] N. Barnett and J. P. Crutchfield. Computational mechanics of bivariate processes: Shannon information measures

and decompositions. *In preparation*, 2015. 5

[23] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E*, 77(2):026110, 2008. 5

[24] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015. 5

[25] S. I. Walker, H. Kim, and P. C. W. Davies. The informational architecture of the cell. *arXiv preprint arXiv:1507.03877*, 2015. 5

[26] U. Lee, S. Blain-Moraes, and G. A. Mashour. Assessing levels of consciousness with symbolic analysis. *Phil. Trans. Roy. Soc. London A: Math. Phys. Engin. Sci.*, 373(2034):20140117, 2015. 5