Black-boxing and cause-effect power

William Marshall¹, Larissa Albantakis¹, Giulio Tononi¹,*

¹Department of Psychiatry, Center for Sleep and Consciousness, University of Wisconsin, Madison, WI, USA
*Corresponding author: gtononi@wisc.edu

Reductionism assumes that causation in the physical world occurs at the micro level, excluding the emergence of macro-level causation. We challenge this reductionist assumption by demonstrating that, once causal power is properly defined and measured, it is possible for a macro level to “beat” the micro level. Here, this was done through a principled, well-defined measure of intrinsic cause-effect power – integrated information (Φ). Simple systems were evaluated for Φ across different spatial and temporal scales by systematically considering all possible black boxes. These are macro elements that consist of one or more micro elements over one or more micro time steps. Cause-effect power was evaluated based on the inputs and outputs of the black boxes, ignoring the internal mechanisms that support their function. We show that a macro, black-box system can have higher Φ than its micro constituents either by implementing more selective mechanisms (lower degeneracy) or by composing a system having more common inputs and outputs (higher integration). We also show that, for a given micro system, one can identify local maxima of Φ across several spatiotemporal scales. These local maxima correspond to macro levels of organization at which causal features of physical systems emerge, and provide a natural focus for scientific inquiries.

I. Introduction

Reductionist approaches in science usually assume that the optimal causal model of a physical system is at the finest possible scale. Coarser causal models are seen as convenient approximations due to limitations in measurement accuracy or computational power (Kim, 2000; Nagel and Hawkins, 1961). The reductionist view is based on the conjecture that the micro level of causal interaction is causally complete, leaving no room for additional causation at a macro level (Kim, 1993). The reductionist assumption is most obvious in fields such as particle physics (Nakamura et al., 2010), neuroscience (Markram, 2012), and nanotechnology (Bhushan and Marti, 2010), but it can also be seen in the social sciences (Imai et al., 2011), where researchers endeavor to ‘look inside the black box’.

Others have argued for the occurrence of genuine emergence at various macro levels (Fodor, 1974), such as the emergence of mind above and beyond the individual neurons (or atoms) that constitute the brain (Tononi, 2008), and for the autonomy of the special sciences such as chemistry (Scerri and McIntyre, 1997), biology (Dupré, 2009), and sociology (Elder-Vass, 2010), above and beyond the underlying physics. However, arguments in favor of emergence have either been vague, or they have emphasized the possibility that macro variables may have greater descriptive power than micro variables, rather than greater causal power (List and Menzies, 2009).

Recently, we challenged the reductionist assumption by introducing explicit measures of cause-effect power, which we used to show that true causal emergence can indeed occur (Hoel et al., 2013, 2016). We simulated simple physical systems constituted, at the micro level, of collections of logic gates. For these systems, we applied measures of effectiveness (Hoel et al., 2013) and integrated information (Hoel et al., 2016) both to the fine grained micro elements and to the same elements coarse grained (averaged) into macro elements and macro time intervals. In this way, we could demonstrate that, under certain conditions involving degeneracy (causal convergence) and/or indeterminism (causal divergence) at the micro level, a macro level constituted of coarse-grained macro elements and macro time intervals can “beat” the micro level in terms of cause-effect power.

Coarse graining micro elements and micro time intervals, in the sense of averaging over subsets of them, may increase cause-effect power when micro
elements are all roughly of the same kind and all their inputs and outputs can be treated as equivalent. In our simulated examples (Hoel et al., 2013, 2016), we coarse grained (nearly) identical elements (‘neurons’) into groups (‘neuronal groups’) and averaged over their states. However, in many cases the macro elements of interest cannot be obtained by coarse graining, because they are constituted of heterogeneous micro elements that are often compartmentalized and have highly specific functions, which would be muddled by averaging. For example, take the neuron, usually considered as the fundamental unit in much of neuroscience. Clearly, a neuron cannot be obtained by coarse graining, because it is constituted of a great diversity of specific molecules, organized in highly specific and hierarchical ways, performing highly specific functions. Indeed, it is the very specificity of the internal micro elements that makes the reductionist assumption seem inevitable in these cases: while we can treat a neuron as a black box for ease of understanding and for convenience, it would seem that its full causal power can only be captured by considering all the molecules that constitute the black box, in exquisite and specific detail (Markram, 2006).

Here we further challenge the reductionist assumption by generalizing the causal analysis employed for coarse graining (Hoel et al., 2016) to black-boxing (Tononi, 2010): we first analyze a system of heterogeneous, specific micro elements at the micro level; then we repeat the analysis at the macro level by grouping subsets of those micro elements inside black boxes (macro elements). Black boxes are characterized exclusively by their overall input-output function (Ashby and others, 1956; Bunge, 1963). Contrary to coarse-grained macro elements (Hoel et al., 2013, 2016), the heterogeneous micro elements inside the black box are hidden, not averaged. As an example of a black box, Fig. 1 shows, on the left, a simple, schematic neuron as constituted of a number of specific micro elements (synapses S, cell body C, axon hillock A that interact internally in specific ways). On the right, the neuron is treated as a single macro element, a black box, that receives inputs (spike or no spike for each input), produces a single output (spike or no spike), and conceals its micro elements inside.

To quantify the cause-effect power at the micro level and all possible black-boxed macro levels, and to assess the possibility of causal emergence (Hoel et al., 2013), we use integrated information (Φ) (Oizumi et al., 2014; Tononi, 2015). As a measure of intrinsic cause-effect power, Φ captures several features that are often overlooked (Oizumi et al., 2014): the dependence of cause-effect power on the specific state the system is in (state-dependency); the cause-effect power of the system’s parts (composition); whether the whole system is causally irreducible to its parts (integration); and what defines the system’s causal borders (exclusion). These features make Φ particularly suited for assessing the cause-effect power intrinsic to a system, independent of external observers. As demonstrated through several examples, the Φ value of systems of black-box macro elements can in fact be greater than that of their corresponding micro elements.

![Figure 1: A schematic neuron considered as a number of 'micro' elements (left), or as a black box (right). The neuron receives inputs at its synapses (S), which are passed on to the cell body (C) and then to the axon hillock (A), which outputs to other neurons.](image)

**II. Theory**

Integrated information (Φ) measures the intrinsic cause-effect power of a physical system (Albantakis and Tononi, 2015; Oizumi et al., 2014) by evaluating five requirements: the system’s capacity to make a difference to itself (intrinsically), composition, information, integration, and exclusion. Loosely defined, Φ quantifies to what extent a system’s cause-effect structure, which specifies how all the system’s parts constrain each other’s past and futures states, is integrated, that is, irreducible to subsystems (more below). The measure Φ, which was developed as part of integrated information theory (IIT) (Oizumi et al., 2014), can also be employed as a general measure of complexity that captures to what extent a system is both integrated and differentiated (Albantakis and Tononi, 2015; Albantakis et al., 2014; Edlund et al.,...
A physical system is a set of elements, for example neurons in the brain, or logic gates in a computer, such that each element has at least two internal states, inputs that can influence these states, and outputs that in turn are influenced by these states. Furthermore, it must be possible to manipulate, observe, and partition among elements, in order to evaluate their cause-effect power. The causal properties of a physical system can be characterized by randomly perturbing its elements into all possible states according to a maximum entropy distribution and observing their subsequent state transitions. Through this process, one obtains the transition probability matrix (TPM) for the physical system. During the perturbations, elements outside the physical system under consideration are held fixed; the states of these elements are considered “background conditions” (Oizumi et al., 2014). From the TPM, integrated information (Φ) can be calculated through further manipulations, observations, and partitions (see Supplementary S1).

Given the TPM of a candidate physical system, the first step is to identify all its mechanisms—the parts of the system which have cause-effect power within the system itself (intrinsically). To determine the set of mechanisms within the physical system, we test the entire power-set of system elements (composition).

To have cause-effect power, a set of elements in its current state must selectively constrain the potential past and future states of the system (information). In the following, we distinguish between mechanisms consisting of a single element (first order mechanisms) and those composed of multiple elements (high-order mechanisms), which play an important role in integrating the whole system. Determinism and degeneracy are two causal properties of a system that directly influence, through selectivity, the cause-effect power of its mechanisms. The system’s past and future states are constrained maximally in a non-degenerate and deterministic system where the current system state must have come from one particular past state and must go to one particular future state. In a degenerate system, multiple past states could precede the current state. In a non-deterministic system, multiple future states could follow the current state. Thus, everything else being equal, a less degenerate and more deterministic system typically has mechanisms with higher cause-effect power.

The cause-effect power of a mechanism within the system is then quantified by its Φ value (Albantakis and Tononi, 2015; Albantakis et al., 2014; Marshall et al., 2016; Oizumi et al., 2014; Tononi, 2015), which reflects its irreducibility: the degree to which the mechanism, in its current state, constrains the past and future states of the system above and beyond the mechanism’s parts (integration). The intrinsic cause-effect power of the whole system is quantified by its integrated information Φ (Albantakis and Tononi, 2015; Albantakis et al., 2014; Marshall et al., 2016; Oizumi et al., 2014; Tononi, 2015), which captures its overall irreducibility: the degree to which its cause-effect power is reduced by partitioning the system in two, such that one part is independent from the rest of the system (with respect to either its causes or its effects). For Φ to be high, the partition must affect many mechanisms that constrain the system in a selective, irreducible manner. Φ is an observer-independent quantity, as it only takes into account the differences that a system makes to itself. If Φ = 0, then there is at least one part of the system that remains unconstrained by the mechanisms of the rest: from the intrinsic perspective, then, there is no unified system, even though an external observer can treat it as one.

Finally, from the intrinsic perspective, a set of elements can only contribute to a single system, the one that is maximally irreducible (Φmax), even though many subsets and supersets of elements may have Φ > 0 (exclusion). This causal exclusion principle ensures that any element contributes its cause-effect power only once, in such a way that the overall cause-effect power of the system is maximal. Finding the set of elements that maximizes Φ defines the borders of the system from its own intrinsic perspective.

The causal exclusion principle also applies across spatiotemporal levels (Kim, 2000): a micro element (or micro interval) can only contribute its cause-effect power as such, or as a constituent of a macro element (or interval), but not multiple times (at multiple spatiotemporal scales). As with the system’s borders, it is the intrinsic cause-effect structure of the system that defines its optimal spatiotemporal scale, which is the one that maximizes Φmax (the scale at which the system makes the greatest difference for itself). To evaluate cause-effect power at macro
scales, micro elements can be grouped either by coarse-graining as in Hoel et al., (2013, 2016) or, more generally, by black-boxing, as will be demonstrated in this study.

**Black-boxing**

In typical usage, a black box is an object into which inputs impinge and from which outputs emerge, but its internal workings are not available for inspection (Ashby and others, 1956; Bunge, 1963). For our purposes, a ‘black-box element’ is a physical macro element that can be manipulated, observed, and partitioned, which is constituted of several micro elements (spatial), operating over several micro time steps (temporal). To qualify as a black box, it must satisfy the following conditions:

(i) It must have at least one input, one output, and two or more states that can be read from its output (element)

(ii) The internal micro elements are hidden (black box)

(iii) The internal micro elements contribute causally to the output (integration)

(iv) There cannot be any overlap between internal micro elements of multiple black boxes (exclusion)

Specifically:

(i) The inputs and outputs of a black box are defined in terms of the internal micro elements which receive direct input from other elements/black boxes (e.g. synapses S in Fig. 1) and directly output to other elements/black boxes (e.g. the axon hillock A in Fig. 1). For this work we allow for inputs to arrive at multiple micro elements, but restrict outputs to leave from only a single micro element within the black box. Furthermore, the inputs are taken to arrive at the beginning of the macro time step, while the outputs are taken to depart at the end of the macro time step. In principle, this framework could be extended to multiple output elements, and to a more general treatment of time.

(ii) The state of a black-box element is equivalent to the state of its output elements at the end of its macro time step. In line with the spirit of black boxes, the micro elements within the black box are “hidden” from other black boxes, meaning they can only influence the state of other elements via their black box’ output. Any inter-box interactions at the micro level are considered external to the system and are therefore ‘frozen’ as background conditions (Oizumi et al., 2014); see also Supplementary S2). The transition probabilities associated with a black-box element are determined as usual by causal analysis, perturbing the inputs of the black box into all possible states according to a maximum entropy distribution. At the end of the macro time step, the state of the black box is observed from its output elements. In this way, one can determine the cause-effect power that the inputs (i.e., outputs from other black-box elements) have on the state of the black-box element. Crucially, for the duration of the macro time step, the internal elements are allowed to evolve naturally. A black-box element also has an internal micro state, corresponding to the state of all its internal elements at the beginning of its macro time step. The internal micro state of the black box may or may not have an effect on the relationship between input and output (see Supplementary S2).

(iii) The requirement that every constituent micro element contributes causally to the output of its black box is a consequence of the integration postulate of IIT, which mandates that cause-effect power be irreducible. Only if a black-box element is irreducible is it meaningful to consider it as a single physical element, as opposed to two or more unrelated elements. This requirement is satisfied implicitly when assessing models using integrated information; any physical system that violates it will be found to be reducible and thus have Φ = 0.

(iv) The requirement for no overlap among the constituents of different black boxes (or equivalently that a micro element cannot be a constituent of more than one black-box element) is a consequence of the exclusion postulate of IIT. In terms of cause-effect power, this means that the cause-effect power of a micro-element cannot be counted multiple times (Oizumi et al., 2014). The importance of the exclusion condition has been independently recognized in the theory of computation; it is only meaningful to say that a physical system implements a computation if the system is constituted of distinct, non-overlapping elements (Chalmers, 1996). If black-box elements were permitted to
overlap, then every open physical system could be said to implement any computation (Chalmers, 1996; Putnam and Putnam, 1988).

Together, the above requirements allow to specify inputs and outputs of each black-box element, to define its macro state and its background conditions, to include within each black box only micro elements that are integrated, and to draw ‘borders’ around each black-box element that exclude any overlap with other black boxes.

In recent work, we provided a proof of principle of causal emergence by showing that systems with coarse-grained macro elements can have greater integrated information than the corresponding systems considered at the micro scale (Hoel et al., 2016). In the following, we take a similar approach and demonstrate black-boxing and its importance for causal emergence based on a set of simple proof-of-principle examples. For the purposes of this work, we shall consider collections of elements that are binary micro elements which cannot be further reduced or split, and the time scale of state transitions to be a micro time step. Time is implicit in the TPM, as micro elements are synchronously updated at discrete micro time steps. In principle, integrated information is defined for any discrete system of elements.

### III. Results

We will present several examples demonstrating how systems of black-box macro element can have greater cause-effect power than their corresponding micro systems. The full mathematical details of the $\Phi$ calculation are detailed elsewhere (Hoel et al., 2016; Marshall et al., 2016; Oizumi et al., 2014; Tononi, 2015); full example analyses are presented in Supplementary S1. All calculations in this work were performed using the PyPhi software package in Python (Mayner et al., 2016). A sample script for black-box analysis is included in Supplementary S3.

#### Degeneracy

A first worked-out example of black-boxing is shown in Fig. 2. Six micro elements – four COPY gates and two AND gates are black-boxed into two macro elements, which correspond to macro COPY gates. As illustrated by the TPM in Fig. 2 (left, bottom), the micro level has a high degree of degeneracy – the convergence of multiple past system states onto the same current system state (i.e. multiple rows leading to the same column in the TPM). Previous work had already demonstrated that reducing degeneracy is one way that a coarse-grained macro level can achieve higher cause-effect power than its corresponding micro level (Hoel et al., 2013, 2016). The example in Fig. 2 shows that black-boxing micro elements into macro elements can also be exploited to counteract degeneracy at the micro level.

In the example, two COPY micro elements input to a single AND micro element, whose current state is OFF. This implies degeneracy in the system since three states of the COPY elements (OFF, OFF), (ON, OFF) and (OFF, ON) all lead to the same state of the AND element (OFF). The micro system in state ‘all OFF’ specifies six first order mechanisms; the AND elements specify mechanisms with $\varphi = 0.167$, and the COPY elements specify mechanisms with $\varphi = 0.25$. There are no high-order mechanisms. The integrated information for the micro system is $\Phi = 0.215$.

We then consider the system at the macro level, after black-boxing each AND micro element with its two inputting COPY micro elements. Each black-box macro element turns out to implement, over two time steps, a macro COPY logic, specifying a first-order mechanism with $\varphi = 0.5$. Again, there are no high-order mechanisms. The macro system has no degeneracy, since no two past states converge onto the same current state (see macro TPM in Fig. 2). For this reason, the integrated information of a COPY element in the macro system is higher ($\varphi = 0.5$) than in the micro system ($\varphi = 0.25$), and so is the overall integrated information for the black-box macro system ($\Phi_{\text{max}} = 0.639$). Thus, appropriately black-boxing micro into macro elements reduces degeneracy and in doing so increases the cause-effect power of the system, as measured by integrated information $\Phi$. 
High-order mechanisms

An intuitive example in which black-boxing may be appropriate is propagation delay—the amount of time between the output of one element and its effect on another element. However, such delays are largely ignored in causal analysis and are taken to be an implicit aspect of the element of interest, i.e., they are black-boxed. In this example, propagation delay is modeled as (one or more) COPY elements that take a single input and then output the same value. Fig. 3 shows the micro structure of an XOR element with a one-step propagation delay, along with the corresponding macro element, a black box with XOR logic.

Next we consider a system of nine micro elements—six COPY and three XOR, which are black-boxed into a macro system of three interconnected XOR elements with a one-step propagation delay (see Fig. 4). The current state of all elements is OFF (0). Note that in this example, black-boxing the system does not decrease the degeneracy of the physical system, as both levels have an equal number of degenerate XOR elements, and the copy elements are non-degenerate. Any increase in cause-effect power through black-boxing would thus have to occur by other means.

Figure 2: Left: Two AND elements that each receive inputs from two COPY elements and send output to two other COPY elements, along with the TPM calculated from systematic perturbation of the elements. The current state of all elements is OFF. The integrated information of the micro system is $\Phi = 0.215$. Right: Black-box elements consisting of AND elements as outputs, and the corresponding COPY elements hidden within. The TPM for this system is found by perturbing the inputs to the black-box element in all possible ways, it is determined that it implements COPY logic over two time steps. The black-box macro system has $\Phi = 0.639$.

Figure 3: An XOR logic gate with one time-step propagation delay. Left: Two COPY elements which each take a single input and relay the values to an XOR element. Right: A black-box element with two inputs and a single output element (dashed red outline). By perturbing the inputs to the black-box element in all possible ways, it is determined that it implements XOR logic over two time steps.
Assessing the cause-effect structure of the micro system, we find that there are only three first-order mechanisms and no high-order mechanisms. The three XOR elements each specify a mechanism with \( \varphi = 0.5 \), all others do not have irreducible cause-effect power so \( \varphi = 0 \) (see Fig. 5). This is because from the intrinsic perspective, a set of elements must have both irreducible causes and effects within the system to be a mechanism for the system (see Theory). The XORs have both causes and effects, since by being in the OFF state, each XOR specifies that its two inputs must have been (OFF, OFF) or (ON, ON) and that its outputs, the COPY elements, must be OFF in the future. However, the six COPY elements, taken individually, lack any effect within the system: by being in the OFF state, the COPY does not constrain the future state of its XOR output, which is still equally likely to be ON or OFF depending on the state of its other input. Two COPY elements in the state (OFF, OFF) that input to the same XOR element do have an irreducible effect within the system, since together they specify that the XOR element they output to will be OFF in the future state. However, the two COPY elements do not have an irreducible cause: in the OFF state, the two COPY elements taken individually specify that their inputs must have been OFF, so the higher-order cause is reducible (Fig. 5, bottom). The lack of either irreducible causes or effects thus prevents the COPY elements from specifying first or high-order mechanisms in the system. The integrated information of the micro physical system is \( \Phi = 0.25 \).

The macro-level physical system with black-box elements also has three mechanisms with \( \varphi = 0.5 \), but they are second-order mechanisms specified by pairs of XOR elements. By being in the state (OFF, OFF), each pair of XOR elements specifies that the past state of the entire model must have been (OFF, OFF, OFF), and that the future state of their common output must be OFF. Neither of the XOR elements in this high-order mechanism can specify these constraints on its own. At the macro level, the collection of mechanisms (cause-effect structure) is more integrated than that of the micro level, with a value of \( \Phi^{\text{max}} = 1.875 \). Consequently, there is cause-effect power that emerges at this macro level of the physical system. Concealing the COPY elements inside the black boxes reveals the high-order interactions between the XOR gates over two time steps. This result agrees with the intuition that all the “real work” is being done by interactions between the XOR elements, and that the COPY elements only implement the propagation delay. Indeed, the integrated information analysis confirms that this is the case not only from our intuitive extrinsic perspective but also from the intrinsic perspective of the system itself.

*Figure 4:* A system of three interconnected XOR elements with one-step propagation delay and all elements in the OFF state. At the micro level, the propagation delay is modeled by COPY elements between the XOR elements. At the micro level the model has only first order mechanisms, and \( \Phi = 0.25 \). When the elements are black-boxed over two time steps, the system has second order mechanisms and \( \Phi = 1.875 \).
Lack of determinism (the effects of noise)

We further explore this example by considering a scenario in which propagation delays occur over noisy channels. Adding noise to the system reduces determinism, another factor that has been previously shown to influence whether the “macro beats the micro” (Hoel et al., 2013, 2016). In this case, the COPY elements take a single input and then output the same value with probability $p$ in $[0.5, 1]$. The original results of Figure 3 correspond to a noiseless channel ($p = 1$), while a completely noisy channel ($p = 0.5$) would have no cause-effect power. The integrated information analysis is performed on both micro and black-box physical system for several different values of $p$ (see Table 1). The number and orders of mechanisms is the same for all values of $p > 0.5$, but the $\Phi$ value of each mechanism decreases as $p$ decreases, and it does so more steeply for the black-box system than for the micro system. Along with $\Phi$ the overall integrated information $\Phi$ of both the black-box and micro systems decrease as the amount of noise increases. Since also $\Phi$ of the black box system declines faster with increasing noise than $\Phi$ of the micro system, the macro level has $\Phi_{\text{max}}$ for $p > 0.6$, whereas the micro level has $\Phi_{\text{max}}$ when $p \leq 0.6$.

In summary, the first example (Fig. 2) demonstrates that, by decreasing degeneracy, a black-boxed system can have more cause-effect power than its corresponding micro system. The second example (Fig. 4) shows that a black-boxed system can beat the micro level independent of degeneracy. Even if the system has the same number of first order mechanisms and the same $\Phi$ values at both the micro and the macro level, the black-boxed system “wins” by having high-order mechanisms. The high-order mechanisms of the black-box system have overlapping constraints, with each mechanism constraining all elements within the system, whereas the first-order mechanisms of the micro system only constrain their respective COPY inputs and outputs, without overlap. A system partition at the micro level only affects a single mechanism, whereas a system partition at the black-box level affects all of the mechanisms in the system, resulting in higher integration. The third example (Table 1) illustrates how indeterminism (noise) can disproportionately affect high-order mechanisms, and that the reduction in cause-effect power due to indeterminism can outweigh the increase due to high-order mechanisms. Overall, several interrelated aspects of cause-effect power have to be considered when assessing if the “macro beats the micro,” including degeneracy, indeterminism and the presence of high-order mechanisms. While the roles of degeneracy and indeterminism were already apparent in the context of coarse-graining, the emergence of high-order mechanisms at macro levels becomes apparent when analyzing black-box elements.
Table 1: Integrated information for various noise levels (p) in the propagation delay network (Fig. 4)

| p   | 1.0  | 0.9  | 0.8  | 0.7  | 0.6  | 0.5  |
|-----|------|------|------|------|------|------|
| Micro Φ | 0.250 | 0.160 | 0.090 | 0.040 | 0.010 | 0.000 |
| Micro φ  | 0.5   | 0.4   | 0.3   | 0.2   | 0.1   | 0    |
| Black box Φ | 1.875 | 1.046 | 0.363 | 0.070 | 0.004 | 0.000 |
| Black box φ | 0.500 | 0.320 | 0.180 | 0.080 | 0.020 | 0.000 |

Finding local and global maxima of cause-effect power

A final example considers a system of micro elements across all possible black-boxings, to establish the global maximum of causal emergence. The micro system is constituted of 55 elements implementing NOR logic; this logic was selected because it is linear and it can be combined to form many other types of logic. Fig. 6, demonstrates how a group of 11 elements implementing NOR logic can be connected in such a way to produce AND/OR logic, or MAJORITY logic at coarser spatiotemporal scales.

For this example, the current state of each of the 55 NOR micro elements is ON. The system is arranged into five interconnected sets of 11 elements, with each set organized according to Fig. 7. Each set of 11 elements receives inputs from three other sets, and outputs to three other sets. We can then describe this example as a micro physical system of NOR elements, a black-boxed system of AND/OR elements, a black-boxed system of MAJORITY elements, or various combinations thereof, as shown in Fig. 6. Many other possible black-boxing schemes were also evaluated but are not shown in the figure.

At the micro level, the system has 55 first order mechanisms, one for each micro elements, and no high-order mechanisms, with φ = 0.239 on average. The micro system is completely non-degenerate—a NOR element in the ON state has only one possible past state (all OFF). The integrated information of this micro physical system is Φ = 0.453.

In Fig. 7, a macro-level black-boxed system with an average spatial grain of 2.75 has 20 macro elements, 15 implementing AND logic and 5 implementing OR logic, operating at a time scale of two steps.

Figure 6: Left: A collection of 11 NOR elements at the micro scale. Middle: A macro scale black-boxing of these elements into four black boxes. Perturbational analysis reveals that three of these elements implement AND logic and the final one OR logic, at a time scale of two time steps. Right: A macro scale black-boxing with one element, implementing MAJORITY logic over its three inputs at a time scale of four time steps. Note that there are only three inputs, but each input arrives at two different micro elements.
Similar to the micro level, there are 20 first order mechanisms (one for each black-box element) but no high-order mechanisms, with $\varphi = 0.112$ on average. As expected based on the results of the first two examples, the integrated information for this black-boxed system is lower ($\Phi = 0.080$) than that of the corresponding micro system. This is because cause-effect power at this level is reduced by an increase in degeneracy (OR element in ON state is highly degenerate; $\varphi$ values are 0.127 lower on average) compared to the micro level in the absence of any counteracting factors, since both levels are equally deterministic and have no high-order mechanisms.

![Image of systems](image)

**Figure 7:** A system of interconnected elements viewed at several different grain sizes, with all elements in the ON state. The systems are arranged with micro elements on the far left, and black-box elements of increasing spatial grain to the right. The legend on the right specifies the input-output function of each element. Each of the three systems on the top row is a local maximum of cause-effect power, corresponding to the three sets of elements in Fig. 6, and the system on the far right is the global maximum. In the bottom row are two examples of the many systems with $\Phi = 0$ at spatial grains between the local maxima.
A macro-level black-boxed system with an average spatial grain of 11 is defined by considering black-box elements implementing MAJORITY logic over four time steps, as in Fig. 7. Compared to the macro level with an average spatial grain of 2.75, this additional black-boxing step further reduces the number of elements, which are also more degenerate (Φ values are 0.023 lower on average). However, this macro system is endowed not only with first-order mechanisms, but with all possible second, third and fourth-order mechanisms. In total, there are 30 of a possible 31 mechanisms from the power set of black-box elements, resulting in high integration, with $\Phi^{\text{max}} = 2.333$. Further analysis indicates that this particular black-box system has the largest Φ value of all possible systems that can be defined from this collection of micro elements. Note that, while this example illustrates a particular state of the system, and cause-effect structures are state dependent, the black-boxing scheme just presented has maximum cause-effect power across many states (not shown).

Fig. 7 also shows additional black-box systems with $\Phi = 0$. One of these black-box system with an average spatial grain of 1.57 has 20 black-box OR elements over two time steps and 15 micro OR elements. A second black-box system with average spatial grain of 3.66 has 10 black-box AND elements over two time steps and 5 black-box AND elements over four time steps. For both of these systems (and many others not shown), the integrated information is $\Phi = 0$, because there is no common time scale over which all the elements in the system have effects on other elements within the system. For any specific time scale, there will be elements that do not causally contribute, and thus the system is not integrated.

In summary, this example demonstrates how to evaluate cause-effect power over many different spatial and temporal scales, to find the global maximum, $\Phi^{\text{max}}$. For this example, the global maximum occurs at a spatial grain of 11 micro elements per macro element, where the elements implement majority logic over 4 time steps. This example also highlights the existence of local maxima of cause-effect power, specifically at the micro level (average spatial grain size of 1, $\Phi = 0.453$) and at a macro level (average spatial grain size of 2.75, $\Phi = 0.080$). While these spatial grains do not specify a global maximum, they nevertheless reveal levels of organization at which the system exhibits non-zero values of cause-effect power, which may shed light on its causal properties when evaluated from the extrinsic perspective. The vast majority of systems of black-box elements, on the other hand, yield $\Phi = 0$.

IV. Discussion

In this work we explore the cause-effect power of simple systems of elements considered both at the micro level and after black-boxing at a macro level. The cause-effect power of these systems was assessed using integrated information (Φ), a measure of the cause-effect power that is intrinsic to a physical system. We show how macro systems based on black boxes can causally emerge, by having greater cause-effect power than their micro element counterparts. This result complements and extends previous work that showed how causal emergence can occur when macro elements are defined by coarse graining micro elements (Hoel et al., 2016).

Integrated Information as intrinsic cause-effect power

Integrated information Φ is a measure of the cause-effect power that is intrinsic to a physical system, that is, how the system constrains its potential past and future states. To properly capture cause-effect power from the intrinsic perspective, Φ also considers composition, specificity, irreducibility, and exclusion (Albantakis and Tononi, 2015; Oizumi et al., 2014).

To fully capture the intrinsic cause-effect power of a system, it is necessary to consider its parts, that is, the compositional structure of its elements. The cause-effect power of the parts of a system is assessed by asking whether and how a subset of elements makes a difference to the system itself. As demonstrated above (Fig. 5), it is possible that a composition of multiple elements will have joint, irreducible causes and effects within a system (high-order mechanisms). Integrated information captures these composite causes and effects by considering the entire power set of elements as potential mechanisms within the system.

Cause-effect power is specific – it depends on the state of the physical system. In IIT, causal specificity is made explicit in informational terms: a mechanism in a state only has cause-effect power if it constrains the repertoire of potential past and
future states of the system in a specific way. This implies that the repertoire of potential past and future states is different for different states of the mechanism – a notion that underlies counterfactual accounts of causality (Lewis, 1974; Pearl, 2009).

Intrinsic cause-effect power is only meaningful for a set of elements that constitutes a single system, not an aggregate of many systems. If a set of elements can be partitioned into two parts with no loss of cause-effect power then there is no intrinsic perspective for the physical system as a whole, but rather, there are several unrelated subsystems each with its own intrinsic perspective. By partitioning the elements of a system in all possible ways, $\Phi$ measures the integration of a physical system, ensuring that the cause-effect power is intrinsic to a unitary system.

Finally, intrinsic cause-effect power is definite; there must be “causal borders” that define the extent and grain size of the system and exclude that overlapping systems and spatiotemporal grains share cause-effect power. Causal exclusion is a mainstay of reductionist accounts of causation, which assume that all causal power resides with micro elements and time steps, excluding all macro levels (Kim, 2000). The exclusion postulate of integrated information theory states that only maxima of cause-effect power count ($\Phi_{\text{max}}$), thereby excluding overlapping sets and spatio-temporal grains. However, the exclusion postulate leaves open the possibility that such maxima may occur at a macro level, as demonstrated here.

**Black-boxing reveals high-order causes and effects at macro spatiotemporal scales**

The two main requirements for high $\Phi$ are that a physical system is differentiated (many specific mechanisms) and integrated (mechanisms with overlapping constraints). Typically, whenever a micro system is mapped into a macro system, there is reduced state differentiation, i.e., the macro system has fewer elements and a smaller state space. This decrease in differentiation means less potential mechanisms that the system can specify, and thus less potential integrated information (Marshall et al., 2016). Thus in order for the macro to beat the micro, the macro must increase cause-effect power, either by having more specific mechanisms, or a more integrated set of mechanisms.

One factor that influences the specificity of a mechanism is degeneracy (causal convergence). The example in Fig. 2 demonstrates how the macro can beat the micro when black-boxing reduces the degeneracy of the system. Decreasing degeneracy leads to an increase in the cause-effect power of mechanisms within the system. If this increase in cause-effect power can overcome the inherent loss of differentiation in macro systems, then the macro can beat the micro. This result was previously demonstrated for coarse-grained macro elements (Hoel et al., 2013, 2016), and we confirm that it also holds for black-box macro elements.

Moreover, as shown in Fig. 4, the macro can beat the micro through increased integration – by revealing high-order causes and effects at macro spatiotemporal scales. This occurs when elements with few effects are concealed within black-box elements, and micro elements with many effects serve as the outputs of black-box elements, resulting in a more densely interconnected set of macro elements, where groups of macro elements share common inputs and common outputs. If creating common inputs and common outputs among elements leads to additional, joint constraints of the possible past and future system states, elements may form high-order mechanisms, resulting in a more integrated cause-effect structure and higher $\Phi$. Being a part of high-order mechanisms, or being constrained by multiple mechanisms, gives an element additional ways to contribute to the cause-effect structure; when an element contributes in multiple ways, cutting that element has a greater effect on the cause-effect structure, making the system more irreducible.

As shown in Fig. 8, black-boxing is most beneficial when there are “causal bottlenecks” in the micro system, that is, when a micro element with a single output connects to a micro element with a single input. In this case, it is impossible for these micro elements to contribute to high-order mechanisms, and they represent a “weak link” in the integration of the system. When the connection between these elements is cut, it will only affect two first-order mechanisms. By black-boxing the system, these elements can become part of a black-box element with several inputs, which outputs to many other elements. Now the black box that contains these micro elements can contribute to high-order mechanisms, and there will be a much greater effect of cutting the connection between them. This factor, combined with decreased degeneracy or higher
determinism, can lead to an overall increase in $\Phi$ compared to the micro level.

**Black-boxing vs. coarse-graining**

Black-boxing is not the only method that allows grouping micro elements into causally emerging macro elements. As demonstrated by Hoel et al., causal power can also be gained when micro elements are coarse-grained (Hoel et al., 2013, 2016). Both black-boxing and coarse-graining necessarily lead to macro systems with a reduced state space. In black-boxing, the macro states of a macro element are determined by the states of its output elements, without reference to the states of its internal, functionally heterogeneous micro elements. In coarse-graining, the macro states of a macro element are averages of the states of its constituting, functionally homogeneous micro elements, without reference to the identity of individual micro elements. In both cases, increases in the system’s selectivity (increased determinism or decreased degeneracy) at the macro level can counteract its inherent disadvantage of fewer potential mechanisms with fewer potential constraints. As described above, black-boxing reveals the causal emergence of high-order macro mechanisms that are not present at a micro scale, which can lead to higher $\Phi$ at the macro level. The main factor underlying causal emergence in coarse-grained systems is a gain in the selectivity of the system’s mechanisms when noisy or degenerate micro elements are grouped into more deterministic macro elements with less convergent inputs and outputs. This is because, everything else being equal, more selective mechanisms make a system more irreducible (Albantakis and Tononi, 2015; Hoel et al., 2016). Ultimately, black-boxing can be considered as the general case, where maximally convergent black boxes (one output for all micro elements of a box) and coarse-grains (an output for each micro element) are the extremes in a continuum of possible macro elements (Fig. 8).

**Local maxima of cause-effect power**

Evaluating cause-effect power of black-boxed systems across many spatiotemporal scales shows that, in general, there can be several local maxima of cause-effect power, between which integrated information decreases or falls to zero (Fig. 7). Importantly, even within a given spatiotemporal grain, there will generally be several local maxima corresponding to overlapping subsets of elements, such that adding or subtracting an element reduces integrated information (Hoel et al., 2016; Oizumi et al., 2014). These local maxima of cause-effect power correspond to organizational macro levels and systems having emergent causal properties. These are natural levels and systems for the special sciences to investigate.

Two prime examples are biological and electronic systems, since they contain many highly specialized components required to perform their function. In biology we can study the molecules within an individual cell, the interactions between networks of cells (nervous system), individual organs (liver, kidneys), whole organism (animals, humans), and communities of organisms (swarms, societies). In electronics we can study individual transistors inside of a computer, circuits of logic gates, larger computer components (processor, memory), a computer as a whole, as well as computer networks. Note that the typical approach of studying biological or electronic systems at a particular (macro) spatiotemporal level is precisely to treat its next-lower level components as black boxes. Here we have proposed a rigorous theoretical framework to evaluate cause-effect power and the causal properties of such a black-box system. If an organizational level corresponds to a local maximum of integrated information, then there will be causal properties that emerge at that level, and that there is knowledge to be gained by studying the system accordingly. Finally, while local maxima reveal causal properties to an investigator studying the
system, the global maximum specifies the set of elements and spatiotemporal grain at which the system has most cause-effect power upon itself – from its own intrinsic perspective. According to integrated information theory (Oizumi et al., 2014), the set of elements at the spatial-temporal grain that defines the global maximum of cause-effect power corresponds to a physical substrate of consciousness.

Acknowledgements

This work has been supported by the Templeton World Charities Foundation (Grant #TWCF 0067/AB41).

References

Albantakis, L., and Tononi, G. (2015). The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. Entropy 17, 5472–5502.

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. (2014). Evolution of integrated causal structures in animals exposed to environments of increasing complexity. PLoS Comput Biol 10, e1003966.

Ashby, W.R., and others (1956). An introduction to cybernetics. Introd. Cybern.

Bhushan, B., and Marti, O. (2010). Scanning probe microscopy—principle of operation, instrumentation, and probes. In Springer Handbook of Nanotechnology, (Springer), pp. 573–617.

Bunge, M. (1963). A general black box theory. Philos. Sci. 346–358.

Chalmers, D.J. (1996). Does a rock implement every finite-state automaton? Synthese 108, 309–333.

Dupré, J. (2009). It is not possible to reduce biological explanations to explanations in chemistry and/or physics. Contemp. Debates Philos. Biol. 32–47.

Edlund, J.A., Chaumont, N., Hintze, A., Koch, C., Tononi, G., and Adami, C. (2011). Integrated information increases with fitness in the evolution of animats. PLoS Comput Biol 7, e1002236.

Elder-Vass, D. (2010). The causal power of social structures: Emergence, structure and agency (Cambridge University Press).

Fodor, J.A. (1974). Special sciences (or: the disunity of science as a working hypothesis). Synthese 28, 97–115.

Hoel, E., Albantakis, L., Marshall, W., and Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. J. Conscious. Sci.

Hoel, E.P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. Proc. Natl. Acad. Sci. 110, 19790–19795.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. Am. Polit. Sci. Rev. 105, 765–789.

Kim, J. (1993). Supervenience and mind: Selected philosophical essays (Cambridge University Press).

Kim, J. (2000). Mind in a physical world: An essay on the mind-body problem and mental causation (MIT press).

Lewis, D. (1974). Causation. J. Philos. 70, 556–567.

List, C., and Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. J. Philos. 106, 475–502.

Markram, H. (2006). The blue brain project. Nat. Rev. Neurosci. 7, 153–160.

Markram, H. (2012). The human brain project. Sci. Am. 306, 50–55.
Marshall, W., Gomez-Ramirez, J., and Tononi, G. (2016). Integrated Information and State Differentiation. Front. Psychol. 7.

Mayner, W., Marshall, W., and Marchman, B. (2016). pyphi: 0.8.1.

Nagel, E., and Hawkins, D. (1961). The structure of science. Am. J. Phys. 29, 716–716.

Nakamura, K., Group, P.D., and others (2010). Review of particle physics. J. Phys. G Nucl. Part. Phys. 37, 75021.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput Biol 10, e1003588.

Pearl, J. (2009). Causality (Cambridge university press).

Putnam, H., and Putnam, H. (1988). Representation and reality (Cambridge Univ Press).

Scerri, E.R., and McIntyre, L. (1997). The case for the philosophy of chemistry. Synthese 111, 213–232.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. Biol. Bull. 215, 216–242.

Tononi, G. (2010). Information integration: its relevance to brain function and consciousness. Arch. Ital. Biol. 148, 299–322.

Tononi, G. (2015). Integrated information theory. Scholarpedia 10, 4164.
Black-boxing and cause-effect power: 
Supplementary Information

William Marshall\textsuperscript{1}, Larissa Albantakis\textsuperscript{1}, Giulio Tononi\textsuperscript{1,*}

\textsuperscript{1}Department of Psychiatry, Center for Sleep and Consciousness, University of Wisconsin, Madison, WI, USA
\textsuperscript{*}Corresponding author: gtononi@wisc.edu

Supplementary S.1 – Full analysis of cause-effect power

Here we present the in-depth results of the integrated information analysis of the example systems from the main text. All calculations were performed using the PyPhi software package in Python (Mayner et al., 2016).

For a given system in a specific state, this process involves identifying its cause-effect structure, the set of mechanisms in the system. Each member of the power set of system elements is tested as a potential mechanism. For a micro-level system this includes only compositions of micro elements, and cannot include the macro causes and effects of coarse-grained or black-boxed macro elements. Conversely, if it is a macro-level system, only compositions of macro elements are tested and the causes and effects of individual micro elements are ignored. If a set of elements has irreducible cause and effect power within the system, then it specifies a mechanism. The complete specification of a mechanism includes its cause and effect purviews (the elements over which it has maximally irreducible power to constrain the past and future states), the cause-effect repertoires that specify those constraints, and its integrated information value ($\phi$).

Next, all possible directed partitions of the candidate set of micro elements are considered, to find the one that least affects the cause-effect structure. After each potential partition, the cause-effect structure is recalculated, and the result is compared to the cause-effect structure of the whole system. The partition that makes the least difference to the cause-effect structure is the minimum information partition (MIP), and the difference it makes, as measured using an extended earth movers distance (Oizumi et al., 2014), defines the integrated information of the system ($\Phi$).

Example 1 – Degeneracy

To describe the cause-effect structure, we first assign labels to each of the micro elements in the system, as shown in Fig. S1.

| Unpartitioned | Mechanism | Past Purview | Future Purview | $\phi$ |
|---------------|-----------|--------------|----------------|-------|
|               | (0)       | (5)          | (2)            | 0.25  |
|               | (1)       | (5)          | (2)            | 0.25  |
|               | (2)       | (0, 1)       | (3, 4)         | 0.167 |
|               | (3)       | (2)          | (5)            | 0.25  |
|               | (4)       | (2)          | (5)            | 0.25  |
|               | (5)       | (3, 4)       | (0, 1)         | 0.167 |
For this system, the MIP is to cut all connections from (1) to (0, 2, 3, 4, 5). Note that there are other equivalent MIPs, we focus on this specific cut without loss of generality. Under the partition, element 1 no longer has an effect and thus does not specify a mechanism, and the mechanism specified by element 2 is altered, its past purview is reduced from (0, 1) to only (0). Comparing the whole and partitioned cause-effect structure, we find that the resulting integrated information value for the micro system is $\Phi = 0.215$.

A search over all subsets of micro elements reveals that in the current state (all OFF), all other potential candidate systems at the micro level have $\Phi = 0$. We then consider all possible black-box systems at various macro scales. The macro system with highest $\Phi$ has two black-box elements over two time steps. The black-box elements are constituted of micro elements $A = (0, 1, 2)$ and $B = (3, 4, 5)$, with corresponding output elements 2 and 5. Performing a causal analysis, we find that these black-box elements implement COPY logic over their input. The cause-effect structure of the black-box system is

| Mechanism | Past Purview | Future Purview | $\Phi$ |
|-----------|--------------|----------------|--------|
| 0         | (5)          | (2)            | 0.25   |
| 1         | (5)          | ()             | 0      |
| 2         | (0)          | (3, 4)         | 0.167  |
| 3         | (2)          | (5)            | 0.25   |
| 4         | (2)          | (5)            | 0.25   |
| 5         | (3, 4)       | (0, 1)         | 0.167  |

The MIP for this system is to cut all connections from (0) to (1, 2, 3, 4, 5), that is, cut the outputs of micro element 0 (by symmetry, cutting the outputs of elements 1, 3 and 4 would be equivalent). After the partition, both mechanisms have had their irreducible cause-effect power diminished, and the result is $\Phi = 0.639$.

| Mechanism | Past Purview | Future Purview | $\Phi$ |
|-----------|--------------|----------------|--------|
| A         | B            | B              | 0.5    |
| B         | A            | A              | 0.5    |
| partitioned | Mechanism | Past Purview | Future Purview | $\Phi$ |
| A         | B            | B              | 0.167  |
| B         | A            | A              | 0.25   |
Example 2 – High-order mechanisms

For this example, we also assign labels to the micro elements of the system, as shown in Fig. S1. There are three mechanisms in the cause-effect structure; they are all first order mechanisms and each one corresponds to an element implementing XOR logic.

| Mechnism | Past Purview | Future Purview | $\varphi$ |
|----------|--------------|----------------|----------|
| (0)      | (2, 7)       | (1, 8)         | 0.5      |
| (3)      | (1, 5)       | (2, 4)         | 0.5      |
| (6)      | (4, 8)       | (5, 7)         | 0.5      |

For this system, the MIP is to cut all connections from (0, 1, 2, 3, 4, 5, 6, 7) to (8). Under the partition, the mechanism specified by element 0 is altered, its future purview is reduced from (1, 6) to only (1). The resulting integrated information value of the micro system is $\Phi = 0.5$.

| Mechnism | Past Purview | Future Purview | $\varphi$ |
|----------|--------------|----------------|----------|
| (0)      | (2, 7)       | (1)            | 0.5      |
| (3)      | (1, 5)       | (2, 4)         | 0.5      |
| (6)      | (4, 8)       | (5, 7)         | 0.5      |

A search of all possible subsets of micro elements is performed to find the “best” micro system (i.e. the one with maximal $\Phi$). For this system, the subsets (0, 1, 2, 3), (3, 4, 5, 6) and (0, 6, 7, 8) all have an equivalent largest value of $\Phi$, for concreteness we will discuss the system (0, 1, 2, 3). In this case, the values of the other elements are fixed in their current state as background conditions. The cause-effect structure of this subsystem has four first order mechanisms.

| Mechnism | Past Purview | Future Purview | $\varphi$ |
|----------|--------------|----------------|----------|
| (0)      | (2)          | (1)            | 0.5      |
| (1)      | (0)          | (3)            | 0.5      |
| (2)      | (3)          | (0)            | 0.5      |
| (3)      | (1)          | (2)            | 0.5      |

The MIP is to cut the connections from (0) to (1, 2, 3). After the partition, elements 0 and 1 no longer specify mechanisms (the effect of mechanism 0 and the cause of mechanism 1 have been destroyed). The integrated information of the micro system is $\Phi = 1$.

| Mechnism | Past Purview | Future Purview | $\varphi$ |
|----------|--------------|----------------|----------|
| (0)      | (2)          | ()             | 0        |
| (1)      | ()           | (3)            | 0        |
| (2)      | (3)          | (0)            | 0.5      |
| (3)      | (1)          | (2)            | 0.5      |

A search is then performed over all possible systems of black-box elements. The system with the largest value of integrated information contained three black-box elements, and operated over two time steps. The three black-box macro elements were constituted of micro elements $A = (0, 2, 7)$, $B = (1, 3, 4)$ and $C = (5, 6, 8)$ with corresponding output elements (0), (4) and (6). This black-box system had three high-order mechanisms.

| Mechnism | Past Purview | Future Purview | $\varphi$ |
|----------|--------------|----------------|----------|
| (A, B)   | (A, B, C)    | (C)            | 0.5      |
The MIP for this macro system cuts connections from \((1, 3)\) to \((0, 2, 4, 5, 6, 7, 8)\). After the partition, all of the mechanisms have been destroyed. Mechanisms \((A, B)\) and \((B, C)\) no longer have irreducible causes or effects, while the set of elements \((A, C)\) has an effect but no cause. The integrated information of this model is \(\Phi = 1.875\).

| Mechanism | Past Purview | Future Purview | \(\varphi\) |
|------------|--------------|----------------|----------|
| \((A, B)\) | ()           | ()             | 0        |
| \((A, C)\) | ()           | (B)            | 0        |
| \((B, C)\) | ()           | ()             | 0        |

### Example 3 – Local and Global Maxima

In this example, because there are so many elements we will not assign numbers to the elements. Instead, we will simply refer to each element based on the number of inputs it has, e.g., \(\text{NOR}(3, 1)\) for a \(\text{NOR}\) element with three inputs and one output. Each of the 55 elements specifies a first order mechanism, summarized in the table below.

| Multiplicity | Mechanism | Past Purview | Future Purview | \(\varphi\) |
|--------------|-----------|--------------|----------------|----------|
| 30           | \(\text{NOR}(1, 1)\) | \(\text{NOR}(1, 6)\) | \(\text{NOR}(2, 1)\) | 0.25     |
| 15           | \(\text{NOR}(2, 1)\) | \(2\times\text{NOR}(1, 1)\) | \(\text{NOR}(3, 1)\) | 0.125    |
| 5            | \(\text{NOR}(3, 1)\) | \(3\times\text{NOR}(2, 1)\) | \(\text{NOR}(1, 6)\) | 0.5      |
| 5            | \(\text{NOR}(1, 6)\) | \(\text{NOR}(3, 1)\) | \(6\times\text{NOR}(1, 1)\) | 0.25   |

The minimum information partition for this system is to cut the connections from a \(\text{NOR}(1, 1)\) element to the rest of the system. It doesn’t matter which \(\text{NOR}(1, 1)\) element, they all have the same effect on their respective future purview. The result of the MIP is that one mechanism is destroyed, and another is altered. The integrated information of this system is \(\Phi = 0.25\).

| Multiplicity | Mechanism | Past Purview | Future Purview | \(\varphi\) |
|--------------|-----------|--------------|----------------|----------|
| 29           | \(\text{NOR}(1, 1)\) | \(\text{NOR}(1, 6)\) | \(\text{NOR}(2, 1)\) | 0.25     |
| 1            | \(\text{NOR}(1, 1)\) | \(\text{NOR}(1, 6)\) | ()             | 0        |
| 14           | \(\text{NOR}(2, 1)\) | \(2\times\text{NOR}(1, 1)\) | \(\text{NOR}(3, 1)\) | 0.125    |
| 1            | \(\text{NOR}(2, 1)\) | \(\text{NOR}(1, 1)\) | \(\text{NOR}(3, 1)\) | 0.125    |
| 5            | \(\text{NOR}(3, 1)\) | \(3\times\text{NOR}(2, 1)\) | \(\text{NOR}(1, 6)\) | 0.5      |
| 5            | \(\text{NOR}(1, 6)\) | \(\text{NOR}(3, 1)\) | \(6\times\text{NOR}(1, 1)\) | 0.25    |

A search of all possible subsets of micro elements is performed to find the system with highest \(\Phi\). It was determined that if any of the micro elements is frozen as part of the background conditions, its outputs will be causally inactivated, and the system is no longer integrated, resulting in \(\Phi = 0\). Thus the micro system that contains all 55 elements is the only one that is integrated, and thus the micro model with the maximum value of \(\Phi\).

One easily interpretable option for a macro system is to define black-box elements that implement AND and OR logic (Fig. 6). This system has an average spatial grain size of 2.75. There is a symmetry in the system, so that the mechanisms specified by each OR gate are the same, and the mechanisms specified by each AND gate are the same (the OR elements output to six AND elements and take inputs from three
AND elements, the AND elements take inputs from two OR elements and output to one OR element). At this macro scale, the system has 15 black-box elements implementing AND logic and 5 black-box elements implementing OR logic, over two time steps, and each specifies a first order mechanism.

| Multiplicity | Mechanism | Past Purview | Future Purview | $\phi$ |
|--------------|-----------|--------------|----------------|-------|
| 5            | OR        | 3*AND        | 6*AND          | 0.071 |
| 15           | AND       | 2*OR         | OR             | 0.125 |

The MIP for this system is to cut the outputs of a NOR(2, 1) element that is one of the input elements of an AND black-box element. In this case, the mechanisms have the same cause-effect power, but one of the AND mechanisms and one of the OR mechanisms have different purviews (i.e. they are moved in cause-effect space).

| Multiplicity | Mechanism | Past Purview | Future Purview | $\phi$ |
|--------------|-----------|--------------|----------------|-------|
| 4            | OR        | 3*AND        | 6*AND          | 0.071 |
| 14           | AND       | 2*OR         | OR             | 0.125 |
| 1            | AND       | OR           | OR             | 0.125 |

A further search over all possible black-box elements is performed to find the global maximum of cause-effect power. This occurs for a system with five black-box elements {A, B, C, D, E} over 4 time steps. Each black box implements a MAJORITY function over its three inputs, with a specialized connectivity pattern shown in Fig. 6. Of the 31 possible mechanisms from the power set of 5 elements, 30 specify irreducible causes and effects.
The minimum information partition of this network is to cut the outputs of one of the NOR(1, 1) elements. By the symmetry in the system, there is an equivalent MIP in each of the black-box elements; however, due to the specialized connectivity structure, not all NOR(1, 1) elements are equivalent. For black-box element A, the MIP is to cut the NOR(1, 1) element that receives input from D and outputs to the NOR(2, 1) elements along with the NOR(1, 1) elements that receives input from C. As a result of the MIP, two of the mechanisms are destroyed (BCD, and ABCD) and 15 others are modified (shown in bold), the $\Phi$ value is 2.333.
Supplementary S.2 – The role of micro elements

While the micro elements that constitute a black-box macro element do not directly contribute to the cause-effect power of a physical system, in certain circumstances the current state of the micro elements will have an indirect effect on the cause-effect structure. This is not the case for any of the examples presented in the main text, so here we give some simple example to explore this phenomenon.

The role of micro elements: As background conditions

As stated in the main text, the cause-effect power of a black-box element is determined by its inputs and outputs, the micro elements within the black box are “hidden” from other black boxes, meaning they can only influence the state of other elements via their black box’ output. From the perspective of the causal analysis, any inter-box interactions at the micro level are considered external to the system and are therefore ‘frozen’ as background conditions.

Consider the network shown in Fig. S1, which is a slight modification of the degeneracy example from the main text (Fig. 2) with an additional connection from micro element (A) to micro element (B) highlighted in blue. Analyzing the same black-box elements as in Fig. 2, this additional connection between (A) and (B) is a micro-level effect between the two black boxes, and is thus frozen as a background condition. If the state of micro element (A) is ON, then once its output to (B) is frozen, micro element (B) implements COPY logic over its remaining input, completely recovering the system presented in the main text. However, if the micro state of element (A) is OFF, then micro element (B) becomes inactivated – there is no input that can affect its state and as a result Φ = 0.

Figure S2: Two examples of the state of internal micro elements indirectly affecting cause-effect power. Left: The connection from (A) to (B) is an inter black-box connection at the micro level and must be frozen as a background condition, as, in this grouping, (A) is not considered an output element of the black box. Right: The output element feeds back into the black box, and as a result the input-output function depends on the state of the XOR element.

The role of micro elements: Modulating the input-output relationship of a black box

Another way that the state of micro elements can indirectly influence cause-effect power is by modulating the input-output relationship of a black box. For all of the examples in the main text, the function of black boxes is independent of its micro state, because there is a feed forward path from input to output. However, if there is feedback within a black-box element, then the state of micro elements may affect its function. As a simple example, consider the black-box element in Fig S2, where the output element feeds back into the XOR element. The input-output function of this black-box element depends on the internal state of the micro XOR element; when the XOR element is OFF it implements COPY logic over three time steps, when the XOR element is ON, it implements NOT logic.
Supplementary S.3 – Python Script for Fig. 2 example

```python
# Black box example (Fig. 2) - Results calculated using PyPhi package in Python
# A system of two AND elements (C and F) which each take inputs from, and give
# outputs to two COPY elements. The current state is all OFF.

# Load the pyphi package - make sure it is installed and up-to-date
import pyphi

# Load the micro system from the pre-made examples.
network = pyphi.examples.blackbox_network()

# The current state of the system is all OFF
current_state = (0, 0, 0, 0, 0, 0)

# We want to evaluate the micro system with all elements included
all_nodes = (0, 1, 2, 3, 4, 5)

# Create a subsystem to analyze
micro_subsystem = pyphi.Subsystem(network, current_state, all_nodes)

# Find the set of mechanisms (unpartitioned_constellation) and integrated
# information (phi) for the micro subsystem
micro_result = pyphi.compute.big_mip(micro_subsystem)

# its Phi value is 0.215 and it has 6 mechanisms
micro_result.phi
len(micro_result.unpartitioned_constellation)

# Next we want to evaluate the blackbox system shown in Fig. 2.
# We will group A, B, C into a black-box element with output C, and D, E, F
# into a black-box element with output F.
partition = ((0, 1, 2), (3, 4, 5))
output_indices = (2, 5)
blackbox = pyphi.macro.Blackbox(partition, output_indices)

# Create a subsystem to analyze, at a time scale of two micro time steps
time_scale = 2
blackbox_subsystem = pyphi.macro.MacroSubsystem(network, current_state, all_nodes,
blackbox=blackbox, time_scale=time_scale)

# Find the set of mechanisms (unpartitioned_constellation) and integrated
# information (phi) for the blackbox subsystem
blackbox_result = pyphi.compute.big_mip(blackbox_subsystem)

# its Phi value is 0.639 and it has 2 mechanisms
blackbox_result.phi
len(blackbox_result.unpartitioned_constellation)
```