Quantifying causal emergence shows that macro can beat micro

Erik P. Hoel, Larissa Albantakis, and Giulio Tononi[1]

Department of Psychiatry, University of Wisconsin, Madison, WI 53719

Edited by Michael S. Gazzaniga, University of California, Santa Barbara, CA, and approved October 22, 2013 (received for review August 6, 2013)

Causal interactions within complex systems can be analyzed at multiple spatial and temporal scales. For example, the brain can be analyzed at the level of neurons, neuronal groups, and areas, over tens, hundreds, or thousands of milliseconds. It is widely assumed that, once a micro level is fixed, macro levels are fixed too, a relation called supervenience. It is also assumed that, although macro descriptions may be convenient, only the micro level is causally complete, because it includes every detail, thus leaving no room for causation at the macro level. However, this assumption can only be evaluated under a proper measure of causation. Here, we use a measure [effective information (*EI*)] that depends on both the effectiveness of a system's mechanisms and the size of its state space: *EI* is higher the more the mechanisms constrain the system's possible past and future states. By measuring *EI* at micro and macro levels in simple systems whose micro mechanisms are fixed, we show that for certain causal architectures *EI* can peak at a macro level in space and/or time. This happens when coarse-grained macro mechanisms are more effective (more deterministic and/or less degenerate) than the underlying micro mechanisms, to an extent that overcomes the smaller state space. Thus, although the macro level supervenes upon the micro, it can supersede it causally, leading to genuine causal emergence—the gain in *EI* when moving from a micro to a macro level of analysis.

In science, it is usually assumed that, the better one can characterize the detailed causal mechanisms of a complex system, the more one can understand how the system works. At times, it may be convenient to resort to a "macro"-level description, either because not all of the "micro"-level data are available, or because a rough model may suffice for one's purposes. However, a complete understanding of how a system functions, and the ability to predict its behavior precisely, would seem to require the full knowledge of causal interactions at the micro level. For example, the brain can be characterized at a macro scale of brain regions and pathways, a meso scale of local populations of neurons such as minicolumns and their connectivity, and a micro scale of neurons and their synapses (1). With the goal of a complete mechanistic understanding of the brain, ambitious programs have been launched with the aim of modeling its micro scale (2).

The reductionist approach common in science has been successful not only in practice, but has also been supported by strong theoretical arguments. The chief argument starts from the intuitive notion that, when the properties of micro-level physical mechanisms of a system are fixed, so are the properties of all its macro levels—a relation called "supervenience" (3). In turn, this relation is usually taken to imply that the micro mechanisms do all of the causal work, i.e., the micro level is causally complete. This leaves no room for any causal contribution at the macro level; otherwise, there would be "multiple causation" (4). This "causal exclusion" argument is often applied to argue against the possibility for mental causation above and beyond physical causation (5), but it can be extended to all cases of supervenience, including the hierarchy of the sciences (6).

Some have nevertheless argued for the possibility that genuine emergence can occur. Purported examples go all of the way from the behavior of flocks of organisms (7) to that of ant colonies (8), brains (9), and human societies (10). Unfortunately, it remains unclear what would qualify some systems as truly emergent and others as reducible to their micro elements. Also, most arguments in favor of emergence have been qualitative (11). A convincing case for emergence must demonstrate that higher levels can be causal above and beyond lower levels ["causal emergence" (*CE*)]. So far, the few attempts to characterize emergence quantitatively (12) have not been based on causal models.

Here, we make use of simple simulated systems, including neural-like ones, to show quantitatively that the macro level can causally supersede the micro level, i.e., causal emergence can occur. We do so by perturbing each system through its entire repertoire of possible causal states ("counterfactuals," in the general sense of alternative possibilities) and evaluating the resulting effects using "effective information" (*EI*) (13). *EI* is a general measure for causal interactions because it uses perturbations to capture the effectiveness/selectivity of the mechanisms of a system in relation to the size of its state space. As will be pointed out, *EI* is maximal for systems that are deterministic and not degenerate, and decreases with noise (causal divergence) and/or degeneracy (causal convergence).

For each system, we completely characterize the causal mechanisms at the micro level, fixing what can happen at any macro level (supervenience). Macro levels are defined by coarse graining the micro elements in space and/or time, and this mapping defines the repertoire of possible causes and effects at each level. By comparing *EI* at different levels, we show that, depending on how a system is organized, causal interactions can peak at a macro rather than at a micro spatiotemporal scale. Thus, the macro may be causally superior to the micro even though it supervenes upon it. Evaluating the changes in *EI* that arise from coarse or fine graining a system provides a straightforward way of quantifying both emergence and reduction.

Theory

In what follows, we consider discrete systems *S* of connected binary micro elements that implement logical functions (mechanisms) over their inputs. We first introduce a state-dependent measure of causation, the "cause" and "effect information" of a single

Significance

Properly characterizing emergence requires a causal approach. Here, we construct causal models of simple systems at micro and macro spatiotemporal scales and measure their causal effectiveness using a general measure of causation [effective information (*EI*)]. *EI* is dependent on the size of the system's state space and reflects key properties of causation (selectivity, determinism, and degeneracy). Although in the example systems the macro mechanisms are completely specified by their underlying micro mechanisms, *EI* can nevertheless peak at a macro spatiotemporal scale. This approach leads to a straightforward way of quantifying causal emergence as the supersedence of a macro causal model over a micro one.

Author contributions: E.P.H., L.A., and G.T. designed research; E.P.H. and L.A. performed research; E.P.H., L.A., and G.T. analyzed data; and E.P.H., L.A., and G.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. E-mail: gtononi@wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1314922110/-/DCSupplemental.

system state s_0, before we describe the state-independent *EI* of the system *S*.

State-Dependent Causal Analysis. The micro mechanisms of *S* specify its state-to-state transition probability matrix (TPM) at a micro time step *t*. Building upon the perturbational framework of causal analysis developed by Judea Pearl (14; see also ref. 18), the TPM can be obtained by perturbing *S* at t_0 (13) into all possible *n* initial states with equal probability $1/n$ $[do(S = s_i) \quad \forall i \in 1 \ldots n\ ]$. Perturbing the system in this way corresponds to the unconstrained repertoire (probability distribution) of possible causes U^C, and determines the probability of the resulting states at t_{+1}, corresponding to the unconstrained repertoire of possible effects U^E. Although U^C is thus identical to the uniform distribution *U* [with $p(s) = 1/n, \ \forall s \in S\ ]$, U^E is typically not uniform. A current system state $S = s_0$ is associated with the probability distribution of past states that could have caused it ("cause repertoire $S_P|s_0$," obtained by Bayes' rule), and the probability distribution of future states that could be its effects ("effect repertoire $S_F|s_0$"). A system's mechanisms and current state thus constrain both the repertoire of possible causes U^C and that of possible effects U^E. An informational measure of the causal interactions in the system (15) can then be defined as the difference [here Kullback–Leibler divergence (D_{KL}) (16)] between the constrained and unconstrained distributions:

$$\text{Cause information}(s_0) = D_{\mathrm{KL}}\left((S_P|s_0), U^C\right),$$

$$\text{Effect information}(s_0) = D_{\mathrm{KL}}\left((S_F|s_0), U^E\right).$$

Cause/effect information depends on two properties: (*i*) the size of the system's state space (repertoire of alternatives), because both are bounded by $\log_2(n)$; (*ii*) the effectiveness of the system's mechanisms in specifying past and future states. To isolate effectiveness from size, we define the following normalized coefficients: Cause coefficient$(s_0) = \frac{\text{Cause Information}(s_0)}{\log_2(n)}$, Effect coefficient$(s_0) = \frac{\text{Effect Information}(s_0)}{\log_2(n)}$.

The "cause coefficient" describes to what extent a state is sufficient to specify its past causes, and the "effect coefficient" indicates how necessary a state is to specify its future effects (Fig. 1*B*). In turn, the effect coefficient itself is a function of two terms, "determinism" and "degeneracy" (see *Effect Coefficient and Effectiveness (Eff) Expressed as Determinism and Degeneracy* for derivation):

$$\begin{aligned} \text{Effect coefficient}(s_0) &= \text{Determinism coefficient}(s_0) \\ &\quad - \text{Degeneracy coefficient}(s_0) \\ &= \frac{1}{\log_2(n)} \sum_{s_F \in U^E} p(s_F|s_0) \log_2\left(n \cdot p(s_F|s_0)\right) \\ &\quad - \frac{1}{\log_2(n)} \sum_{s_F \in U^E} p(s_F|s_0) \log_2\left(n \cdot p(s_F)\right). \end{aligned}$$

The determinism coef. is the difference $D_{\mathrm{KL}}((S_F|s_0), U)$ between the effect repertoire and the uniform distribution (U) of system states, divided by $\log_2(n)$, and measures how deterministically (reliably) s_0 leads to the future state of the system: it is "1" (complete determinism) when the current state leads to a single future state with probability $p = 1$, and is "0" (complete indeterminism or noise) if it could be followed by every future state with $p = 1/n$. The degeneracy coef. measures to what degree there is deterministic convergence (not due to noise) from other states onto the future states specified by s_0. In broad terms, degeneracy refers to multiple ways of deterministically achieving the same effect or function (17, 18). The degeneracy coef. is 1 (complete degeneracy) when s_0 specifies the same future state as all other states, and 0 when s_0 specifies a unique future state (no degeneracy).

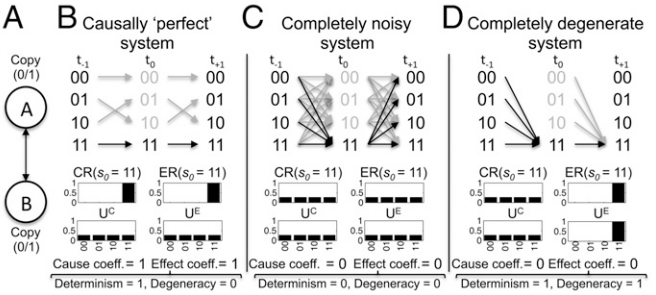

A
B Causally 'perfect' system
C Completely noisy system
D Completely degenerate system
Copy (0/1)
A
B
Copy (0/1)
t-1
t0
t+1
00
01
10
11
CR(s0 = 11)
ER(s0 = 11)
UC
UE
1
0.5
0
Cause coeff. = 1 Effect coeff. = 1
Determinism = 1, Degeneracy = 0
Cause coeff. = 0 Effect coeff. = 0
Determinism = 0, Degeneracy = 0
Cause coeff. = 0 Effect coeff. = 0
Determinism = 1, Degeneracy = 1

Fig. 1. Cause and effect coefficients in example systems with different causal architectures. (*A*) The systems consist of two interconnected binary COPY gates with possible states 0 and 1. (*B*) A causally perfect system, in which each state has one cause and one effect. Thus, $s_0 = [11]$ has a cause and effect coefficient (coef.) of 1. Moreover, there is no divergence (determinism coef. = 1) and no convergence (degeneracy coef. = 0). (*C* and *D*) In both the completely indeterministic and completely degenerate systems, state $s_0 = [11]$ is completely insufficient to specify past system states and completely unnecessary to specify future states (cause and effect coefficient = 0). Note that the degeneracy coef. is 0 in the completely noisy system, because all convergence is due to noise alone.

Both cause and effect coefficients are minimal (0) in a completely noisy or completely degenerate system (Fig. 1 *C* and *D*) and maximal (1) in a deterministic, nondegenerate system (*Bounds of Cause and Effect Coefficients and Effectiveness Eff(S)*). The contribution of a single state to the system's determinism and degeneracy are best demonstrated by decomposing the effect coefficient. Although the cause coefficient also reflects the degeneracy and determinism of the system, it is not subdivided further here.

State-Independent Causal Analysis. A state-independent informational measure of a system's causal architecture can be obtained by taking the expected value of cause or effect information over all system states, a quantity called effective information (*EI*):

$$\begin{aligned} EI(S) &= \langle \text{Cause Information}(s_0) \rangle = \sum_{s_0 \in U^E} p(s_0) D_{\mathrm{KL}}\left((S_P|s_0), U^C\right) \\ &= \langle \text{Effect Information}(s_0) \rangle = \frac{1}{n} \sum_{s_0 \in U^C} D_{\mathrm{KL}}\left((S_F|s_0), U^E\right). \end{aligned}$$

The two terms are identical, because the system is assumed to be time invariant ($\langle t_{-1} \rightarrow t_0 \rangle = \langle t_0 \rightarrow t_{+1} \rangle$), and cause and effect information are related via Bayes' rule. *EI* is also the mutual information (*MI*) between all possible causes and their effects, $MI(U^C;U^E)$ (*Effective Information EI(S) Expressed in Terms of Cause and Effect Information and Mutual Information MI*).

As a measure of causation, *EI* captures how effectively (deterministically and uniquely) causes produce effects in the system, and how selectively causes can be identified from effects. As with the state-dependent measures, the effectiveness (*Eff*) of the causal interactions within a system can be captured by normalizing *EI* by the system's size: $Eff(S) = EI(S)/log_2(n)$. Also as in the state-independent case, effectiveness can be split into two components, determinism and degeneracy:

$$\begin{aligned} Eff(S) &= \langle \text{Determinism coefficient}(s_0) \rangle \\ &\quad - \langle \text{Degeneracy coefficient}(s_0) \rangle \\ &= \langle D_{\mathrm{KL}}((S_F|s_0), U) \rangle / \log_2(n) - D_{\mathrm{KL}}\left(U^E|U\right) / \log_2(n). \end{aligned}$$

Thus, *Eff(S)* = 1 if *EI* is maximal for a given system size, and decreases with indeterminism (divergence due to noise) or degeneracy (deterministic convergence), with $Eff(S) = 0$ for completely noisy or degenerate systems (Fig. 1 *C* and *D*). In a system with perfect effectiveness (Fig. 1*B*), each cause has a unique effect, and each effect has a unique cause. Thus, such a system [where $Eff(S) = 1$] is perfectly retrodictive/predictive, in the sense that not only the unique future trajectory, but also the unique past trajectory of all states can be deduced from the TPM (complete causal reversibility).

Levels of Analysis. A finite, discrete system S can be considered at various levels, from the most fine-grained micro causal model S_m through various coarse-grained causal models S_M. All macro levels S_M are assumed to be "supervenient" on the micro level S_m: given the micro elements of S_m and the causal relationships between them, all other members of $\{\mathbf{S}\}$—the set of all possible causal models of system S—are fixed as well (19). Although S_m fixes S_M, any S_M may be fixed by a number of different lower level descriptions, a property known as "multiple realizability" (20).

Groupings. Micro elements are binary and labeled by Latin letters {A, B, C…}, macro elements by Greek letters {α, β, γ…}. Micro states are labeled {1, 0} and macro states {"on," "bursting," "quiet"…). Micro elements can be grouped into macro elements spatially, temporally, or both. Micro states are grouped into macro states through a mapping $M : S_m \rightarrow S_M$. The mapping must be exhaustive and disjunctive over micro elements (all of the states of one micro element must be mapped to the states of the same macro element; note that a macro element can consist of a single micro element as long as the state space of the system is reduced). Moreover, the mapping must be such that no micro-level information is available at the macro level (the identity of the micro elements grouped into a macro element is lost). For example, the grouping of the four states of two micro elements into the two states of one macro element as [[00, 01, 10] = off, [11] = on] is permitted, whereas the grouping [[00, 01], [10, 11]] is not, because distinguishing 01 from 10 requires knowing the identity of the micro elements.

Level-Specific Perturbations. Causal analysis at the micro level S_m, requires setting S into all possible micro states with equal probability (i.e., testing all micro alternatives) and determining the resulting effects. When moving to a macro level S_M, S must similarly be set into all possible macro states with equal probability (i.e., testing all macro alternatives). To causally assess any macro state, then, one must set S into all of the n_{micro} micro states $\{s_m\}$ that are grouped into the corresponding macro state s_M, and average over the effects. This is done using a "macro perturbation": $do(S_M = s_M) = \frac{1}{n_{micro}} \sum_{s_{m,i} \in s_M} do(S_m = s_{m,i})$. Using such macro perturbations, one can obtain cause/effect information and EI for every coarse grain of S_m. EI at each macro level is then equivalent to the MI between the set of macro causes and their macro effects.

Causal Emergence/Reduction. Finally, by assessing $EI(S)$ over all coarse grains of S_m, one can ask at which level of $\{\mathbf{S}\}$ causation reaches a maximum. This provides an analytical definition of causal emergence, expressed in bits: $CE = EI(S_M) - EI(S_m)$.

Thus, if $EI(S)$ is maximal for a macro-level S_M rather than the micro-level S_m, then $CE > 0$ and causal emergence occurs. If for every macro-level $CE < 0$, causal reduction holds. Although the focus here is on emergence/reduction relative to the micro-level S_m, the above measure can of course be used to compare different macro levels.

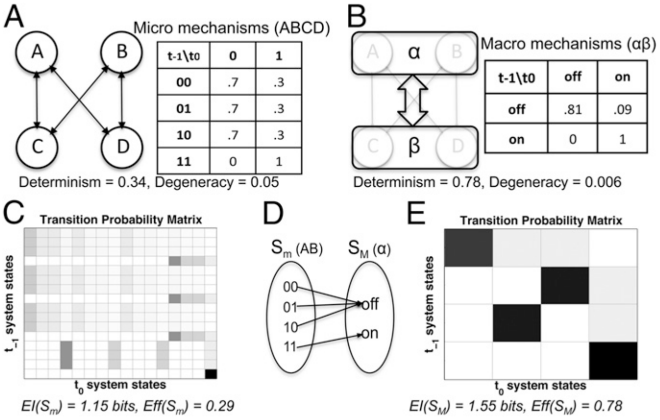
As mentioned above, $EI(S)$ depends on both the size of the system's repertoire of states and on the effectiveness of its mechanisms. When moving from one system level to another, both terms change as the state space becomes smaller or larger, and the individual states become more or less selective with respect to the past, and more or less determined or degenerate with respect to the future. The respective informational contributions of repertoire size and effectiveness to $\Delta EI(S)$ can be expressed separately as follows: $\Delta I_{Eff} = (Eff(S_M) - Eff(S_m)) \cdot log_2(n_M)$, $\Delta I_{Size} = Eff(S_m) \cdot (log_2(n_M) - log_2(n_m))$, where $n_{m/M}$ is the state repertoire size of $S_{m/M}$. It follows that $\Delta EI = \Delta I_{Eff} + \Delta I_{Size} = CE$. A positive ΔI_{Eff} can thus be due to the macro reducing the degeneracy of the micro level, increasing the determinism of the micro level, or both. Notably, coarse graining the micro-level S_m into macro-level S_M implies that ΔI_{Size} is always negative. Hence, for causal emergence to occur [$EI(S_M) > EI(S_m)$], the increase in effectiveness ΔI_{Eff} must outweigh the decrease in ΔI_{Size}.

Results

Causal analysis was performed across all coarse grains of a system [only the S_M with maximal $EI(S)$ is shown in the figures] with a custom-made Python program. Data plots were created using MATLAB. Below, we consider examples of spatial, temporal, and spatiotemporal emergence (see Fig. S1 for an example of spatial reduction).

Spatial Causal Emergence. As a proof-of-principle example, consider a system of four binary elements S_m = {ABCD} (Fig. 2*A*). Each micro mechanism is an AND-gate (two inputs) operating over some intrinsic noise. The 16 × 16 S_m TPM was constructed by setting the system into all possible micro states from [0000] to [1111] with equal probability (Fig. 2*B*). At the micro level S_m, effective information $EI(S)$ = 1.15 bits, out of maximally 4 bits, with effectiveness $Eff(S_m)$ = 0.29. The macro level S_M (Fig. 2*D*), composed of two elements {α, β}, each with states {"on," "off"}, is a coarse graining of S_m as defined by the mapping $\boldsymbol{M}$ in Fig. 2*C*. The 4 × 4 S_M TPM was obtained by setting the system into all possible macro states from [off, off] to [on, on] with equal probability (Fig. 2*E*). For the macro level, $EI(S_M)$ = 1.55 bits, higher than $EI(S_m)$ = 1.15 bits. Thus, $CE(S)$ = 0.40 bits, demonstrating that in this case the macro S_M beats the micro S_m and constitutes the optimal causal model of system S. This is because the TPM for S_M is much closer to perfect effectiveness [$Eff(S_M)$ = 0.78] and the increase in effectiveness gained by grouping ΔI_{Eff} = 0.97 bits outweighs the loss in size ΔI_{Size} = −0.57 bits. In this example, the gain in effectiveness ΔI_{Eff} at the macro level comes primarily (91%) from counteracting noise [determinism coef. (S_m) = 0.34; (S_M) = 0.78] and less so (9%) from reducing degeneracy [degeneracy coef. (S_m) = 0.05; (S_M) = 0.006].

The higher effectiveness of the macro level is also evident comparing S_m and S_M in a state-dependent manner. As an example, the cause/effect distributions for S_m in state {ABCD} = [0001] are compared with the corresponding S_M state {αβ} = [off, off] in Fig. 3. Comparing the cause/effect distributions of S_m = [0001] against the unconstrained repertoires (using D_{KL}) yields 0.83 bits of cause information and 0.43 bits of effect information. For the macro S_M, cause information is 2 bits and effect information



Micro mechanisms (ABCD)

t-1\t0	0	1
00	.7	.3
01	.7	.3
10	.7	.3
11	0	1

Determinism = 0.34, Degeneracy = 0.05

Macro mechanisms (αβ)

t-1\t0	off	on
off	.81	.09
on	0	1

Determinism = 0.78, Degeneracy = 0.006

$EI(S_m)$ = 1.15 bits, $Eff(S_m)$ = 0.29

$EI(S_M)$ = 1.55 bits, $Eff(S_M)$ = 0.78

Fig. 2. Spatial causal emergence (counteracting indeterminism). (*A*) The micro level S_m of system S is composed of identical noisy micro mechanisms. (*B*) The micro TPM. (*C*) A macro causal level S_M and its TPM are defined by the mapping $\boldsymbol{M}$ (shown for AB to α, CD to β is symmetric). (*D*) S_M and its macro mechanisms. (*E*) By reducing indeterminism and increasing effectiveness Eff, the macro beats the micro in terms of EI despite the reduced repertoire size (CE = 0.40 bits).

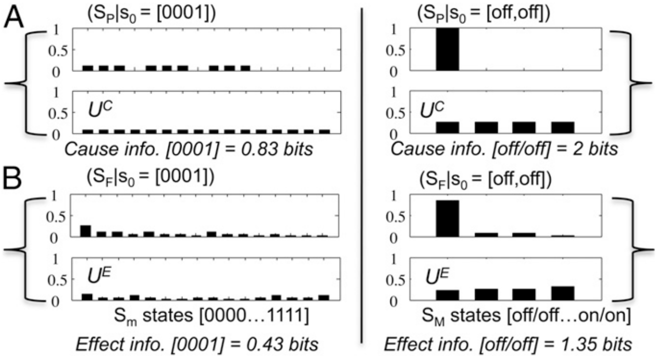


Fig. 3. State-dependent cause/effect information. (*A*) The cause information of S_m in micro state {ABCD} = [0001] is calculated as the difference (D_{KL}) between the cause repertoire of state [0001] and the unconstrained micro repertoire U^C (*Left*). The cause information of S_M in the supervening macro state {αβ} = [off/off] (*Right*) is the difference (D_{KL}), between the cause repertoire of [off/off] and the unconstrained macro repertoire U^C. (*B*) Effect information. The higher cause and effect information at the macro level is due to an increase in determinism and decrease in degeneracy, reflecting higher selectivity.

1.35 bits. The macro beats the micro because {αβ} = [off, off] is both more selective and more reliable than {ABCD} = [0001].

Causal emergence may arise not only from macro gains in determinism (as above), but also from reducing degeneracy. In Fig. 4, micro elements A–F are deterministic AND gates connected in a way that ensures high degeneracy (Fig. 4*A*, determinism coef. = 1; degeneracy coef. = 0.6), resulting in $Eff(S_m) = 0.4$ and $EI(S_m) =$ 2.43 bits (Fig. 4*C*). The optimal macro groups the six micro AND gates into three macro COPY gates (αβγ) (Fig. 4*B*). Both macro and micro are deterministic, but by eliminating degeneracy $\Delta I_{Eff} = 1.79$ bits $> -\Delta I_{Size} = 1.22$ bits. As a result, $Eff(S_M) = 1$, EI $(S_M) = 3$ bits, and the macro emerges over the micro ($CE = 0.57$ bits).

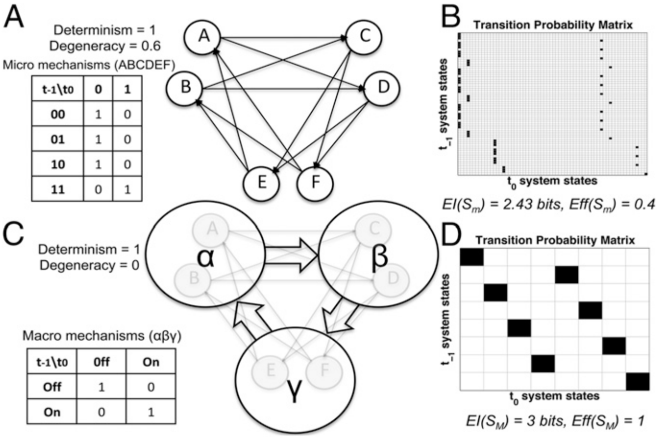
Temporal Causal Emergence. The same principles allowing for emergence through spatial groupings hold for temporal groupings, which coarse grain micro time steps (t_x) into macro time steps (T_x). The example in Fig. 5 shows micro elements that, upon receiving an input "burst" of two spikes, respond with an output burst of two spikes. Thus, elements implement second-order Markov mechanisms over both inputs and outputs (Fig. 5*A*). Fig. 5*B* shows that causal interactions assessed over one micro time step are weak [$EI(S_m) = 0.16$ bits; $Eff(S_m) = 0.03$] because they fail to capture the second-order mechanisms. By contrast, causal analysis over two micro time steps (Fig. 5*C*) gives $EI = 1.38$ bits and $Eff(S_m) = 0.34$. The temporal grouping of micro into macro states α = {A_t, A_{t+1}} and β = {B_t, B_{t+1}} (Fig. 5*D*) is analogous to the spatial grouping in Fig. 2: {00, 01, 10} = {off} and {11} = {on}. Over macro time steps, the system becomes fully deterministic and nondegenerate, $EI(S_M) = 2$ bits, $Eff(S_M) =$ 1, and $CE(S) = 0.62$ bits (Fig. 5 *E* and *F*).

Spatiotemporal Causal Emergence. In general, emergence may occur simultaneously over space and time (Fig. 6). As in Fig. 5, the nine neural-like micro elements in Fig. 6*A* are second-order Markov mechanisms, integrating inputs and outputs over two micro time steps, t_{-2} t_{-1}, and t_0 t_{+1}, respectively [compare to longer time constants of NMDA receptors (21)]. Moreover, in the examples above, the micro elements within a macro element were not connected and were causally equivalent. To demonstrate that this is not a requisite for causal emergence, in Fig. 6, the micro elements are fully connected and causally heterogeneous (self-connections not drawn). All elements are spontaneously active (1) with heterogeneous probabilities: p(A/D/G) = 0.45; p(B/E/H) = 0.5; p(C/F/I) = 0.55. The elements are structured into three groups {ABC, DEF, GHI} due to different intragroup and intergroup mechanisms: within each group, if the sum of intragroup connections Σ(intra) = 0 (for two time steps), all elements stay 0 (for the next two time steps). However, if the sum of intergroup connections Σ(inter) = 6 from one or both of the other two groups over two time steps (burst of synchronous activity), p(1) is raised by 0.5 for the next two time steps (see Fig. S2 for macro and micro TPMs of a spatial system with equivalent rules). At the macro-level S_M (Fig. 6*B*), the three groups of neurons become macro elements, and two micro time steps (t_x) are grouped into one macro time step (T_x). In neural terms, these macro elements could represent "minicolumns" having three states: "inhibited" (all minicolumn neurons silent at T_x), "receptive" (some firing at T_x), or "bursting" (all firing at T_x). Macro causal interactions can be summarized as follows: if a macro element is inhibited, only receiving a burst can move it to the receptive or (more unlikely) the bursting state; otherwise, it stays inhibited. As in previous examples, the coarse-grained S_M has higher $EI(S_M) = 3.51$ bits and $Eff(S_M) = 0.74$ than S_m [$EI(S_m) = 0.59$ bits; $Eff(S_m) = 0.033$]. In this case, spatiotemporal causal emergence [$CE(S) = 2.92$ bits] is due to an increase in determinism that far outweighs a slight increase in degeneracy and the decrease in size.

Discussion

This paper provides a principled way of assessing at which spatiotemporal grain size the causal interactions within a system reach a maximum. Causal interactions are evaluated by effective information (*EI*), a measure that is sensitive both to the effectiveness of the system's mechanisms and to the size of its state space. Examples with simulated systems demonstrate that, after coarse graining the micro mechanisms in both space and time, *EI* can be higher at a macro level than at a micro level. In these cases, the macro mechanisms, rather than the micro ones, can be said to be doing the causal work within a system.

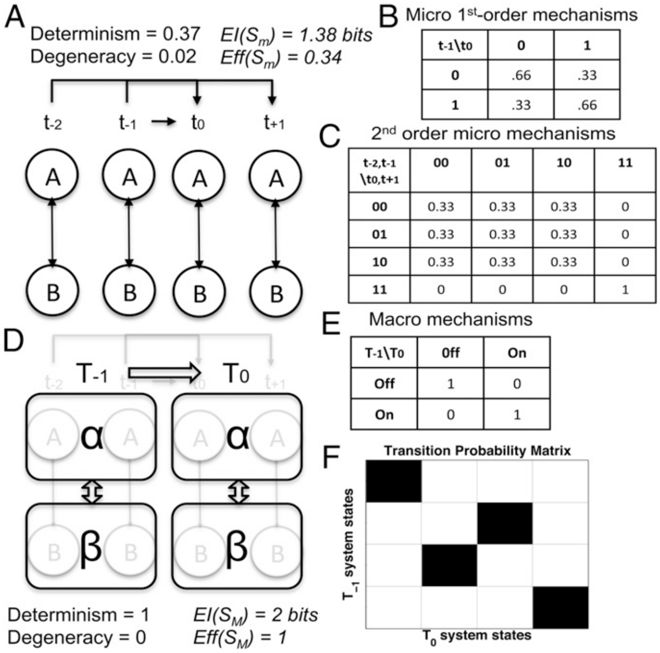
Effective Information, Effectiveness, and Emergence. As shown here, *EI* corresponds to the "effectiveness" of a system's mechanisms multiplied by repertoire size, expressed in bits. Effectiveness *Eff* (*S*) is the average of the effect coefficients over all system states. The effect coefficient measures to what extent the current system state is necessary to specify the system's future state. This, in turn, is a function of determinism minus degeneracy. On the cause side, the equivalent to the effect coefficient is the cause coefficient, which measures to what extent the current state is sufficient to specify the system's past state. For a particular current state, cause and effect coefficients may differ: for example, a state may have many causes but only one effect. However, the average of the effect coefficients over system states, i.e., effectiveness,



t-1\t0	0	1
00	1	0
01	1	0
10	1	0
11	0	1

t-1\t0	Off	On
Off	1	0
On	0	1

Fig. 4. Spatial causal emergence (counteracting degeneracy). (*A*) A degenerate S_m with deterministic AND gates. (*B*) The cycle of AND gates is mapped onto a cycle of COPY gates at the macro level. (*C*) The deterministic but degenerate micro TPM. (*D*) The deterministic macro TPM with zero degeneracy. By eliminating degeneracy and achieving perfect effectiveness, the macro beats the micro (*CE* = 0.57 bits).



t-1\t0	0	1
0	.66	.33
1	.33	.66

t-2,t-1 \t0,t+1	00	01	10	11
00	0.33	0.33	0.33	0
01	0.33	0.33	0.33	0
10	0.33	0.33	0.33	0
11	0	0	0	1

T-1\T0	Off	On
Off	1	0
On	0	1

Fig. 5. Temporal causal emergence. (*A*) S_m is composed of second-order Markov mechanisms A and B: at t_0, each mechanism responds based on the inputs at t_{-2} and t_{-1}, and outputs over t_0 and t_{+1}. (*B*) Causal analysis over one micro time step gives an incomplete view of the system. (*C*) A causal analysis over two micro time steps reveals the second-order Markov mechanisms. (*D*) The optimal macro system S_M groups two micro time steps into one macro time step for macro elements {α,β}. (*E*) Each coarse grained macro mechanism effectively corresponds to a deterministic COPY gate. (*F*) The macro one-time step TPM S_M has $Eff(S_M) = 1$, and the micro two-time step TPM has $Eff(S_m) = 0.34$; $CE = 0.62$ bits.

corresponds to the average of the cause coefficients (weighted by the probability of the effects). In other words, within a time-invariant system the average selectivity of the causes corresponds to the average selectivity of the effects. Note that, in principle, other measures of causation that, like *EI*, reflect causal structure (selectivity, determinism, degeneracy) and system size, should demonstrate causal emergence as well.

The main result obtained in the simulations is that coarse graining, both in space and in time, can yield a higher value of *EI*. This happens even though the micro has, by definition, a larger state space than the macro—an advantage with respect to *EI*. Given this inherent advantage of the micro, it is understandable why the default scientific strategy for analyzing systems has been one of reduction (*Causal Reduction*). However, the examples presented above show that the inherent loss in *EI* due to the macro's smaller repertoire size can be offset if the macro achieves a greater gain in effectiveness. In turn, greater effectiveness stems from macro mechanisms constructed from their constituting micro mechanisms in such a way that, at the macro level, determinism is increased and/or degeneracy is decreased. Genuine causal emergence can then be said to occur whenever there is a gain in *EI* ($CE > 0$) at the optimal macro level. If instead there is a loss in *EI* ($CE < 0$), causal reduction is appropriate, and the micro level is the optimal level of causal analysis. The causal approach pursued here suggests that qualitative or noncausal accounts of emergence may have been hindered by not being able to characterize how and why a macro level can actually have greater causal effectiveness than a micro level (22, 23).

Micro Macro Mappings and Repertoires of Alternatives. The present approach makes it possible to compare causation at the micro and macro levels in a fair manner. First, the simulated examples are such that the macro supervenes strictly upon the micro: once the micro is defined, all macro levels are fixed. Specifically, no extra causal ingredients are added at the macro level, such as rules that apply to the macro only (24). Furthermore, the mapping of micro into macro elements is such that the identity of micro elements is lost; otherwise, the macro level would have access to micro-level information that could offset its reduced repertoire size. Finally, when causation is evaluated a uniform distribution of alternatives is imposed independently at the micro and macro levels. For this uniform distribution of perturbations to be imposed at the macro level, the probability of the underlying micro perturbations must be modified by averaging the micro states that map into the same macro state. The modified distribution of micro perturbations yielding a uniform distribution of macro perturbations makes *EI* sensitive to the causal structure at each level, ultimately allowing the supervening macro *EI* to exceed the micro *EI*.

Emergence as an Intrinsic Property of a System. *EI* is a causal measure, because it requires perturbing the system in all possible ways and evaluating the resulting effects on the system. It is also an informational measure, because its value depends on the size of the repertoire of alternatives. Indeed, in the present approach, causation and information are necessarily linked (25), hence the term "effective information." Finally, measuring *EI* reveals an "intrinsic" property of the system, namely the average effectiveness/selectivity of all possible system states with respect to the system itself. Effectiveness/selectivity can be assessed at multiple spatiotemporal grains, and the particular spatiotemporal grain at which *EI* reaches a maximum is again an intrinsic property of the system. This in no way precludes an observer from profitably investigating the system's properties at other macro levels, at the micro level, or at multiple levels at once (e.g., neuroscientists studying the brain at the level of ion channels, individual neurons, local field potentials, or functional magnetic resonance signals). However, causal emergence implies that the macro level with highest *EI* is the one that is optimal to characterize, predict, and retrodict the behavior of the system—the one that "carves nature at its joints" (26).

The search for the macro level at which *EI* is maximal has a parallel in information theory: channel capacity is an intrinsic property defined as the maximal amount of information that can be transmitted along the channel at a certain rate, found by searching over all possible input distributions (27). Finding the optimal level of coarse graining for causal emergence is based on a similar search, with several differences. First, *EI* is evaluated using perturbations over the system itself, rather than across a channel (the system is its own input and output). Second, the probability distributions over micro states that can be considered must conform to a proper mapping of micro into macro elements (or time intervals). Additional connections of causal emergence to established measures, such as reversibility

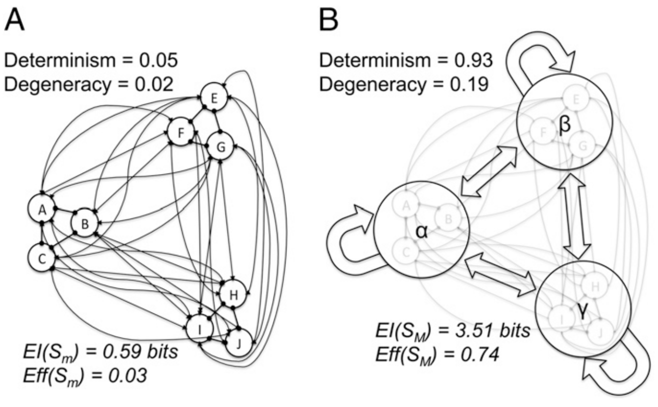


Fig. 6. Spatiotemporal causal emergence. (*A*) A "neuronal" system merging the temporal characteristics of the system in Fig. 5 with a differentiated spatial structure (Fig. S2). Regular and rounded arrows indicate intergroup and intragroup connections, respectively. (*B*) Each macro element receives inputs from itself and the other macro element. The macro level beats the micro level, leading to spatiotemporal emergence [$CE(S) = 2.92$ bits].

and lumping in Markov processes (28), or epsilon machines (29), are a potential subject for future work.

Causal Exclusion and Its Implications. Causal analysis as presented here endorses both supervenience (no extra causal ingredients at the macro level) and causal exclusion [for a given system at a given time, causation occurs at one level only, otherwise causes would be double counted (4)]. However, causal analysis also demonstrates that EI can actually be maximal at a macro level, depending on the system's architecture. In such cases, causal exclusion turns the reductionist assumption on its head, because to avoid double-counting causes, optimal macro causation must exclude micro causation. In other words, macro mechanisms can always be decomposed to their constituting micro mechanisms (supervenience); however, if there is emergence, macro causation does not reduce to micro causation, in which case the macro wins causally against the micro and takes its place (supersedence). The notion of irreducibility among levels (does the macro beat the micro?) is complemented by the notion of irreducibility among subsets of elements within a level [is the whole more than its parts (15, 25)?]. From the perspective of a system, emergence ($CE > 0$) implies causal "self-definition" at the optimal macro level—the one at which its causal interactions "come into focus" (30) and "the action happens."

Applicability to Real Systems. Measuring EI exhaustively, across all micro/macro levels, is not feasible for complex physical or biological systems (*Applicability—Network Motifs as Indicators of Emergence*). However, some useful guidelines can be derived from the above analysis: (*i*) if $Eff(S_m) \geq Eff(S_M)$, then causal emergence is impossible and causal reduction holds; (*ii*) if $EI(S_m) > \log_2(n_M)$, where n_M is the state repertoire size of S_M, causal reduction holds; (*iii*) if for some coarse graining, Eff increases drastically, causal emergence is to be suspected (as $\Delta I_{Eff} >> -\Delta I_{Size}$). Therefore, systems that already are close to maximal effectiveness at the micro level (Fig. S1) indicate causal reduction. By contrast, heavily interconnected groups of elements with spontaneous activity and the ability to distinguish between intragroup and intergroup connections, such as the simplified neural system of Fig. 6, are more suitable for emergence.

In real neural systems, one could compare the respective effective information at the micro scale of single neurons over millisecond intervals, the meso scale of neuronal groups over hundreds of milliseconds, and the macro scale of brain regions over several seconds (using tools such as optogenetics and calcium imaging). In this way, classic notions, such that cortical minicolumns may constitute the fundamental units of brain function (31), or that the cortex works by population coding in space (32) or rate coding in time (33) in the face of high intertrial variability (34), could then be tested rigorously using a measure of effectiveness. Examining small motifs that are overrepresented in complex networks [such as brains (35)] could determine whether the network as a whole is biased toward emergence or reduction. Heuristic assessments of the likelihood of emergence could also rely on the analysis of wiring diagrams, which can offer an estimate of degeneracy, combined with knowledge of the amount of intrinsic noise in a system, which can provide an estimate of determinism.

Conclusions

The approach to emergence investigated here provides theoretical support for the intuitive idea that, to find out how a system works, one should find the "differences that make [most of] a difference" to the system itself (25) (cf. ref. 36). It also suggests that complex, multilevel systems such as brains are likely to "work" at a macro level because, in biological systems, selectional processes must deal with unpredictability and lead to degeneracy (18). This may also apply to some engineered systems designed to compensate for noise and degeneracy. More broadly, this view of causal emergence suggests that the hierarchy of the sciences, from microphysics to macroeconomics, may not just be a matter of convenience but a genuine reflection of causal gains at the relevant levels of organization.

ACKNOWLEDGMENTS. We thank M. Boly, C. Cirelli, A. Hashmi, C. Koch, L. Morton, A. Nere, M. Oizumi, and L. Shapiro for helpful discussions, and P. Rana for assisting with the Python software. This work has been supported by Defense Advanced Research Planning Agency Grant HR 0011-10-C-0052 and The Paul G. Allen Family Foundation.

1. Sporns O, Tononi G, Kötter R (2005) The human connectome: A structural description of the human brain. *PLoS Comput Biol* 1(4):e42.
2. Markram H (2006) The blue brain project. *Nat Rev Neurosci* 7(2):153–160.
3. Davidson D (1980) Mental events. *Readings in Philosophy of Psychology*, ed Block N (Harvard Univ Press, Cambridge, MA), Vol 1, pp 107–119.
4. Kim J (1993) *Supervenience and Mind: Selected Philosophical Essays* (Cambridge Univ Press, Cambridge, UK).
5. Kim J (2000) *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (MIT Press, Cambridge, MA).
6. Bontly T (2002) The supervenience argument generalizes. *Philos Stud* 109:75–96.
7. Seth A (2008) Measuring emergence via nonlinear Granger causality. *ALIFE* 2008:545–552.
8. Hölldobler B, Wilson E (2009) *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies* (W. W. Norton, New York).
9. Sperry R (1983) *Science and Moral Priority: Merging Mind, Brain, and Human Values* (Columbia Univ Press, New York).
10. Sawyer R (2005) *Social Emergence: Societies as Complex Systems* (Cambridge Univ Press, Cambridge, UK).
11. Broad C (1925) *The Mind and Its Place in Nature* (Routledge & Kegan Paul, London).
12. Bar-Yam Y (2004) A mathematical theory of strong emergence using multiscale variety. *Complexity* 9:15–24.
13. Tononi G, Sporns O (2003) Measuring information integration. *BMC Neurosci* 4:31.
14. Pearl J (2000) *Causality: Models, Reasoning and Inference* (Cambridge Univ Press, Cambridge, UK).
15. Albantakis L, Hoel EP, Koch C, Tononi G (2013) Intrinsic Causation and Consciousness. *Association for the scientific study of consciousness conference (ASSC17)*. Available at www.theassc.org/files/assc/docs/ASSC17-PB-070113-online-version-with-Addendum.pdf. Accessed November 2, 2013.
16. Kullback S (1997) *Information Theory and Statistics* (Dover Publications, New York).
17. Edelman GM (1987) *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic Books, New York).
18. Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. *Proc Natl Acad Sci USA* 96(6):3257–3262.
19. Stalnaker R (1996) Varieties of supervenience. *Philos Perspect* 10:221–241.
20. Fodor J (1974) Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28:97–115.
21. Jahr CE, Stevens CF (1990) Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J Neurosci* 10(9):3178–3182.
22. Bedau M (1997) Weak emergence. *Noûs* 31:375–399.
23. Chalmers D (2006) Strong and weak emergence. *The Reemergence of Emergence*, eds Clayton P, Davies P (Oxford Univ Press, Oxford), pp 244–256.
24. Butterfield J (2012) Laws, causation and dynamics at different levels. *Interface Focus* 2(1):101–114.
25. Tononi G (2012) Integrated information theory of consciousness: An updated account. *Arch Ital Biol* 150(2-3):56–90.
26. Hamilton E, Cairns H (1961) *The Collected Dialogues of Plato: Including the Letters* (Pantheon Books, New York).
27. Shannon CE (1997) The mathematical theory of communication. 1963. *MD Comput* 14(4):306–317.
28. Kemeny J, Snell J (1976) *Finite Markov Chains* (Springer, New York).
29. Shalizi C, Crutchfield J (2001) Computational mechanics: Pattern and prediction, structure and simplicity. *J Stat Phys* 104:817–879.
30. Alexander S (1920) *Space, Time, and Deity: The Gifford Lectures at Glasgow, 1916–1918* (Macmillan, London).
31. Buxhoeveden DP, Casanova MF (2002) The minicolumn hypothesis in neuroscience. *Brain* 125(Pt 5):935–951.
32. Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233(4771):1416–1419.
33. London M, Roth A, Beeren L, Häusser M, Latham PE (2010) Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* 466(7302):123–127.
34. Knoblauch A, Palm G (2005) What is signal and what is noise in the brain? *Biosystems* 79(1-3):83–90.
35. Sporns O (2010) *Networks of the Brain* (MIT Press, Cambridge, MA).
36. Fitelson B, Hitchcock C (2010) *Probabilistic Measures of Causal Strength* (Oxford Univ Press, Oxford, UK).

Supporting Information

Hoel et al. 10.1073/pnas.1314922110

Effect Coefficient and Effectiveness (*Eff*) Expressed as Determinism and Degeneracy

The state-dependent effect coefficient $(s_0)=\frac{\text{effect information}(s_0)}{\log_2(n)}$ can be described as a function of two terms, the determinism and degeneracy coefficient. To derive these two terms, the effect information (s_0), the distance between the effect repertoire $(S_F|s_0)$ and the unconstrained repertoire of effects U^E, is split into the distance between $(S_F|s_0)$ and the uniform distribution U with $p(s_U)=1/n$, and a residual term:

$$\text{Effect Information}(s_0)=D_{\text{KL}}\left((S_F|s_0),U^E\right)=\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_F)}\right)\qquad\textbf{[S1]}$$

$$=\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_U)}+\frac{p(s_U)}{p(s_F)}\right)\qquad\textbf{[S2]}$$

$$=\sum_{s_F\in U^E}p(s_F|s_0)\left(\log_2\left(\frac{p(s_F|s_0)}{p(s_U)}\right)-\log_2\left(\frac{p(s_F)}{p(s_U)}\right)\right)\qquad\textbf{[S3]}$$

$$=\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_U)}\right)-\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F)}{p(s_U)}\right)\qquad\textbf{[S4]}$$

$$(\text{using } p(s_U)=1/n)=\sum_{s_F\in U^E}p(s_F|s_0)\log_2(n\cdot p(s_F|s_0))-\sum_{s_F\in U^E}p(s_F|s_0)\log_2(n\cdot p(s_F))\qquad\textbf{[S5]}$$

$$=D_{\text{KL}}((S_F|s_0),U)-\sum_{s_F\in U^E}p(s_F|s_0)\log_2(n\cdot p(s_F)),\qquad\textbf{[S6]}$$

where s_F denotes a state of the system S_F at t_{+1} with probability $p(s_F)$ according to the unconstrained distribution of effects U^E. s_0 is the present system state. The determinism coefficient is then the left term in lines **S5** and **S6** divided by log2(n):

$$\text{Determinism coefficient}(s_0)=\frac{\sum_{s_F\in U^E}p(s_F|s_0)\log_2(n\cdot p(s_F|s_0))}{\log_2(n)}=\frac{D_{\text{KL}}((s_F|s_0),U)}{\log_2(n)},\qquad\textbf{[S7]}$$

the degeneracy coefficient the right term:

$$\text{Degeneracy coefficient}(s_0)=\frac{\sum_{s_F\in U^E}p(s_F|s_0)\log_2(n\cdot p(s_F))}{\log_2(n)},\qquad\textbf{[S8]}$$

as defined in the main article.

The effectiveness (*Eff*) of a system assesses the causal relations in a system in a state-independent manner, irrespective of the size of the system's state space:

$$Eff(S)=\frac{EI(S)}{\log_2(n)}=\frac{\langle\text{Effect Information}(s_0)\rangle}{\log_2(n)}=\frac{\sum_{s_0\in U^C}p(s_0)D_{\text{KL}}\left((S_F|s_0),U^E\right)}{\log_2(n)},\qquad\textbf{[S9]}$$

where the effective information $EI(S)$ is the average effect information of all system states s_0, distributed according to U^C, the unconstrained repertoire of causes, which is identical to the uniform distribution U; thus, here $p(s_0)=1/n$. $EI(S)$ can then be divided in the same way as the state-dependent effect information:

$$EI(S)=\langle\text{Effect Information}(s_0)\rangle,\qquad\textbf{[S10]}$$

$$=\left\langle D_{\text{KL}}((S_F|s_0),U)-\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F)}{p(s_U)}\right)\right\rangle,\qquad\textbf{[S11]}$$

$$=\langle D_{\text{KL}}((S_F|s_0),U)\rangle-\left\langle\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F)}{p(s_U)}\right)\right\rangle,\qquad\textbf{[S12]}$$

$$=\langle D_{\text{KL}}((S_F|s_0),U)\rangle-\sum_{s_0\in U^C}p(s_0)\sum_{s_F\in U^E}p(s_F|s_0)\log_2\left(\frac{p(s_F)}{p(s_U)}\right),\qquad\textbf{[S13]}$$

$$=\langle D_{\text{KL}}((S_F|s_0),U)\rangle-\sum_{s_F\in U^E}p(s_F)\log_2\left(\frac{p(s_F)}{p(s_U)}\right),\qquad\textbf{[S14]}$$

$$=\langle D_{\text{KL}}((S_F|s_0),U)\rangle-D_{\text{KL}}\left(U^E,U\right).\qquad\textbf{[S15]}$$

The last equality is due to the fact that $p(s_F)$ is the probability of state s_F to occur at t_{+1} following U^E, the unconstrained distribution of effects (future states) obtained by setting the system S at t_0 into all possible states s_0 with equal probability $p(s_0)=1/n$.

Both, indeterminism and degeneracy at the micro level may be indicative of causal emergence (*Discussion*, main text). Note that, in previous work, it was suggested that a convergence of two causes onto the same effect—an instance of degeneracy—may actually disqualify the micro level from causation (1, 2) (although see ref. 3).

1. Yablo S (1992) Mental causation. *Philos Rev* 101:245–280.
2. List C, Menzies P (2009) Non-reductive physicalism and the limits of the exclusion principle. *J Philos* CVI(9):475–502.
3. Shapiro L, Sober E (2012) Against proportionality. *Analysis* 72:89–93.

Effective Information *EI*(*S*) Expressed in Terms of Cause and Effect Information and Mutual Information *MI*

The effective information of a system, $EI(S)$, can be obtained as the expected value of the cause or effect information. Moreover, $EI(S)$ is identical to the mutual information $MI(U^C;U^E)$: the MI between the system S set to all possible counterfactuals (system states) with equal probability (unconstrained repertoire of causes, U^C) and the resulting distribution of system states at the next time step (unconstrained repertoire of effects, U^E). Note that EI was originally introduced as a measure of causal influence of one subset of a system over another (1), whereas here it captures the overall effectiveness of system S onto itself (see refs. 2 and 3 for related measures).

In the following derivation, we start from the definition of $EI(S)$ as the average effect information of all system states s_0 as counterfactual causes [distributed according to U^C with equal probability $p(s_0)=1/n$ for all system states]:

$$\mathrm{EI}(S)=\langle \text{Effect Information}(s_0)\rangle=\sum_{s_0\in U^C} p(s_0)D_{\mathrm{KL}}\left((S_F|s_0),U^E\right)= \qquad \textbf{[S1]}$$

$$(\text{using } p(s_0)=1/n\ \forall s_0)=\frac{1}{n}\sum_{s_0\in U^C} D_{\mathrm{KL}}\left((S_F|s_0),U^E\right). \qquad \textbf{[S2]}$$

Using Bayes' rule and time invariance, we then show that the average effect information is indeed equivalent to the mutual information $MI(U^C;U^E)$ and to the expected value of the cause information, which is the average cause information of each accessible state at t_0, weighted by $p(s_0)$ according to U^E :

$$\begin{aligned}EI(S)&=\langle \text{Effect Information}(s_0)\rangle=MI\left(U^C;U^E\right)\\&=\langle \text{Cause Information}(s_0)\rangle.\end{aligned} \qquad \textbf{[S3]}$$

In detail:

$$EI(S)=\langle \text{Effect Information}(s_0)\rangle=\sum_{s_0\in U^C} p(s_0)D_{\mathrm{KL}}\left((S_F|s_0),U^E\right)= \qquad \textbf{[S4]}$$

$$=\sum_{s_0\in U^C} p(s_0)\sum_{s_F\in U^E} p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_F)}\right)= \qquad \textbf{[S5]}$$

$$=\sum_{s_0\in U^C}\sum_{s_F\in S_F} p(s_0)p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_F)}\right)= \qquad \textbf{[S6]}$$

$$(\text{Bayes' rule})=\sum_{s_0\in U^C}\sum_{s_F\in U^E} p(s_0,s_F)\log_2\left(\frac{p(s_0,s_F)}{p(s_0)p(s_F)}\right)= \qquad \textbf{[S7]}$$

$$=MI\left(U^C;U^E\right)= \qquad \textbf{[S8]}$$

$$(\text{time invariance})=\sum_{s_P\in U^C}\sum_{s_0\in U^E} p(s_P,s_0)\log_2\left(\frac{p(s_P,s_0)}{p(s_P)p(s_0)}\right)= \qquad \textbf{[S9]}$$

$$(\text{Bayes' rule})=\sum_{s_P\in U^C}\sum_{s_0\in U^E} p(s_0)p(s_P|s_0)\log_2\left(\frac{p(s_P|s_0)}{p(s_P)}\right)= \qquad \textbf{[S10]}$$

$$=\sum_{s_0\in U^E} p(s_0)\sum_{s_P\in U^C} p(s_P|s_0)\log_2\left(\frac{p(s_P|s_0)}{p(s_P)}\right)= \qquad \textbf{[S11]}$$

$$=\sum_{s_0\in U^E} p(s_0)D_{\mathrm{KL}}\left((S_P|s_0),U^C\right)=\langle \text{Cause Information}(s_0)\rangle. \qquad \textbf{[S12]}$$

MI is originally a statistical measure of how much information is shared between a source and a target (4). In the present context, MI is applied between two time steps of a system that is first perturbed into all counterfactuals (alternative states) with equal probability and then observed at the next time step. Because of the system perturbations, MI here is a causal measure. In other words, $EI(S)$ is the MI between the set of all possible causes U^C and the set of all their effects U^E. Usually, however, MI is calculated for observed distributions of system states and thus not a causal measure, but a statistical measure of correlation.

1. Tononi G, Sporns O (2003) Measuring information integration. *BMC Neurosci* 4:31.
2. Ay N, Polani D (2008) Information flows in causal networks. *Adv Complex Syst* 11(1): 17–41.
3. Korb KB, Nyberg EP, Hope L (2011) *Causality in the Sciences*, eds Illari P, Russo F, Williamson J (Oxford Univ Press, Oxford), pp 628–652.
4. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience, Hoboken, NJ).

Bounds of Cause and Effect Coefficients and Effectiveness *Eff(S)*

In the following, we will show that the cause and effect coefficients, as well as the effectiveness $Eff(S)$, are bounded between 0 and 1 ($\in[0\ldots1]$) :

$$\begin{aligned}\text{Cause coefficient}(s_0)&=\frac{\text{Cause information}(s_0)}{\log_2(n)}\\&=\frac{D_{\mathrm{KL}}\left((S_P|s_0),U^C\right)}{\log_2(n)},\end{aligned} \qquad \textbf{[S1]}$$

$$\begin{aligned}\text{Effect coefficient}(s_0)&=\frac{\text{Effect information}(s_0)}{\log_2(n)}\\&=\frac{D_{\mathrm{KL}}\left((S_F|s_0),U^E\right)}{\log_2(n)},\end{aligned} \qquad \textbf{[S2]}$$

$$\begin{aligned}Eff(S)=\frac{EI(S)}{\log_2(n)}&=\frac{\frac{1}{n}\sum_{s_0\in U^C}D_{\mathrm{KL}}\left((S_F|s_0),U^E\right)}{\log_2(n)}\\&=\langle \text{Effect coefficient}(s_0)\rangle.\end{aligned} \qquad \textbf{[S3]}$$

The lower bound (0) is given by the fact that the Kullback–Leibler divergence (D_{KL}) is always nonnegative (Gibbs' inequality). Because the cause and effect information are expressed in terms of D_{KL} and the state-independent effective information $EI(S)$ is just an average of the state-dependent values, neither of the three coefficients can be negative. It thus remains to show that cause and effect coefficients cannot exceed 1.

The cause information (s_0) is the D_{KL} between the cause repertoire $(S_P|s_0)$ and U^C, the unconstrained cause repertoire, which is identical to the uniform distribution with $p(s_P)=1/n\ \forall s_P$. It follows that

$$\begin{aligned}\text{Cause information}(s_0)&=D_{\mathrm{KL}}\left((S_P|s_0),U^C\right)\\&=\sum_{s_P\in U^C} p(s_P|s_0)\log_2\left(\frac{p(s_P|s_0)}{p(s_P)}\right)=\end{aligned} \qquad \textbf{[S4]}$$

$$=\sum_{s_P\in U^C} p(s_P|s_0)\log_2(n\cdot p(s_P|s_0)) \qquad \textbf{[S5]}$$

$$(\text{since } p(s_P|s_0)\le 1)\le\sum_{s_P\in U^C} p(s_P|s_0)\log_2(n)=\log_2(n), \qquad \textbf{[S6]}$$

and thus

$$\text{Cause coefficient}(s_0) \le 1. \quad \textbf{[S7]}$$

The effect information (s_0) is the D_{KL} between the effect repertoire $(S_F|s_0)$ and U^E, the unconstrained effect repertoire. U^E is in general not identical to the uniform distribution. However,

$$p(s_F) = \sum_{s_0 \in U^C} p(s_F|s_0) \cdot p(s_0), \quad \textbf{[S8]}$$

where $p(s_0) = 1/n \; \forall s_0$ and thus:

$$p(s_F|s_0) \le n \cdot p(s_F), \; \forall s_F. \quad \textbf{[S9]}$$

Using Eq. **S9**, if follows that:

$$\text{Effect information}(s_0) = D_{\mathrm{KL}}\left((S_F|s_0), U^E\right) = \sum_{s_F \in U^E} p(s_F|s_0) \log_2\left(\frac{p(s_F|s_0)}{p(s_F)}\right) = \quad \textbf{[S10]}$$

$$(\text{using Eq. } \mathbf{S9}) \le \sum_{s_F \in U^E} p(s_F|s_0) \log_2\left(\frac{n \cdot p(s_F)}{p(s_F)}\right) = \sum_{s_F \in U^E} p(s_F|s_0) \log_2(n) \quad \textbf{[S11]}$$

$$= \log_2(n), \quad \textbf{[S12]}$$

and thus

$$\text{Effect coefficient}(s_0) \le 1. \quad \textbf{[S13]}$$

Finally, because the effect coefficient $(s_0) \in [0 \ldots 1] \, \forall s_0$, also its average over all system states, the state-independent effectiveness $\mathit{Eff}(S) \in [0 \ldots 1]$.

Causal Reduction

To complement the examples of causal emergence in the main text, we here provide an example in which causal reduction is called for. In Fig. S1, a macro mechanism works as an XOR logic gate (as an isolated part of a larger circuit board) with inputs X, Y, and output Z (Fig. S1*A*). At the macro level, the system (XOR,X, Y,Z) generates 2 bits of *EI* over one macro time step T_x (the XOR operates after a "decision" period where it processes the input) and $\mathit{Eff}(S_M) = 0.5$. The macro XOR gate is actually composed of (supervenes upon) nine deterministic micro logic gates (COPY, NOT, AND, OR). In this case, however, causal interactions are stronger at the micro level and over a single micro time step t_x [$EI(S_M) = 7.43$ bits and $\mathit{Eff}(S_M) = 0.83$]. Thus, CE = −5.43 bits, corresponding to negative causal emergence, i.e., reduction. Note that in this case the micro circuit is deterministic and minimally degenerate (0.17), so the macro cannot offset the loss of effective information due to its reduced size by a gain in determinism or a reduction in degeneracy.

To demonstrate this case of causal reduction, we have assumed that a deterministic micro circuit underlies the above macro circuit. In general, however, real digital circuits are often built from many stochastic analog micro elements in a highly degenerate manner, to compensate for noise at the lower level and to create deterministic macro elements. In this way, digital circuits and other engineered systems follow similar design principles as the more physiological examples presented in the main text. Consequently, there is the potential for either causal emergence or reduction in digital circuits, depending on the underlying micro level, just as in physiological systems.

More generally, the notion of causal reduction ($CE < 0$) stands in contrast to previous accounts of reduction that focused on the relationship between scientific theories and whether or not they are reducible to one another (1). In the present account based on causal analysis, the focus is instead on the relationship between micro and macro levels of mechanisms. This account reveals why there is a bias in favor of reductionism in mechanistic scientific explanations. The bias is understandable given that, everything else being equal, the micro would always beat the macro: being more detailed by definition, the micro has an inherent advantage in how informative its causal mechanisms are. This inherent advantage is captured quantitatively in causal analysis because the micro can benefit from both ΔI_{Eff} and ΔI_{Size}, whereas the macro can only gain from ΔI_{Eff}.

1. Nagel E (1961) *The structure of science: problems in the logic of scientific explanation* (Harcourt, Brace & World, New York).

Causal Emergence in a System with Causally Heterogeneous Elements

Although the examples in the main text (with the exception of Fig. 6) all have macro elements with underlying unconnected and causally equivalent micro elements, this is not a necessity for causal emergence. In Fig. S2*A*, the six micro elements are fully interconnected and causally heterogeneous. The elements are structured into two groups {ABC, DEF} due to different intragroup and intergroup mechanisms: within each group, if the sum of intragroup connections Σ(intra) = 0, all elements stay 0 (inactive) the next time step. However, if the sum of intergroup connections Σ(inter) = 3 (synchronous activity from the other group), all elements turn 1, unless they are all 0, in which case they become spontaneously active (1) with probabilities: p(A/D) = 0.45; p(B/ E) = 0.5; p(C/F) = 0.55. Because the micro transition probability matrix (TPM) is noisy, $EI(S_m) = 1.13$ bits and $\mathit{Eff}(S_m) = 0.19$ (Fig. S2*B*). The optimal macro grouping S_M (Fig. S2*C*) has a more deterministic TPM (Fig. S2*D*), $EI(S_M) = 1.84$ bits and $\mathit{Eff}(S_M) = 0.58$. Thus, the macro supersedes the micro [$CE(S)$ = 0.72 bits] despite its reduced repertoire size, because it counteracts noise by responding almost deterministically to synchronous activity over intergroup connections.

The neural-like system of Fig. 6 in the main text has equivalent spatial properties to the example system of Fig. S2 (fully connected, causally heterogeneous elements, sensitive to differences in intraconnections and interconnections). In addition, it has the same temporal properties as the system shown in Fig. 5 (main text), with second-order Markov mechanisms at the micro level. The system's states space at the micro level thus contains 2^{18} states, which prohibited an exhaustive search for the optimal macro level. Nevertheless, the spatiotemporally emergent macro grouping shown in Fig. 6*B* (main text) is assumed to be the optimal macro grouping based on the results obtained from the examples of Fig. S2 and Fig. 5 (main text).

Applicability—Network Motifs as Indicators of Emergence

Measuring *EI* exhaustively, across all micro/macro levels, is not feasible for large systems. This is because, assuming N binary elements, $B_N - 1$ (Nth Bell number) possible groupings of those micro elements into macro elements exist, each of which entails $\prod_{j=1}^{k}(B_{m(j)+1} - 1)$ possible groupings of micro into macro states, where k is the number of macro elements with $m(j)$ micro elements each. The number of *EI* computations to determine the spatiotemporal grain with maximal *EI* thus increases dramatically with N ($N = 1$, 1; $N = 2$, 5; $N = 3$, 27; $N = 4$, 180 computations, etc.) if calculated exhaustively.

In large, complex networks where an exhaustive causal analysis is unfeasible, overrepresented network motifs could already indicate whether the network as a whole is biased toward emergence or reduction. For example, the two most common network motifs shared by the gene networks in *Escherichia coli* and the brain of *Caenorhabditis elegans* are the feedforward loop and the bifan (1). Both these network motifs mimic in their connectivity precisely the micro element groups that made up the optimal (winning) macro elements in our chosen examples. In Fig. 2 (main text), the first spatial example, the macro elements are bifans, whereas in Fig. 6 (main text), the first temporal example, the macro elements are feedforward loops. These are perhaps the simplest possible functionally relevant macro elements. Both the bifan and the feedforward loop show causal convergence (degeneracy) in either space or time. A greater than random prevalence of these or similar network motifs, paired with some amount of intrinsic noise in the system, may indicate that the system operates at a macro level.

1. Milo R, et al. (2002) Network motifs: Simple building blocks of complex networks. *Science* 298(5594):824–827.

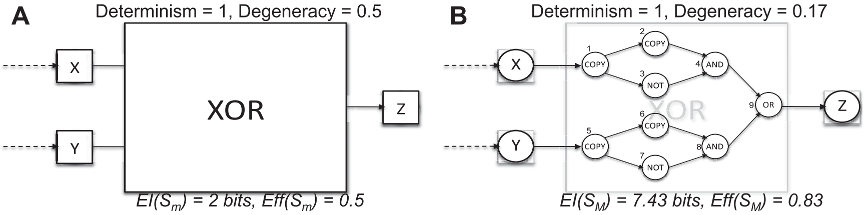

A
Determinism = 1, Degeneracy = 0.5
X
Y
XOR
Z
$EI(S_m)$ = 2 bits, $Eff(S_m)$ = 0.5
B
Determinism = 1, Degeneracy = 0.17
X
Y
1 COPY
2 COPY
3 NOT
4 AND
5 COPY
6 COPY
7 NOT
8 AND
9 OR
XOR
Z
$EI(S_M)$ = 7.43 bits, $Eff(S_M)$ = 0.83

Fig. S1. Causal reduction. (*A*) A part of a larger circuit is presented, which performs a macro XOR logic function over its inputs X, Y, and outputs to Z. (*B*) At the micro level, the XOR consists of nine deterministic logic gates. The system is deterministic at both the micro and the macro level. Moreover, the degeneracy coefficient at the micro level is lower than at the macro level. Therefore, in this case, the micro beats the macro, leading to causal reduction. $CE(S) = -5.43$ bits.

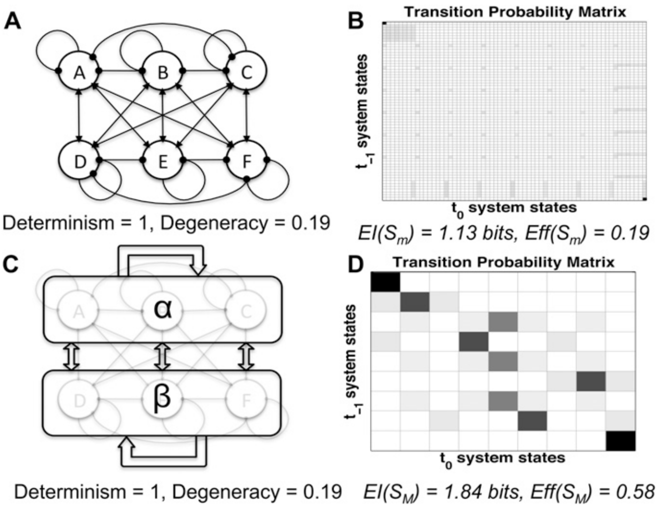

A
A
B
C
D
E
F
Determinism = 1, Degeneracy = 0.19
B
Transition Probability Matrix
t_{-1} system states
t_0 system states
$EI(S_m)$ = 1.13 bits, $Eff(S_m)$ = 0.19
C
A
α
C
D
β
F
Determinism = 1, Degeneracy = 0.19
D
Transition Probability Matrix
t_{-1} system states
t_0 system states
$EI(S_M)$ = 1.84 bits, $Eff(S_M)$ = 0.58

Fig. S2. Causal emergence in a system with differentiated connectivity. (*A*) Micro system S_m with six elements. Regular and rounded arrows indicate intergroup and intragroup connections, respectively. (*B*) Noisy micro-level TPM. (*C*) Macro system S_M. Each macro element receives inputs from itself and the other macro element. (*D*) More deterministic macro-level TPM. $CE(S) = 0.72$ bits.