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:

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

$$= \sum_{s_F \in U^E} p(s_F|s_0)\log_2\left(\frac{p(s_F|s_0)}{p(s_F)}\right) \quad \textbf{[S1]}$$

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

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

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

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

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

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):

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

the degeneracy coefficient the right term:

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

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:

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

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:

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

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

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

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

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

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

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]:

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

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

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 :

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

In detail:

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

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

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

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

$$= MI\big(U^C; U^E\big) = \quad \textbf{[S8]}$$

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

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

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

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

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 ($\in[0 \ldots 1]$) :

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

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

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

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

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

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

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

and thus

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

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:

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

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

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

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

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

and thus

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

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 $Eff(S) \in [0 \ldots 1]$.

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 $Eff(S_M) = 0.5$. 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 $Eff(S_M) = 0.83$]. 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 $Eff(S_m) = 0.19$ (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 $Eff(S_M) = 0.58$. 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.

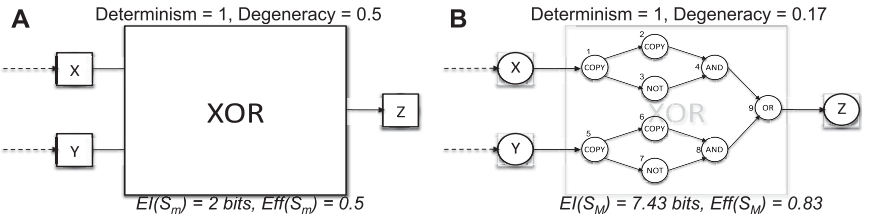

Fig. S1. Causal reduction. (*A*) A part of a larger circuit is presented, which performs a macro XOR logic function over its inputs X, Y, and outputs to Z. (*B*) At the micro level, the XOR consists of nine deterministic logic gates. The system is deterministic at both the micro and the macro level. 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.

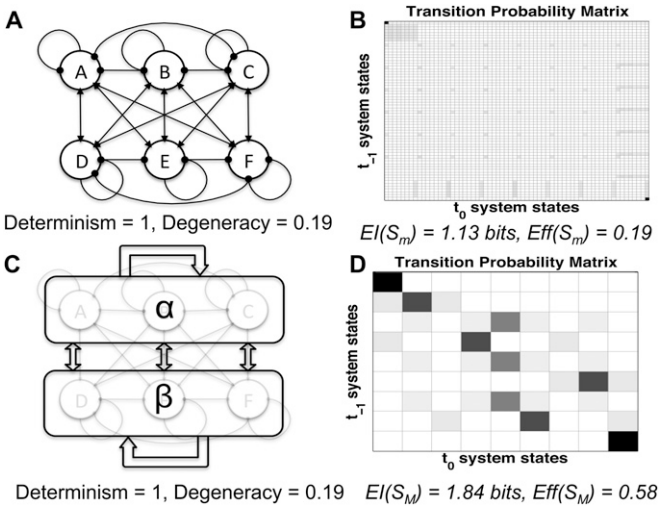

Fig. S2. Causal emergence in a system with differentiated connectivity. (*A*) Micro system S_m with six elements. Regular and rounded arrows indicate intergroup and intragroup connections, respectively. (*B*) Noisy micro-level TPM. (*C*) Macro system S_M. Each macro element receives inputs from itself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