PERSPECTIVE
https://doi.org/10.1038/s41593-021-00911-8

nature neuroscience

Check for updates

Causal reductionism and causal structures

Matteo Grasso[1,2], Larissa Albantakis[1,2], Jonathan P. Lang[1] and Giulio Tononi[1] ✉

Causal reductionism is the widespread assumption that there is no room for additional causes once we have accounted for all elementary mechanisms within a system. Due to its intuitive appeal, causal reductionism is prevalent in neuroscience: once all neurons have been caused to fire or not to fire, it seems that causally there is nothing left to be accounted for. Here, we argue that these reductionist intuitions are based on an implicit, unexamined notion of causation that conflates causation with prediction. By means of a simple model organism, we demonstrate that causal reductionism cannot provide a complete and coherent account of 'what caused what'. To that end, we outline an explicit, operational approach to analyzing causal structures.

Everybody understands that an organism is more than the sum of its parts. Biology offers endless examples of how complex functions 'emerge' from the organization of simple parts: we could not even begin to describe and understand the behavior of living organisms, not to mention their psychology, if we did not refer to emergent levels of organization and function[1,2]. Yet almost everybody also seems to believe that, causally, only the elementary constituents matter[3]. This intuitive belief in 'causal reductionism' is based on reasoning that seems straightforward and unassailable: if we know what causes each element of a system to do what it does individually, we know all we need to know to predict the system's behavior as a whole, so there is no room for additional causes to do anything. To illustrate with a neural example, suppose neurons A and B happen to fire. If we establish the cause of A firing (a 'first-order' occurrence) and that of B firing (another 'first-order' occurrence), why would we bother to establish the cause of A and B firing together (a 'second-order' occurrence)? Obviously, what happened to A and B separately necessarily determines what happened to A and B together[4]. Causally, then, an organism is nothing more than a pack of unit-level mechanisms: if we figure out 'what causes what' at the most elementary level of organization of a system, we do not need to worry about any further causes. As Francis Crick[5] put it, 'You, your joys and your sorrows, your memories and your ambitions, your sense of identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules…You're nothing but a pack of neurons[5].'

The adequacy of reductionist approaches to understand the brain and its relation to behavior has been questioned repeatedly (Box 1). However, when it comes to accounting for actual causes and effects ('what caused what?'), reductionism prevails. We think that this deep-seated, widespread intuition of causal reductionism is fundamentally wrong. The main reason is that reductionism fails to acknowledge the existence of composite mechanisms that are irreducible to their elementary constituents. And if these mechanisms exist, there must be a cause for what they do. As we will see, being blind to the existence of composite mechanisms, reductionism inevitably misses causes and effects that clearly play a role in biology and elsewhere. At heart, the problem with reductionist intuitions is that they rely on an implicit, unexamined notion of causation, which leads to the conflation of causation with prediction (Box 1). Indeed, once the state of each individual neuron is accounted for, there is nothing left to predict, but, as we will see, there can be plenty left to be caused. If we characterize causation explicitly and operationally, the existence of irreducible high-order mechanisms with specific causes and effects becomes as obvious as that of first-order ones, revealing causal structures that are as organized as organisms are[6–8]. (Throughout, the 'order' of a mechanism refers to its number of elementary constituents.)

Here we wish to illustrate this point with an example that demonstrates, as simply as possible, why causal reductionism is incoherent and how causal structures can be analyzed.

A simple example: causal reductionism

The example relies on simple, simulated organisms termed 'frogs' that have adapted to a world inhabited by three kinds of bugs: left-bugs, right-bugs and super-bugs. Left-bugs and right-bugs are small and are preyed upon by frogs as food. Super-bugs are large, being composed of a left-bug and a right-bug fused by the tail, and prey upon frogs. We will consider three species of evolved frogs: F3, F2 and F1. Studying the behavior of some of these frogs shows that they have adapted to this world as follows (Fig. 1): they typically jump forward and left when they detect a left-bug, jump forward and right when they detect a right-bug and jump 'over' to escape any super-bug they detect.

To provide a mechanistic account of the frogs' behavior, we must open their head and study their brain as neuroscientists would do[9].

F3 frogs. In one species of frogs, F3, we find three sensors (S_L, S_C and S_R), three central neurons (C_L, C_C and C_R) and two motors (M_L and M_R; Fig. 2). By stimulating and recording inputs and outputs of the central neurons, we discover that each of them fires preferentially for a different type of bug, which leads to different actions (the neurons' responses are probabilistic, which allows for more flexibility in the frog's behavior). C_L is a left-bug detector on the input side and a jump-left command neuron on the output side: it fires when S_L turns ON while S_C remains OFF, and when it fires, it triggers M_L with high probability (specifically, the probability that M_L fires if $C_LC_C = 10$ is 0.8). C_R is a right-bug detector and a jump-right command neuron: it fires when S_R turns ON with S_C OFF, and when it fires, it triggers M_R with high probability (like in the case of M_L, the probability that M_R fires if $C_CC_R = 01$ is 0.8). Finally, C_C is a super-bug detector: it fires when S_L is ON, S_C is OFF and S_R is ON, and when it fires, it triggers both M_L and M_R (probability (p) = 1), making the frog jump over the super-bug (specifically, the firing of C_C increases the probability of the frog to jump 'over' from 0.8, when $C_LC_CC_R = 101$, to 1, when $C_LC_CC_R = 111$).

[1]Department of Psychiatry, University of Wisconsin-Madison, Madison, WI, USA. [2]These authors contributed equally: Matteo Grasso, Larissa Albantakis. ✉e-mail: gtononi@wisc.edu

Box 1 | Causation, prediction and supervenience

'By choosing to ignore the science to pursue his own selfish goals, he caused a lot of pain to a lot of people.' In this and similar examples, explaining human behavior and cognition calls for high-level causal language—reductionist descriptions simply will not do[2]. This is not just a matter of practicality. Many brain functions are multiply realizable (there are many ways to implement the same function[21–24]) and degenerate (there are many ways to cause the same effect[25,26]), as confirmed by studies showing functional similarities across individuals[27] and species[28]. Even in relatively simple systems, different neural circuits or circuit states produce similar outputs, such as rhythmic oscillations and resulting motor patterns, both across and within individuals[29]. Conversely, the individual constituents of neural networks demonstrate multi-functionality and context dependency[25,26,30]. Indeed, the need for functional[31–34], computational[1,2,35] and dynamical descriptions[26,36,37] of brain processes is now widely recognized. An account based on micro-units, such as individual neurons, taken one by one (first-order), might in principle predict what happens, but will not explain why or allow for meaningful inferences. Nevertheless, the necessity of macro-, high-order descriptions and their use in 'causal' explanations does not dispel the reductionist intuition that only micro-, first-order causes truly matter. It would still seem that, once the state of every neuron in a neural network is accounted for in causal terms, there is nothing else left to be caused[3,4]. This widely shared reductionist intuition about causation seems to be based, ultimately, on two related notions: prediction and supervenience.

Causation and prediction. Prediction has pride of place in science, especially in physics. In fact, the success of dynamical laws at predicting the temporal evolution of physical systems might seem to make notions of causation unnecessary[38], in line with long-standing philosophical skepticism[39–41]. However, the need to distinguish true causation from mere correlation also has a long history[19,42,43] and neuroscientists are regularly warned not to conflate the two. The main reason is that, in biology, we cannot rely on general laws to predict the behavior of biological systems constituted of many heterogeneous parts. To that end, we need to uncover the specific mechanisms governing their interactions through systematic observations and manipulations.

Within neuroscience, causal methods have become an essential part of research, at multiple levels. Electrical stimulation and recording of neural circuits have long been used in neurophysiology[44], complemented now by refined tools such as optogenetic manipulations of specific cell types and even individual cells or synapses[45]. At the system level, transcranial magnetic stimulation and high-density electroencephalogram recordings have proved useful in evaluating the brain's effective connectivity[46]. Network analysis methods such as dynamic causal modeling[47] make explicit reference to causal models based on known anatomical pathways to obtain better inferences about multiregional interactions[48].

Even so, the notion still lingers that what ultimately matters for science is the ability to predict. In fact, an accurate causal model is a powerful tool to predict the state of individual units, such as neurons in a network, from the state of their inputs. If we do so for every unit, no extra work is needed to predict the state of the entire network and any subsets of interest. In other words, first-order prediction is all we need for high-order prediction. This reinforces the reductionist intuition that there are no high-order causes, because they would not add anything to first-order predictions.

The problem, however, is that prediction and causation are ultimately two different notions, and can easily be dissociated, as illustrated in Box 2. In the context of a neural network, prediction can be understood as our ability to derive the future state of a system's units based on knowledge of their past state and of the system's mechanisms ('horizontal determination'). Instead, causation should be understood as the ability of a mechanism to 'take' or 'make' a difference, as demonstrated through observation and manipulation. Interventionist, counterfactual notions of causation have been developed formally and can be applied to any system that can be described by a causal Bayesian network[7,19].

Causation and supervenience. The other leg on which the reductionist intuition rests is the notion of supervenience[49]—the assumption that once the state of all micro-, first-order units is fixed (at a given time step), the state of all macro-, high-order subsets is fixed, too (at the same time step). If micro-, first-order causes are sufficient to fix the state of all micro-, first-order units, and if these in turn are sufficient to fix the state of macro-units and high-order subsets in accordance with supervenience, then, goes the intuition, the latter cannot provide any additional causation[50].

Once again, the problem with this reasoning is that supervenience is different from causation[51]. Supervenience can be understood as our ability to derive properties of macro-units or high-order subsets of units from those of micro-, first-order units ('vertical determination'). As already stated, causation should be understood as the ability to 'make' (or 'take') a difference. Whether a high-order mechanism can make (or take) a difference in a way that is irreducible to the difference made (or taken) by its parts should be assessed through causal structure analysis[7] rather than ruled out based on unexamined intuitions. As illustrated in the main text, the fact that the state of high-order subsets is fixed once that of the first-order units is fixed, is orthogonal to the question of whether they have irreducible causes or effects. In other words, causation can be compositional, as long as it is irreducible, even though compositional mechanisms supervene on first-order ones. As mentioned in Box 3, a similar demonstration of causal irreducibility can be provided for macro-units, such as groups of neurons or mini-columns, with respect to macro-causes and effects. In sum, causation is different from prediction and supervenience, it addresses a separate question ('what caused what?'), and requires its own formal and operational characterization[7,19].

In this simple case, we seem to have a straightforward, mechanistic account of the frog's behavior. Each central neuron is a first-order mechanism that functions as a specialized bug detector and triggers the appropriate motor response. Moreover, for each neuron, we can tell what caused it to fire (or not) and what effect its firing had. Finally, this first-order causal account seems complete: once we have accounted for the cause of each individual neuron firing (or not), there is nothing else that can or needs to be caused. Indeed, once we can predict what each neuron will do, we can predict what the frog will do. Causal reductionism thus seems to work just fine, strengthening the intuition that, as long as we can figure out the causes and effects of each individual neuron, we can have a complete causal account and a perfect prediction.

F2 frogs. But now consider a second species of frog, F2, whose behavior is the same as that of F3 frogs, but which turns out to have a slightly different brain (Fig. 2). In fact, F2 frogs have a more efficient brain with just two central neurons (C_L and C_R),

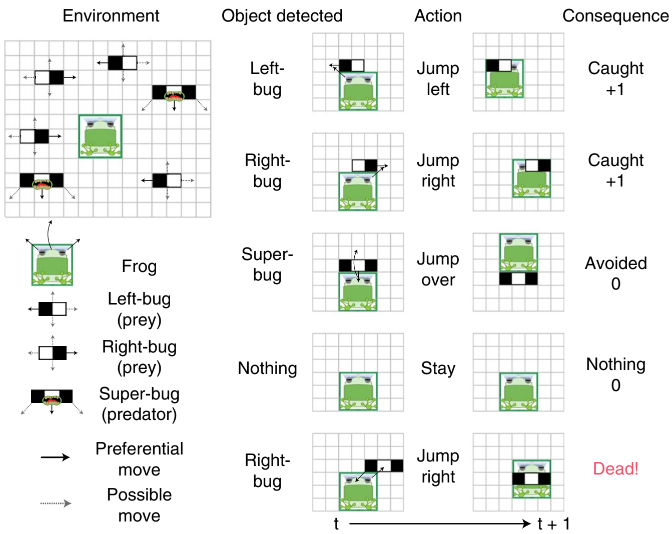


Fig. 1 | Simulated environment and frog behavior. The frog is indicated as a 3 × 3 green square, left-bugs and right-bugs (prey) are indicated as 2 × 1 segments, super-bugs (predators) are indicated as 3 × 1 segments; the possible actions of each type of bug are indicated in the bottom-left quadrant; on the right is a description of possible scenarios in which the frog catches small bugs, avoids super-bugs, stays still or is captured by a super-bug.

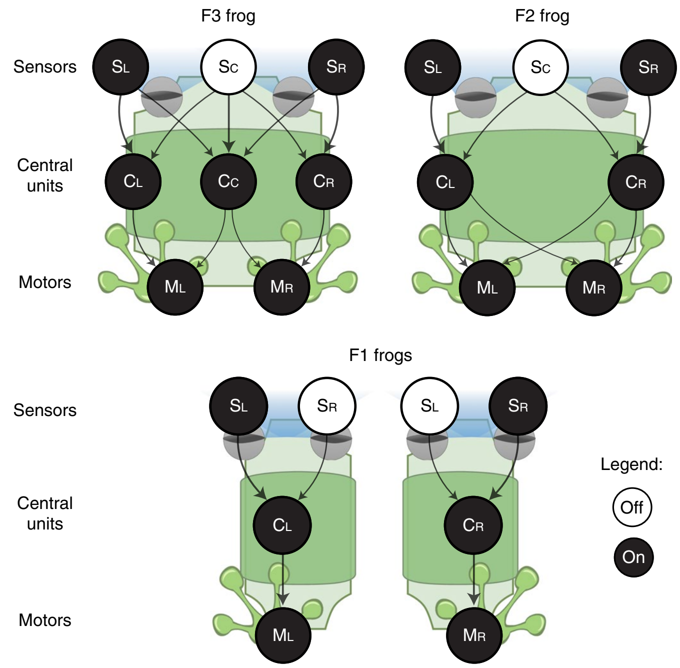


Fig. 2 | Wiring of F3 and F2 frogs and of a pair of F1 frogs. F3, F2 and F1 frogs differ in their internal wiring. The activation pattern indicated by black (ON) and white (OFF) nodes corresponds to the most likely system state upon encountering a super-bug.

which provides them with some evolutionary advantage[10]. C_L is still a left-bug detector that triggers a left jump, and C_R is still a right-bug detector that triggers a right jump. However, F2 frogs have managed to get rid of C_C, the super-bug detector, while preserving the same super-bug detection and avoidance functions: what happens is that a super-bug triggers the firing of both C_L and C_R, which in turn trigger the firing of both M_L and M_R, making the frog jump over with $p = 1$.

There is surely no mystery as to how the brain of the F2 frog works (a typical explanation would be that C_L and C_R together provide a 'distributed representation' of a super-bug). But how does causal reductionism treat F2 frogs? As with F3 frogs, the firing of neuron C_L is caused by a left-bug and that of C_R by a right-bug. When a super-bug appears, the input of C_L, corresponding to the left side of the super-bug, causes it to fire, just as it would if it were detecting a left-bug. Similarly, the input of C_R, corresponding to the right side of the super-bug, causes it to fire as if it were detecting a right-bug. Knowing what C_L and C_R do, we can also predict with perfect accuracy that M_L and M_R will be triggered, making the frog jump over, just as we can predict the behavior of the F3 frog based on C_L, C_C and C_R.

However, if we follow causal reductionism, there is a crucial difference in causal terms between F2 and F3 frogs: the super-bug as such never shows up as a cause in F2 frogs, because both C_L and C_R have already been caused by their separate inputs, and reductionism does not consider high-order mechanisms such as C_LC_R. This is unlike F3 frogs, where the super-bug as a whole was the cause of C_C firing. Thus, a first-order, reductionist causal account sees the super-bug as a cause (mediated by S_L, S_C and S_R) in F3 frogs, but excludes it as a cause in F2 frogs. The assumption is that once all individual neurons are causally accounted for, there is nothing else left to be caused. This inability to see C_LC_R in F2 frogs as a high-order mechanism—a super-bug detector—and consequently to see the super-bug as a cause now seems absurd, because these frogs can obviously detect and respond to super-bugs just as well as F3 frogs. While causal reductionism provides an irreducible causal account for avoiding super-bugs in F3 frogs, the avoidance behavior of F2 frogs is instead 'explained away'. Yet, both F3 and F2 frogs have evolved a mechanism to detect and avoid super-bugs. Perhaps something is not right, then, with the way causal reductionism conceives of causation.

F1 frogs. To drive the point home from a different angle, consider yet another species, F1 frogs. These are frogs reduced to their bare minimum—essentially 'half-frogs' with just two sensors and a single central neuron. F1 frogs come in two varieties: left-F1 frogs detect left-bugs and jump forward and left to catch them, while right-F1 frogs do the same for right-bugs. Neither type of F1 frog can detect super-bugs, which is bad, but we assume here that as a species they manage to survive by sheer numerosity.

Now consider what happens, in mechanistic terms, if by chance two F1 frogs, of the left and right variety, find themselves side by side in front of a super-bug (Fig. 2; for a detailed description of the behavior of F1 frogs, see the figures in Supplementary Note 1). The C_L neuron in the left-F1 frog will detect a left-bug (corresponding to the left side of the super-bug), while the C_R neuron in the right-F1 frog will detect a right-bug (corresponding to the right side of the super-bug). Because both M_L and M_R will be activated, both varieties of F1 frogs will leap sideways and effectively avoid the super-bug.

In this instance, a reductionist account would find the cause of C_L firing (which in turn triggers M_L), the cause of C_R firing (triggering M_R) and no other cause. In the case of two F1 frogs, then, the reductionist account happens to capture everything: there are two separate causes leading two separate F1 frogs to a fortuitous escape. And it is intuitively obvious that there cannot be any other cause, because F1 frogs conspicuously lack anything resembling a super-bug detector. The problem is that causal reductionism cannot distinguish this case from that of an F2 frog: unlike the two F1 frogs, F2 frogs have evolved an efficient, second-order mechanism, C_LC_R, whose activation has a clear cause, the detection of a super-bug, and a clear effect, the escape response. In sum, causal reductionism is blind to the obvious causal similarity between F3 and F2 frogs, both of which possess a super-bug detector—first-order for F3 and

Box 2 | Dissociation between causation and prediction: an example

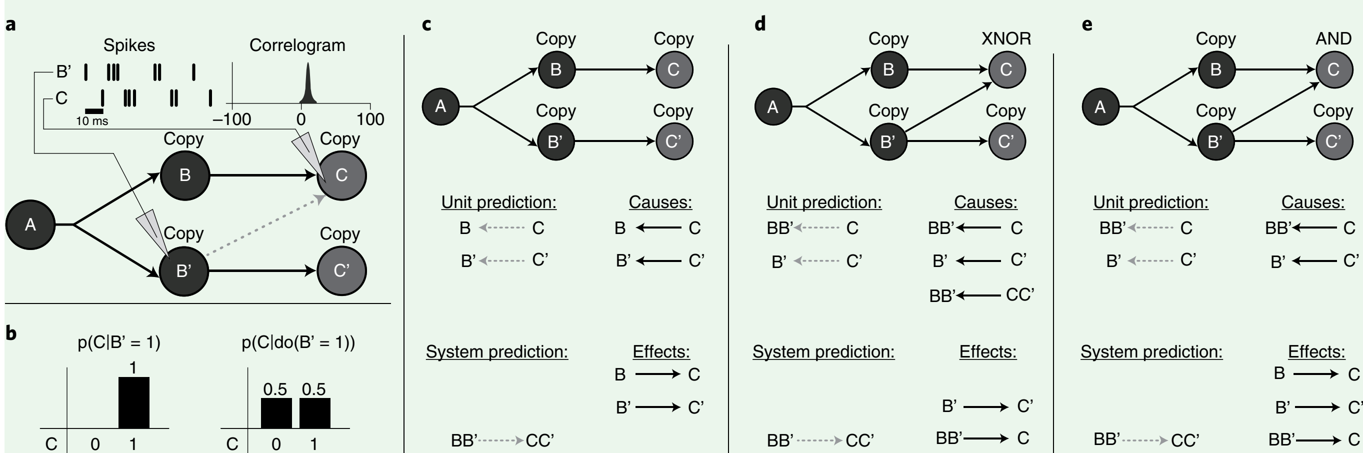
In the box figure, **a** shows a schematic 'neural' example in which, if we recorded the state of neuron B' (firing or not), we would unfailingly predict the ensuing state of neuron C (gray dotted arrow). However, an intervention such as optogenetic stimulation (indicated by the 'do-operator' do(B' = 1) in **b**) would reveal that changing the state of neuron B' does not affect the state of unit C, which is only affected by perturbations of neuron B. Furthermore, a physical partition (for example, severing an anatomical connection) between B and C would abolish this effect, but one between B' and C would not. In other words, unlike prediction, causation must be evaluated by physical interventions (perturbations and partitions). A causal model, represented by directed arrows, makes explicit the difference between prediction and causation. It also enables better predictions when circumstances change. Thus, we can predict that C will continue to be affected by A if B' is inactivated, but not if B is inactivated.

The dissociation between prediction and causation is also relevant when it comes to high-order causation of the kind described in the main text. In **c**–**e** (top row), the same causal model is illustrated with different activation functions for neuron C (copy, XNOR or AND). As argued above, if we can predict the next state of every individual neuron based on the state of its inputs (first-order, unit prediction), we also predict higher-order states (system prediction). In the example, predicting the state of neurons C and C' individually also predicts the state of the set CC' in all three cases. However, this is not true for causation. Causal structure analysis shows that, if C and C' are copy units (**c**), C = 1 has cause B = 1 and C' = 1 has cause B' = 1, but there is no irreducible cause for CC' = 11.

By contrast (**d**), if C is an XNOR unit (it fires if its inputs are equal—00 or 11), then C = 1 has its cause in BB' = 11, C' = 1 has its cause in B' = 1, and CC' = 11 also has an irreducible cause in BB' = 11. This is because CC' = 11 requires BB' to have been in the specific state 11, while C = 1 could also have been caused by BB' = 00 and C' = 1 alone only requires B' to have fired.

If C is an AND unit (**e**), C also has its cause in BB' = 11 and C' has its cause in B' = 1. In this case, however, despite the joint input to C from B and B', CC' = 11 has no irreducible cause, because CC' does not 'take a difference' in BB' that is not already taken by C.

The example in **d** also shows that, in general, it is not possible to predict the next state of a neuron or set of neurons based on the output of each individual input unit, for example, due to nonlinear interactions (individual inputs to an XNOR have no predictive information about its next state). On the other hand, with a causal model, we can achieve 'holistic' system prediction on the output side (**c**–**e**, bottom row): based on the output of BB' as a whole (= 11), we can predict the next state of CC' (= 11). In doing so, we also predict the next state of each individual unit. However, like unit prediction, system prediction does not capture the structure (composition) of effects: BB' has an irreducible effect on C in **d** and **e**, but not in **c**. Thus, both unit and system prediction fail to capture the structure (composition) of irreducible causes and effects. In summary, to demonstrate causation, we need to show that something takes or makes a difference, as assessed through perturbations and partitions. Unlike prediction, high-order causation must be assessed in its own right and does not automatically follow from first-order causation.



second-order for F2. It is also blind to the causal difference between F2 frogs, which have evolved a super-bug detector, and the pair of F1 frogs, which obviously have not.

The essential point made by these simple examples is this: causal reductionism rightly recognizes the importance of causal irreducibility, but it does so implicitly and intuitively rather than explicitly and operationally. As shown by F3 frogs, it is reasonable to assume the irreducibility of first-order mechanisms constituted of individual units, such as the F3 super-bug detector C_C. As illustrated by F1 frogs, it is also reasonable to assume that a high-order mechanism is not a mechanism at all if it is fully reducible to two first-order mechanisms (for example, C_LC_R in pairs of F1 frogs). However, this implicit, intuitive approach breaks down in F2 frogs: causal reductionism fails to recognize the super-bug detector C_LC_R as an irreducible high-order mechanism with a cause in its own right.

When it comes to high-order mechanisms constituted of two or more units, then, causal reductionism becomes incoherent with respect to irreducibility and ends up missing obvious causes. (Similar problems arise for 'causal holism', which would exclusively consider the dynamics of the system as a whole and ignore causal structure, just like causal reductionism[11]; Box 2.)

Causal structures

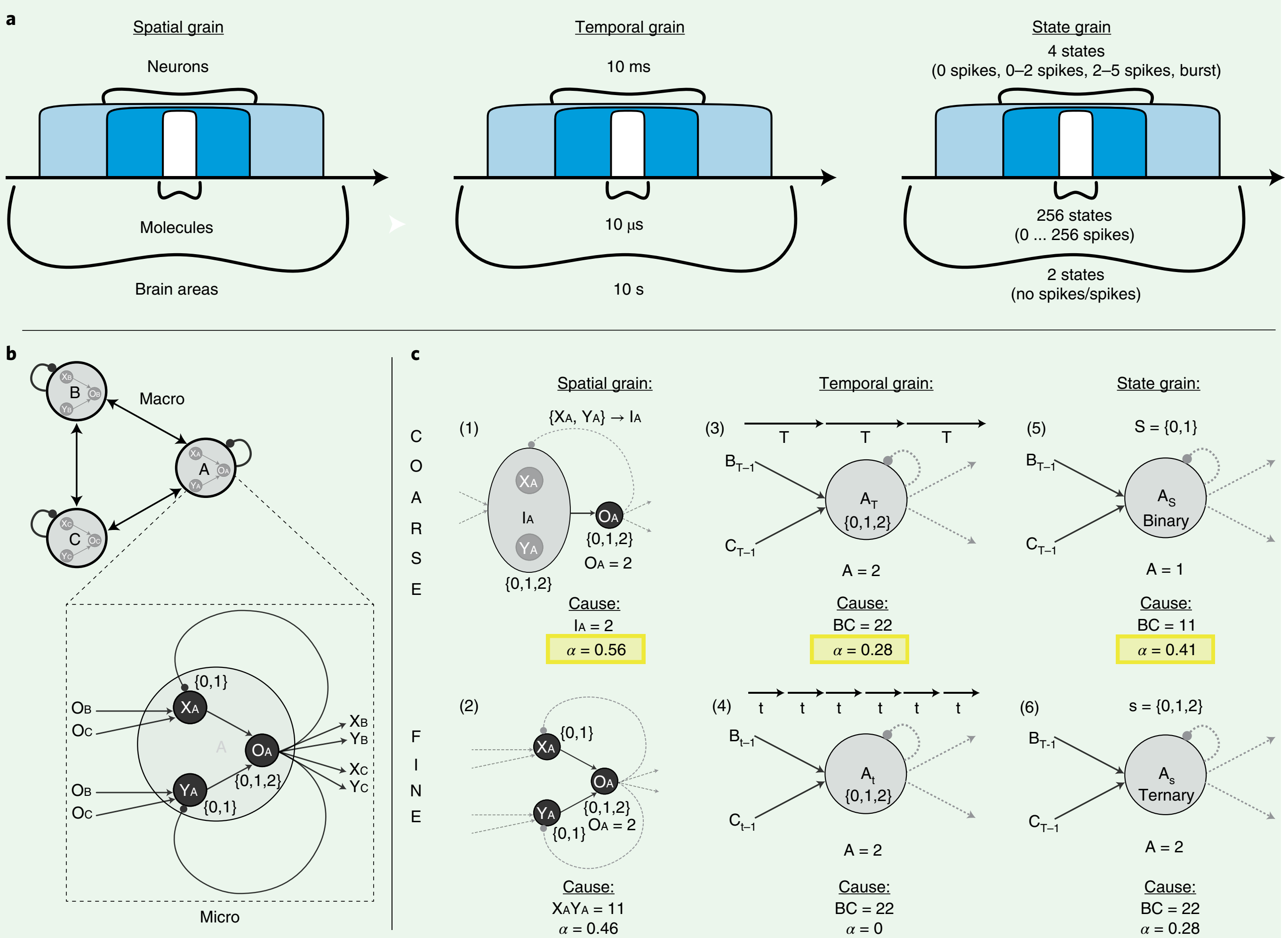
How can we move beyond the seemingly straightforward but ultimately incoherent reductionist notion of causation? What is needed, we believe, is an explicit and operational method to characterize

Box 3 | Macro and micro: causation at different levels of organization

To understand the workings of biological systems, including the brain, we rely on models at multiple grains of description (Box 1). For example, we may consider models based on neurons, finer grains such as neuronal compartments, down to membrane patches, or coarser grains such as mini-columns, macro-columns, brain regions, and so on. We may consider their interactions with a resolution of a few milliseconds, hundreds of milliseconds or seconds. And we may consider neural states at finer or coarser spatial-graining, temporal-graining and state-graining, such as binary subdivisions into firing or not, or finer subdivisions into many discrete levels of firing (**a**).

Macro-units, macro time steps and macro-states can provide a useful way of grouping micro-units, time steps and states to obtain a simpler, more essential description of the system's dynamics[52]. Nevertheless, it is typically assumed, in line with the reductionist intuition and the conflation of causation with prediction and supervenience (Box 1), that all the causal work is done by the micro-units that constitute the system. And indeed, leaving aside practical concerns, analyzing a system at a macro-grain cannot provide better predictions than analyzing it at the micro-grain. Moreover, because the macro-grain supervenes on the micro-grain, every macro-property can be derived from micro-properties.

However, it can be shown that, in causal terms, macro can beat micro[53]. To illustrate this, we analyzed a small example system constituted of three interacting macro-units (corresponding, say, to mini-columns; **b**; for a full description of this system, see Supplementary Note 3). Each of these macro-units is taken as a 'black box'[54,55], constituted at the micro level of two input units (X and Y) with two possible states {0, 1} each, and one output unit (O) with three possible states. The macro-units are self-inhibitory due to inhibitory connections from O to X and Y in each black box; all other connections are excitatory. Based on the formalism of causal structure analysis[7], we evaluated the causal strength (α) with which the black box A or its output element O_A specify their respective cause. We compared coarser spatial, temporal and state grains to finer-grained characterizations of the same system (**c**). With respect to spatial grain (**c1** and **c2**), the two input units X and Y are coarse-grained into one macro-unit (I) with three possible states (corresponding to the sum of X and Y). At this coarser spatial grain, the output unit O_A specifies its cause, the macro-unit $I_A=2$, with higher causal strength (α) than O_A specifies its micro inputs ($X_AY_A=11$) at the micro level. With respect to temporal grain (**c3** and **c4**), the black box A with three possible states per element, evaluated within the macro system ABC, specifies its cause (BC = 22) at the macro time step (corresponding to two micro time steps), but not the at the micro time step ($\alpha=0$). For the state grain, **c5** and **c6** show that the black boxes A, B and C can also be considered as binary rather than ternary units (for example, with states $\{0,1\}_s$ grouped into macro state $\{0\}_S$ and state $\{2\}_s$ into macro state $\{1\}_S$). In this case, the binary unit A = 1 specifies its cause BC = 11 with higher causal strength than the ternary macro-units.



Box 3 | Macro and micro: causation at different levels of organization (continued)

In summary, causal structure analysis reveals that, in the maximally activated state: (i) a stronger cause can be found at the grain of macro-units, compared to micro-units (cause (1) > (2)); (ii) a macro time step can also result in higher causal strength compared to micro-intervals (cause of (3) > (4)); and (iii) reducing the number of states can increase causal strength under certain conditions (cause of (5) > (6)). A quantitative, operational approach to causal analysis that can be applied across spatiotemporal scales can thus identify those macro levels that are particularly relevant for our understanding of a system. By contrast, a reductionist account cannot explain why some spatiotemporal scales seem causally more relevant than others.

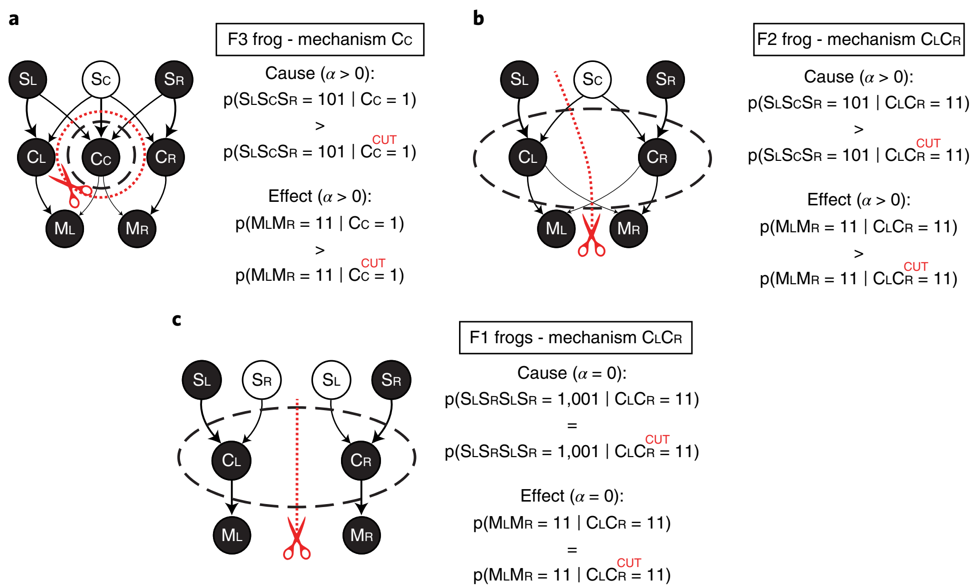


Fig. 3 | Irreducibility analysis for selected mechanisms. For each type of frog, the irreducibility analysis is illustrated for the mechanism intuitively playing the role of the 'super-bug detector': C_C in F3 frogs (**a**), C_LC_R in F2 frogs (**b**) and also C_LC_R in pairs of F1 frogs (**c**).

causal structures. This is possible through a complete and coherent causal analysis, grounded in a mechanistic model of a system, that relies on explicit causal principles. These principles—realization, composition, information, integration and exclusion—are based on the postulates of cause–effect power of integrated information theory (IIT)[12,13].

Among the principles informing such causal analysis, composition, integration and exclusion are especially important for our purposes. The composition principle says that causes and effects are structured: a mechanism can be constituted of any number of units in a state: a first-order mechanism may be constituted of a single neuron that fired, with its own cause and effect; a second-order mechanism may be constituted by a pair of neurons, one that fired and one that did not, again with its own cause and effect, and so on. Causes and effects of separate mechanisms are causally related whenever they overlap over some units[14]. The integration principle says that mechanisms must be irreducible: a mechanism in a state has an actual cause if partitioning it into two or more sub-mechanisms in any way makes the cause less probable (same on the effect side; Fig. 3). The irreducibility measure, α, quantifies how much of a difference a partition makes to the probabilities of causes and effects (Fig. 3) and is assessed by considering the minimum partition of the mechanism[7]. The exclusion principle says that every irreducible mechanism has just one cause and one effect—the one that is maximally irreducible (having the highest α value). Causal reductionism also endorses causal exclusion: if something is fully accounted for causally, then it cannot have further causes. However, the way exclusion is conceived in causal reductionism typically rules out high-order mechanisms, as we saw with F2 frogs. Instead, the causal analysis used here endorses high-order mechanisms (composition), as long as they are irreducible to sub-mechanisms (integration). In other words, irreducible mechanisms do not exclude each other, but causes and effects do: each mechanism must have just one cause and one effect (exclusion).

A complete causal analysis yields a 'causal structure': the set of all causes, effects and causal relations specified by the irreducible mechanisms constituted by a set of units in a particular state[7]. Such a complete analysis, which we call 'causal structure analysis'[7–11], is formulated in probabilistic terms[15,16] and requires perturbing subsets of elements in every possible way, observing their effects on other system elements[17,18], and establishing the consequences of partitions among the elements. This analysis is applicable to any system that can be described by a causal Bayesian network[19] with a discrete number of nodes and finite, discrete states (not restricted to binary variables[7,20]). Here we apply causal structure analysis at a given level of organization. As illustrated in Box 3, the proposed causal formalism also applies across the various levels of organization of a system.

How does causal structure analysis fare with the three frog examples? The fully unfolded causal structure of these three examples is presented in Supplementary Note 2. For the present purposes, the main point is the following: causal structure analysis establishes that both F3 and F2 frogs have irreducible mechanisms for the detection and avoidance of super-bugs, with corresponding causes and effects, whereas pairs of F1 frogs do not.

Specifically, with F3 frogs (Fig. 3a), the analysis identifies three irreducible, first-order mechanisms (C_L, C_C and C_R). For $C_C = 1$, the cause is the sensor state $S_LS_CS_R = 101$, corresponding to the super-bug. Similarly, the effect is the motor state $M_LM_R = 11$, corresponding to jumping 'over'. In this case, then, the cause and effect of

C_C (and the other two first-order mechanisms) turn out just as the causal reductionist would expect. (Note that testing for the irreducibility of a first-order mechanism, such as C_C in F3, requires partitioning all its input and output connections. Partitioning a single element thus amounts to 'disintegrating it,' which captures the intuition that a first-order mechanism cannot be reduced further.)

Not so, however, with frog F2 (Fig. 3b). In this case, causal structure analysis reveals that any partition of C_LC_R into independent mechanisms makes a difference ($\alpha > 0$). Therefore, C_L and C_R together constitute an irreducible second-order mechanism C_LC_R. As in F3 frogs, the cause of $C_LC_R = 11$ turns out to be the sensor state $S_LS_CS_R = 101$ (the super-bug), and the effect is the motor state $M_LM_R = 11$ (jumping 'over'). Unlike causal reductionism, then, causal structure analysis identifies C_LC_R as a second-order mechanism that detects super-bugs and triggers the escape action, which fulfills the same causal requirement of irreducibility as C_C in F3 frogs (in this view, a 'distributed representation' only matters causally if it corresponds to a high-order mechanism that specifies a cause and an effect). Causal structure analysis also identifies causal relations among overlapping causes and effects of second-order and first-order mechanisms, capturing the fact that a super-bug is composed of a left-bug and right-bug bound together (Supplementary Note 2).

Finally, in the case of the two F1 frogs, causal structure analysis easily reveals that C_L in the left-F1 frog and C_R in the right-F1 frog do not constitute a second-order mechanism because they are causally independent (Fig. 3c): a partition between C_L and C_R does not make any difference ($\alpha = 0$), consistent with the absence of any super-bug detection and escape mechanism—not to mention of any full frog.

Conclusion

In neuroscience, as elsewhere, causal reductionism seems at first both intuitive and inescapable. Once all neurons have been caused to fire or not to fire, there seems to be no room for further causation. Moreover, there would seem to be no need for it: if we know what causes each neuron to fire and the current state of the brain, we can predict what the brain will do next without fail (leaving aside indeterminism). And yet, as we have seen using a simple example that can be fully characterized in mechanistic terms, causal reductionism misses out on causes and effects that clearly are important, both conceptually and biologically. For all its intuitive appeal, reductionism lacks a principled, explicit approach to analyzing causal structures. It assumes that first-order mechanisms are causally irreducible but fails to recognize that higher-order mechanisms can be just as irreducible, having their own irreducible cause and effect[11]. In doing so, reductionism conflates causation with prediction (Boxes 1 and 2). Knowledge of first-order mechanisms is indeed enough to predict everything about the dynamics of a system. But only the analysis of causal structures can provide a coherent account of 'what caused what'.

Code availability

The code used for the simulations can be accessed freely at https://github.com/wmayner/pyphi/blob/develop/pyphi/examples.py/.

Received: 20 September 2020; Accepted: 15 July 2021;
Published online: 23 September 2021

References

1. Marr, D. *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 1982).
2. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
3. Bickle, J. Marr and reductionism. *Top. Cogn. Sci.* **7**, 299–311 (2015).
4. Kim, J. *Mind in a Physical World: an Essay on the Mind–Body Problem and Mental Causation* (MIT press, 1998).
Classical philosophical work introducing the causal exclusion argument and employing it in the context of reductive physicalism.
5. Crick, F. *The Astonishing Hypothesis* (Scribner's, New York, 1994).
An explicit endorsement of causal reductionism in the neuroscience literature. Strictly speaking, Crick was making an ontological statement in addition to a causal statement.
6. Albantakis, L. & Tononi, G. The intrinsic cause–effect power of discrete dynamical systems—from elementary cellular automata to adapting animats. *Entropy* **17**, 5472–5502 (2015).
7. Albantakis, L., Marshall, W., Hoel, E. & Tononi, G. What caused what? a quantitative account of actual causation using dynamical causal networks. *Entropy* **21**, 459 (2019).
Formal exposition of causal structure analysis, which is based on an interventional, counterfactual notion of causation. Rather than testing a single counterfactual, causal structure analysis takes into account all possible counterfactuals (system states), allowing for a probabilistic formulation. Further differences with other approaches to actual causation are also discussed, including the distinction between cause and effect, composition, integration and exclusion.
8. Juel, B. E., Comolatti, R., Tononi, G. & Albantakis, L. When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. in *Proceedings of the ALIFE 2019: The 2019 Conference on Artificial Life*, 477–484 (MIT Press, 2019).
9. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. & Pitts, W. H. What the frog's eye tells the frog's brain. *Proc. IRE* **47**, 1940–1951 (1959).
10. Albantakis, L., Hintze, A., Koch, C., Adami, C. & Tononi, G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput. Biol.* **10**, e1003966 (2014).
11. Albantakis, L. & Tononi, G. Causal composition: structural differences among dynamically equivalent systems. *Entropy* **21**, 989 (2019).
12. Oizumi, M., Albantakis, L. & Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput. Biol.* **10**, e1003588 (2014).
13. Tononi, G., Boly, M., Massimini, M. & Koch, C. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* **17**, 450–461 (2016).
The IIT formalism establishes whether a system qualifies as an intrinsic entity—a maximum of intrinsic, structured, specific, irreducible cause–effect power—which is required for a complete account of causation, since only what exists can cause. The IIT analysis of cause–effect power examines potential causes and effects from the intrinsic perspective of a system in a single state (potential causation). By contrast, causal structure analysis examines what actually caused what based on a sequence of states that have happened (actual causation). It should be noted that in this paper we do not consider whether our example systems qualify as intrinsic entities and what that would imply for causation. Instead, we have attempted to illustrate the incoherence of causal reductionism purely within a biological and functional framework.
14. Haun, A. & Tononi, G. Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* **21**, 1160 (2019).
15. Ay, N. & Polani, D. Information flows in causal networks. *Adv. Complex Syst.* **11**, 17–41 (2008).
16. Janzing, D., Balduzzi, D., Grosse-Wentrup, M. & Schölkopf, B. Quantifying causal influences. *Ann. Stat.* **41**, 2324–2358 (2013).
17. Halpern, J. Y. & Pearl, J. Causes and explanations: a structural-model approach. Part I: causes. *Br. J. Philos. Sci.* **56**, 843–887 (2005).
Halpern and Pearl's account is currently the most established approach to actual causation. Unlike causal structure analysis, it does not evaluate causal strength. Instead, it aims to provide a set of contingency conditions under which a simple, counterfactual test may be applied to identify variables that are causally relevant for the occurrence of a particular event.
18. Halpern, J. Y. *Actual Causality* (MIT Press, 2016).
19. Pearl, J. *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2000).
Seminal contribution introducing a causal calculus—a formal framework to evaluate interventions in causal Bayesian networks. The book also offers an overview over methods for 'causal inference'—how to define a causal model from sparse data. While causal structure analysis makes use of interventions and causal Bayesian networks, it is not concerned with causal inference.
20. Gomez, J. D., Mayner, W. G. P., Beheler-Amass, M., Tononi, G. & Albantakis, L. Computing Integrated Information (Φ) in discrete dynamical systems with multi-valued elements. *Entropy* **23**, 6 (2020).
21. Putnam, H. Psychological Predicates. in *Art, Mind and Religion* (eds. W. Capitan & D. Merrill) 37–48 (University of Pittsburgh Press, 1967).
22. Sober, E. The multiple realizability argument against reductionism. *Philos. Sci.* **66**, 542–564 (1999).

23. Aizawa, K. Neuroscience and multiple realization: a reply to Bechtel and Mundale. *Synthese* **167**, 493–510 (2009).
24. Aizawa, K. Multiple realizability by compensatory differences. *Eur. J. Philos. Sci.* **3**, 69–86 (2013).
25. Tononi, G., Sporns, O. & Edelman, G. M. Measures of degeneracy and redundancy in biological networks. *Proc. Natl Acad. Sci. USA* **96**, 3257–3262 (1999).
26. Kelso, J. S. Multistability and metastability: understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 906–918 (2012).
27. Brennan, C. & Proekt, A. A quantitative model of conserved macroscopic dynamics predicts future motor commands. *Elife* **8**, e46814 (2019).
28. Hemberger, M., Pammer, L. & Laurent, G. Comparative approaches to cortical microcircuits. *Curr. Opin. Neurobiol.* **41**, 24–30 (2016).
29. Marder, E., Goeritz, M. L. & Otopalik, A. G. Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Curr. Opin. Neurobiol.* **31**, 156–163 (2015).
30. McIntosh, A. R. Contexts and catalysts. *Neuroinformatics* **2**, 175–181 (2004).
31. Pouget, A., Dayan, P. & Zemel, R. Information processing with population codes. *Nat. Rev. Neurosci.* **1**, 125–132 (2000).
32. Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
33. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).
34. Panzeri, S., Macke, J. H., Gross, J. & Kayser, C. Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* **19**, 162–172 (2015).
35. Frégnac, Y. Big data and the industrialization of neuroscience: a safe roadmap for understanding the brain? *Science* **358**, 470–477 (2017).
36. Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M. & Friston, K. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* **4**, e1000092 (2008).
37. Beer, R. D. Beyond control: the dynamics of brain–body–environment interaction in motor systems. *Adv. Exp. Med. Biol.* **629**, 7–24 (2009).
38. Norton, J. D. Causation as folk science. *Philosophers' Imprint* **3**, 1–22 (2003).
39. Hume, D. *An Enquiry Concerning Human Understanding* (Clarendon Press, 2000). 1748.
40. Russell, B. On the notion of cause. *Proc. Aristotelian Soc.* **13**, 1–26 (1913).
41. Lewis, D. K. *On the Plurality of Worlds*. (Blackwell, 1986).
42. Chicharro, D. & Ledberg, A. When two become one: the limits of causality analysis of brain dynamics. *PLoS ONE* **7**, e32466 (2012).
43. James, R. G., Barnett, N. & Crutchfield, J. P. Information flows? a critique of transfer entropies. *Phys. Rev. Lett.* **116**, 238701 (2016).
44. Selimbeyoglu, A. & Parvizi, J. Electrical stimulation of the human brain: perceptual and behavioral phenomena reported in the old and new literature. *Front. Hum. Neurosci.* **4**, 46 (2010).
45. Zhang, S. et al. Selective attention. Long-range and local circuits for top-down modulation of visual cortex processing. *Science* **345**, 660–665 (2014).
46. Massimini, M., et al. Breakdown of cortical effective connectivity during sleep. *Science* **309**, 2228–2232 (2005).
47. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
48. Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J. & Friston, K. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* **58**, 339–361 (2011).
49. Davidson, D. Mental events. In *Readings in Philosophy of Psychology* (ed. Block, N.) 107–119 (Cambridge, Harvard University Press, 1980).
50. Kim, J. *Physicalism, or Something Near Enough* (Princeton University Press, 2005).
51. Kim, J. Supervenience and supervenient causation. *South. J. Philos.* **22**, 45–56 (1983).
52. Kelso, J. A. Synergies: atoms of brain and behavior. *Adv. Exp. Med. Biol.* **629**, 83–91 (2009).
53. Hoel, E. P., Albantakis, L., Marshall, W. & Tononi, G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* **1**, niw012 (2016).
54. Marshall, W., Albantakis, L. & Tononi, G. Black-boxing and cause-effect power. *PLoS Comput. Biol.* **14**, e1006114 (2018).
55. Albantakis, L., Massari, F., Beheler-Amass, M. & Tononi, G. A macro agent and its actions. Preprint at https://arxiv.org/abs/2004.00058 (2020).

Acknowledgements
We thank M. Boly, A. Cattani, F. Ellia, G. Findlay, B. Juel, W. Marshall, W. Mayner, G. Mindt and R. Verhagen, and especially J. Hendren, for their comments on the manuscript. This project was made possible through support from Templeton World Charity Foundation (nos. TWCF0216 and TWCF0526) and by The Tiny Blue Dot Foundation (UW 133AAG3451). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation and The Tiny Blue Dot Foundation.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00911-8.

Correspondence should be addressed to Giulio Tononi.

Peer review information *Nature Neuroscience* thanks Viktor Jirsa and Klaas Stephan for their contribution to the peer review of this work.