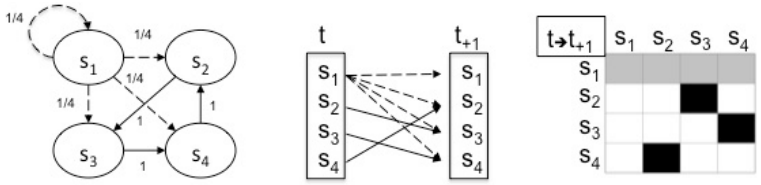
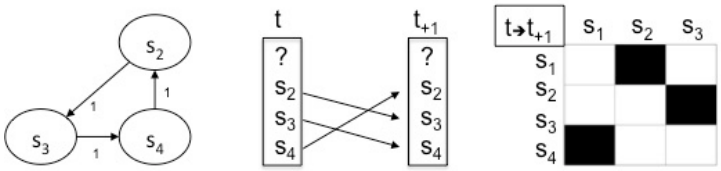
A terminological note: the theory uses the term "macroscale" as an umbrella term for various changes in the description of a system that involve some sort of dimensional reduction. This includes things like black boxing (allowing states to vary under the effects of interventions but not including them in the model), leaving states/elements totally exogenous to the model, and also setting particular initial states or boundary conditions. None of these are the same as the most detailed, full, and microscopic possible model of the system (the microscale, or, the territory). Leaving anything out, just like a map does, counts as a macroscale, especially because similar math turns out to apply to all of them.

To assess causal emergence, we want to look at the causal information gained from deviating from the microscale / fully-detailed model of the system. The theory currently proposes to compare the information generated by a randomized set of interventions on the full description (the microscale) vs. the information generated by a randomized set of interventions on some reduced description (some macroscale). A central reason for this has to do with the theory's aim to measure the information contained in the causal structure. However, let's put aside that reason for now and examine Scott's objection that any difference in information will just be due to a "normalization trick," using some simple example systems.

Consider a Markov process composed of four states, with the transition probabilities and TPM shown below (probabilities in gray scale).



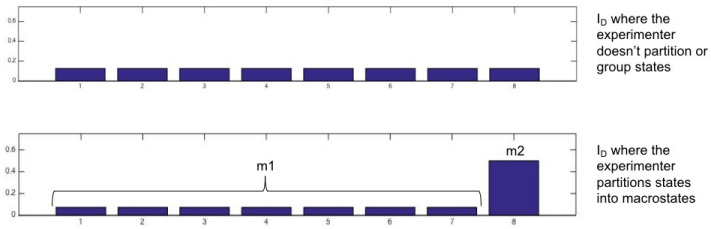
The effective information (EI) of the microscale is 1.32 bits. The bit value comes from applying the full set of possible interventions on the system: an $I_{D,micro}$ of [1/4 1/4 1/4 1/4] over the set $\{s_1, s_2, s_3, s_4\}$. However, for this system, s_1 can be left exogenous, which changes the description of the system to:



This leads to a higher EI value: 1.58 bits. The associated macroscale intervention distribution ($I_{D,macro}$) is [1/3 1/3 1/3] over the set $\{s_2, s_3, s_4\}$. Note that this is also the channel capacity of the system, using the analogy of the I_D being the input to a channel, with the output being future states. Since the macroscale model is derivable from the original system (supervenience holds), there must be some corresponding equivalent intervention distribution at the microscale, $I_{D,equivalent}$. Here this would be an underlying I_D of [0 1/3 1/3 1/3] over the set $\{s_1, s_2, s_3, s_4\}$.

Scott Aaronson implies in his post that this increase of 0.26 bits is just a trick of normalization. He says this because the information generated by $I_{D,equivalent}$ will be the same as that of $I_{D,macro}$: 1.58 bits. Yet, our purpose was to measure the information gained in some reduced description, so to make that comparison you need to have a comparison case where s_1 is actually included. In the $I_{D,equivalent}$ s_1 is neither being intervened upon nor is it ever observed. That is *by definition* excluding s_1 from the model. In fact, *any* intervention distribution that doesn't leave s_1 exogenous, even those that aren't the uniform distribution, will lose information.

Here's another example of this using the same example system Scott gave: a system with eight states, where states s_1-s_7 transition to state s_1, whereas state s_8 transitions to itself. The system is deterministic, and the EI at the microscale (or over the uniform distribution) is 0.54 bits. The system has an obvious macroscale as well, which is coarse-graining states s_1-s_7 into some macrostate m_1. Let's at the macroscale also call s_8 m_2. For this macroscale EI = 1 bit and the $I_{D,macro}$ is [1/2 1/2]. Below is the I_D (at the microscale) that gives you the 0.54 bits. At the bottom is the $I_{D,equivalent}$. It also gives you 1 bit.



Notice the obvious difference? In order to get this increase in information, when applying the I_D the experimenter must carefully partition the microstates up. Any microstate in the group of s_1-s_7 is interchangeable with any other in the group (it's intervened on with the same probability and has the same effect) while s_8 is treated as a unique entity with a separate effect. This is what coarse-graining *is*. As in the first example, any time you're not intervening on states s_1-s_7 like they are interchangeable states and intervening on them in a partitioned manner from the rest of the system, you lose information[1].

So what was called an issue of normalization due to using the uniform distribution is really about picking an appropriate comparison case. As the two examples demonstrate, there are systems where using *any* intervention distribution at the microscale that isn't definitionally equivalent to assessing EI at the macroscale will mean a loss of information. So in such systems you can pick your comparison distribution to be anything you like, and still there is causal emergence.

What about using the $I_{D,equivalent}$ itself as your comparison case? Well, this would mean things like leaving out the majority of states in the "fully-detailed microscale," or partitioning microstates into equivalency classes and intervening on those classes differently. So this isn't the appropriate comparison case for definitional reasons, unless one wants to start purely semantic quibbles about maps vs territories ("It somehow still counts as the fully detailed microscale even when the majority of states aren't included," etc).

Moving away from such semantic debates, the $I_{D,equivalent}$ is also not the appropriate comparison case because we were originally interested in the information contained in the causal relationships at a certain scale. It's clear that the channel capacity at the macroscale can be identical to but never exceed the channel capacity at microscale. But, as can be seen in the above examples, the channel capacity at both the macro and micro scales can be fully accounted for *just* by the information in the causal relationships of the macrostates. Whereas at the microscale, if one looks at just the information in the causal relationships of the microstates, it is much lower than the channel capacity. Now, the channel capacity can indeed be achieved when the experimenter fits their intervention distribution in such a way as to match those higher scale relationships. The way the experimenter fits the distribution is by grouping states, leaving them exogenous, etc, and the paper was demonstrating that this tracks the causal relationships of the higher scale. Indeed, this view indicates that higher scale causal relationships can be viewed as a kind of code for precisely this reason.

---

[1] You can also leave states exogenous in this example, but we've already covered that.