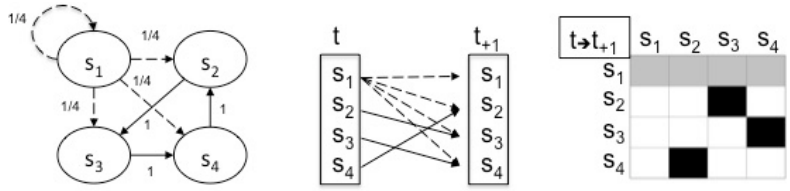
A terminological note: the theory uses the term "macroscale" as an umbrella term for various changes in the description of a system (such as coarse-graining) that involve some sort of dimensional reduction. This includes black boxing (allowing states to vary under the effects of interventions but not including them in the model), or leaving states/elements totally exogenous to the model, and also setting particular initial states or boundary conditions (like fixing states/elements). Why do I call all of these macroscales? Because they are not 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 will work for all of them.

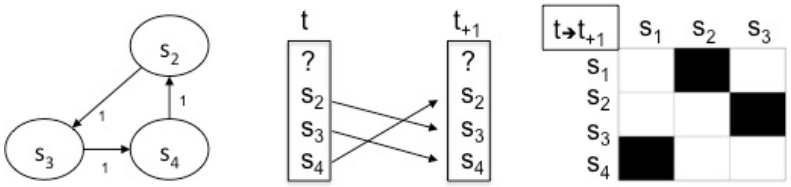
To assess causal emergence, we want to look at the causal information gained from deviating from the microscale / fully-detailed model of the system. The theory currently proposes to compare the information generated by a randomized set of interventions on the full description (the microscale) vs. the information generated by a randomized set of interventions on some changed description (some macroscale). A central reason for using this measure has to do with the theory's aim to capture the information contained in the causal structure. However, even putting aside that reason, let's 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 (the fully detailed system) is 1.32 bits. See the papers for exactly how *EI* is calculated, but the number comes from applying the full set of possible interventions on the system: an I_D of [1/4 1/4 1/4 1/4] over the set $\{s_1, s_2, s_3, s_4\}$. The *EI* is the mutual information between this set of interventions and their effects on the system (the state-transitions they trigger).

For this system, a particular state, s_1, can be left exogenous[1], which changes the description of the system to:



Doing so gives a higher effective information value: 1.58 bits. The macroscale intervention distribution ($I_{D,macro}$) in the system 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

[1]A commenter correctly pointed out that I originally called the first example a case of "black boxing," which is defined in the original paper as when an element is exogenous but still being affected indirectly by intervention. Rather, this is an even simpler case of leaving an element or state exogenous (and not being affected by intervention). Both are macroscales, but thanks to ARaybold for pointing that out.

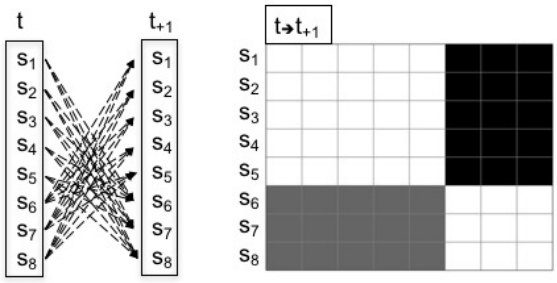
channel, with the output being future states. Since the macroscale model is derivable from the original system (supervenience holds), there must be some corresponding I_D at the microscale that is identical to how the experimenter treats the system at the macroscale. Here this would be an underlying 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 macroscale I_D corresponds to an underlying I_D at the microscale, for which the information generated will be the same: 1.58 bits. This corresponding I_D would here be [0 1/3 1/3 1/3] over the microscale set $\{s_1, s_2, s_3, s_4\}$, and generate the same 1.58 bits. It also generates an output distribution of [0 1/3 1/3 1/3] over the same set $\{s_1, s_2, s_3, s_4\}$.

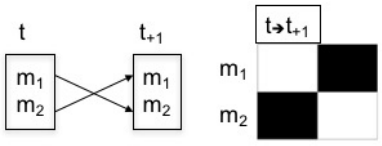
Yet, *to do a comparison of leaving s_1 in or out of the model one must have s_1 both in and out of the model*. In the macroscale-equivalent microscale I_D s_1 is neither being intervened upon nor is it ever observed (it is exogenous to the description). That is **by definition** excluding s_1 from the model. So, there's no sense in which comparing the information gained by excluding s_1 (or lost from including s_1) is based on normalization.

I believe that the other macroscale cases applying a macro-equivalent I_D at the microscale are also obviously changing the model/description of the system in all but name. So Scott's argument misses the point: that changing your description of the system, such as leaving s_1 out in the above case, generates more information. It's not a case of normalization, as the comparisons between the I_Ds are directly based on what happens when we change our description of the system (including a state or not).

Here's an example of this same general point, but with coarse-graining, which is arguably a more subtle case. Consider a Markov process of eight states. States s_1-s_5 transition to states s_6-s_8 with equal probability, whereas states like s_6-s_8 transition to s_1-s_5 with equal probability. The transitions and TPM are shown below, and the *EI* of the system is 0.95 bits.

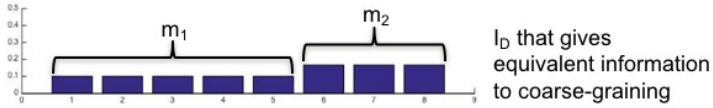


The system has an obvious macro-scale as well, which is when we coarse-grain states s_1-s_5 into some macro-state m_1 and s_6-s_8 into m_1. For this system *EI* = 1 bit and the $I_{D,macro}$ is [1/2 1/2]. I'm aware that this difference between micro and macro is small (only 0.05 bits), but one can make this example as arbitrary large as one wants and that bit difference will increase.



Note also that this 1 bit value at the macroscale is again the channel capacity of the system if we think of the causal structure of the system as a kind of information channel over which we can send interventions. Let's look at the I_Ds, which is what Scott thinks are being normalized. Here's the I_D (at the microscale) that gives you the 0.95 bits:

And here's the I_D (at the microscale) that's equivalent to the $I_{D,macro}$, and gives you 1 bit:



Notice the difference? In order to get this increase in information at the microscale, when applying the I_D the experimenter must carefully partition the microstates into two camps. They must intervene always identically within those groups but always differently between the groups. This is what coarse-graining *is* when you look down to what's going on at the underlying microscale. The sensible comparison case for coarse-graining a model versus not is when an experimenter *doesn't* partition the states into two equivalency classes (i.e., uses the uniform distribution at the top). What the uniform distribution actually corresponds to is the full microscale description of the model.

Since we always ensure that supervenience holds, any change in description of a system means some equivalent change down at the microscale (like leaving a state out of the model, or carefully partitioning up states into groups). The point of the paper is that you can compare before and after that change in description and see the increase. Enacting these same changes, like leaving a state exogenous, or partitioning microstates into different groups, all while still counting that model of the system as the full microscale, is to deny the obvious comparison case of when this isn't done.