A primer on causal emergence: Reply to Scott Aaronson

Erik Hoel

Date: 18 April 2021

URL: https://www.theintrinsicperspective.com/p/a-primer-on-causal-emergence

Main Text

This post was originally inspired by the physicist and blogger Scott Aaronson, who blogged his criticisms about a theory of mine called causal emergence. I unfortunately had little time to reply on his post, and probably didn't do the best job in replying concisely. To see the actual nature of his error, skip down to the section on **Scott Aaronson's criticism**, although this debate does assume you are familiar with some of the theory's terminology. Since Scott's criticisms seemed to reflect a misunderstanding of the theory, it prompted me to do this generalized explainer. Please note this explainer is purposefully designed to not be technical, formalized, or comprehensive. Its goal is to give interested parties a conceptual grasp on the theory.

What's causal emergence?

It's when the higher scale of a system has stronger causal relationships than its underlying lower scale. Causal relationships are just a set of dependencies between some variables, such as states or mechanisms. A causal structure is just the set of these causal relationships. Measuring causal emergence is like you're looking at the causal structure of a system with a camera (the theory) and as you focus the camera (look at different scales) the causal structure snaps into focus. Notably, it doesn't have to be "in focus" at the lowest possible scale, the microscale. Why is this? In something approaching plain English: macrostates can be strongly dependent even while their underlying microstates are only weakly dependent. The goal of the theory is to search across scales until the scale at which variables (like elements or states) are most strongly causally dependent pops out. It uses information theory to do this.

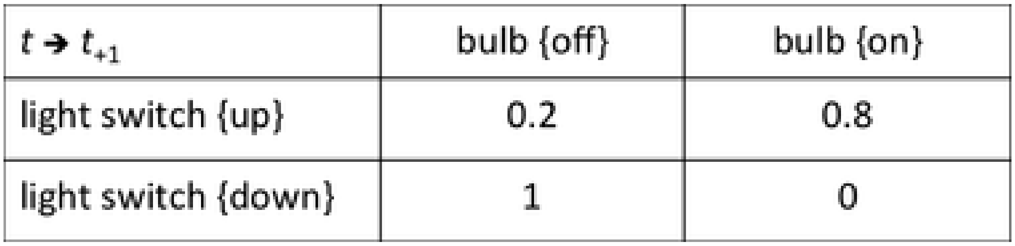
Isn't this against science, reductionism, or physicalism?

Nope. The theory adheres to something called *supervenience*. That's a technical term that means: if everything is fixed about the lower scale of a system, everything about the higher scale must follow. So there's nothing spooky, supernatural, or anti-physicalist about the results. Rather, the theory provides a toolkit to identify appropriate instances of causal emergence or reduction depending on the properties of the system under consideration. It just means that, when thinking about causation, reductionism isn't always best. The higher scales the theory considers are things like coarse-grains (groupings of states or mechanisms) or leaving states or elements out of the system, among others. These are just different levels of description, some of which capture the real causal structure better. In this sense, the theory says that causal interpretations are not relative or arbitrary, but ordered, and there is a best one.

How do you find the strongest causal relationships?

Causation has long been considered a philosophical subject, even though it's at the heart of science in the form of experiments and separating correlation from causation. Causation, much like information, can actually be formalized abstractly and mathematically. For instance, in the 90s and 2000s, a researcher named Judea Pearl introduced something called the do(x) operator. The idea is to formalize causation by modeling the interventions an experimenter makes on a system. Let's say I want to check if there is a causal relationship between a light switch and a light bulb in a room. Formally, I would do(light switch = up) at some time *t*, and observe the effects on the bulb at some time *t*+1.

One of the fundamental ways of analyzing causation is what's called an A/B test, or a randomized trial. For two variables A and B, you randomize those variables and observe the outcome of your experiment. Think of it like injecting noise into the experiment, which then tells you which of those two variables is more effective at producing the outcome. For example, let's say the light bulb flickers into the {off} state while the light switch is in the {up} state 20% of the time. If you do(light switch = up) and then do(light switch = down) at in a 50/50 manner, it reveals the effects of the states. From this, you can construct something (using Bayes' theorem) called a transition table:



$t \rightarrow t_{+1}$	bulb {off}	bulb {on}
light switch {up}	0.2	0.8
light switch {down}	1	0

Note that this reflects the actual causal structure. Flipping the switch {up} really does cause the bulb to turn {on} 80% of the time. Doing the A/B test appropriately screened out everything but the conditional probabilities between the states, such as how often you flipped the switch. And, ultimately, causal relationships are *conditional*. They aren't about the probabilities of the states themselves, but about "if *x* then *y*" classes of statements.

Of course, we can also do A/B/C tests, and so on. What matters is randomizing over everything (creating an independent noise source) so that the result exposes the conditional probabilities between the states. The theory of causal emergence formalizes this as applying an intervention distribution: a probability distribution of do(x) operators. The intervention distribution that corresponds to an A/B test would be [1/2 1/2], and if A/B/C, then [1/3 1/3 1/3]. This is called a maximum entropy, or uniform, distribution.

How does information theory relate to causal structure?

Consider two variables, X and Y. We want to assess the causal influence X has over Y, X → Y. Assume, for now, there are no other effects on Y. If every change in the state of X is followed by a change in the state of Y, then the state of X contains a lot of causal information about Y. So if Y is very sensitive to the state of X, a metric of causal influence should be high. Note this is different than predictive information. You might be able to predict Y given X even if changes in X don't lead to changes in Y (like how sales of swim wear in June could predict sales of air conditioners in July).

To be more formal about assessing X → Y, we inject noise into X and observe the effects on Y. Effective information, or EI, is the mutual information, I(X;Y), between X and Y while intervening to set X to maximum entropy (inject noise). Note that this is the same as applying a uniform intervention distribution over the states of X.

This is a measure of what can be called causal strength, influence, dependency, or constraint. Beyond capturing in an intuitive way how Y is causally dependent on the state of X, here are a few additional reasons that the metric is appropriate: i) setting X to maximum entropy screens off everything but the conditional probabilities to matter in the final value, ii) it's like doing an experimental randomized trial, or A/B test, without prior knowledge of the effects on Y of X's states, iii) it doesn't leave anything out, so if a great many states of X don't impact Y, this will be reflected in the causal influence of X on Y, iv) it's like injecting the maximum amount of experimental information into X, Hmax(X), in order to see how much of that information is reflected in Y, and v) the metric can conceptualized as the expected number of Y/N questions it takes to identify the intervention on X at t given some y at t+1.

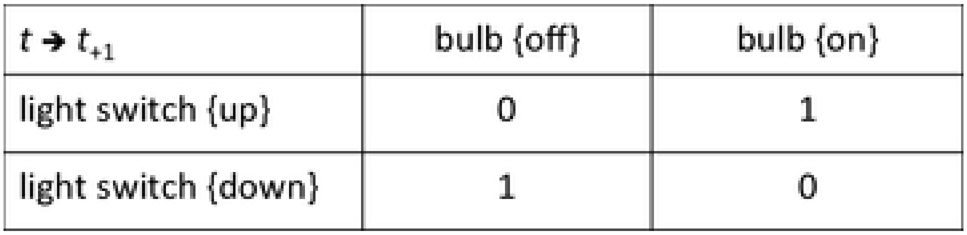
Ultimately, the metric is using information theory to track the counterfactual dependence of Y on X. In traditional causal terminology this is putting a bit value on notions like how necessary and sufficient the state of X is for the state of Y.

EI is low if states of X only weakly determine the states of Y, or if many states of X determine the same states of Y (as those states make Y uncertain about the state of X). It is maximal only if all the causal relationships in X → Y are composed of biconditional logical relationships (iff x then y).

Another way to think about it is that EI captures how much difference the average do(x) makes. In causal relationships where all states transition to the same state, no state makes a difference, so the EI is zero. If all interventions lead to completely random effects, the measure is also zero. The measure is maximal (equal to the logarithm of the number of states) if each intervention has a unique effect (i.e., interventions on X make the maximal difference to Y).

How is the metric applied to systems?

Consider, a simple switch/bulb system where the light doesn't flicker. The relationship of which can be represented by the transition table:



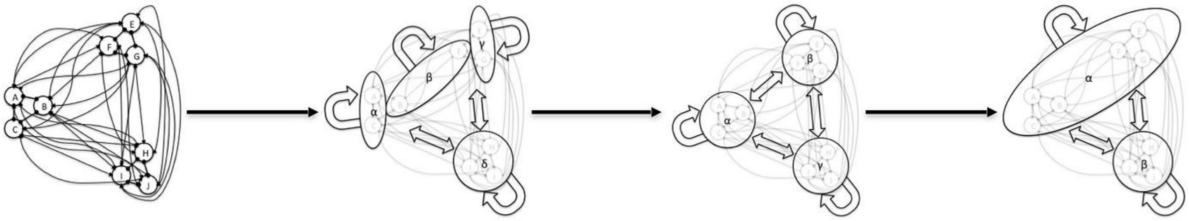
$t \rightarrow t_{+1}$	bulb {off}	bulb {on}
light switch {up}	0	1
light switch {down}	1	0

In this system the causal structure is totally deterministic (there's no noise). It's also non-degenerate (all states transition to unique states). So the switch being {up} is both sufficient and necessary for the bulb being {on}. Correspondingly, the EI is 1 bit. However, for the previous case where the light flickered, the EI would be lower, at 0.61 bits.

Effective information even captures things traditional, non-information-theoretic measures of causation don't capture. For instance, let's say that we instead analyze a system of a light dial (with 256 states) and a light bulb with 256 states of luminance. Both the determinism (1) and degeneracy (0) are identical to the original binary switch/bulb system. But the causal structure overall contains a lot more information: each state leads to exactly one other state and there are hundreds of states. EI in this system is 8 bits, instead of 1 bit.

What does any of this have to do with emergence?

EI can be measured at the macroscale of a system, and compared to the microscale of a system,. What counts a macroscale? Broadly, any description of a system that's not the most detailed microscale. Leaving some states exogenous, coarse-grains (grouping states/elements), black boxes (having states/elements be exogenous when they are downstream of interventions), setting some initial state or boundary condition, all these are macroscales in the broad sense. Moving from the microscale to a macroscale might look something like this:



Interestingly, macroscales can have higher EI than the microscale. Basically, in some systems, doing a full series of A/B tests at the macroscale gives you more information than doing a corresponding full series of A/B tests at the microscale. More generally, you can think about it as how informative a TPM of the system is, and how that TPM gets more informative at higher scales.

Wait. How is that even possible?

There are multiple answers. In a general sense, causal structure is scale-variant. Microscale mechanisms (like NOR gates in a computer) can form a different macroscale mechanism (like a COPY gate). This is because the conditional probabilities of state-transitions change across scales. Consequently, the determinism can increase and the degeneracy can decrease at the higher scale (the causal relationships can be stronger).

Another answer is from information theory. Higher-scale relationships can have more information because they are performing error-correction. As Shannon demonstrated, you can increase how much information is transmitted across a channel by changing the input distribution. The analogy is that intervening on a system at different scales is like trying different inputs into a channel. From the perspective of the microscale, some higher-scale distributions will transmit more information. This is because doing a series of A/B tests to capture the effects of the macroscale states doesn't correspond to doing a series of A/B tests to capture the effects of microscale states. A randomized trial at the macroscale of medical treatments to see their effect on tumors won't correspond to an underlying set of microscale randomized trials, because many different microstates make up the macrostates.

Scott Aaronson's criticism

Scott’s criticism is he believes that the micro and the macro are not being compared fairly. His point is that, *if* you use the same intervention distribution at both the microscale and the macroscale, there can’t be any difference in the effects. To him the use of a uniform distribution therefore seems arbitrary, since a uniform distribution at one scale will not be uniform at another.

It’s worth noting that, definitionally, the measure EI itself is now out the window, since it requires a certain intervention distribution. Moving to a different one means the measure would not longer be able to broken down into the determinism, degeneracy, and size of a system, and remove its potency as a measure of causation.1 But let’s aside whether we *should* use the uniform distribution for a moment.

Instead, let’s just focus on Scott’s actual counterproposal. What he suggests is using the same intervention across all scales. While it may seem like a sensible proposal, it’s wrong for several reasons, once one works through it in detail.

For instance, when applying the macroscale intervention distribution, one would need to weight it by the number of microscale states within each macrostate. This means to calculate the strength of the causal relationship between a light switch and a light bulb, one must know how many atoms are in each microstate of “up” or “down.” This byproduct of Scott’s proposal entails that macroscale causal relationships are totally random in whatever bit number they end up with, since it is based ultimately what microscale intervention distribution you’re using, and can be totally disconnected from any property or relation at the macroscale. Whatever you are now measuring, which is what Scott wants to measure, is not a measure of causal strength, power, or informativeness, at the macroscale.

Scott’s proposal to use an identical intervention distribution at both scales can even lead to comedy. Consider a macroscale where an element of the microscale (consisting of elements ABCD) has been left exogenous. Let’s call the element left exogenous to the model D (and remember there’s also ABC). At the microscale, the causal structure is assessed and EI is measured by intervening on ABCD. So far, so good. Now we come to the macroscale. It’s just a dimension reduction wherein D is left exogenous, that is, not included in the model. So assessing the causal structure here entails intervening on ABC and measuring EI. Here Scott asks us to wait. He claims that the actual fair comparison is to leave D out at the microscale too! That is, to be fair we must compare an intervention distribution over ABC with… an identical one also over ABC… and then call one the “macroscale” and the other the “microscale.” Even though the microscale now leaves out an element, by definition making it not a microscale. That’s not a fair comparison. It’s not even sensible.

Let me state the problem as clearly as I can. Using the intervention distribution of the microscale while intervening at the macroscale *implicitly assumes knowledge of the microscale* (knowing the constituents that make up the macrostates). So there is microscale information in your macroscale interventions. This leads to things like needing to count the number of atoms in a lightswitch to understand what its causal relationship is to a light bulb. I.e., that macroscale causal relationships can’t be defined on their own. On the other side, using the intervention distribution of the macroscale at the microscale *implicitly means intervening on the microscale as if it were a macroscale.* This leads to things like partitioning your probabilities of interventions between micro-states (as with coarse-grains) or leaving micro-states out altogether (as with black boxes). *So under Scott’s proposal you either treat the macro like the micro (rending macroscales definitionally microscales) or you treat the micro like the macro (rending microscales definitionally macroscales).*

This is the definition of comparing apples and oranges. “Microscale interventions” that based on dimension reductions and groupings? “Macroscale interventions” that are weighted in probabilities by their microscale constituents? It is both unattractive on the theory side and also contrary to how people operate when actually intervening to assess causal relationships. Keeping intervention distributions the same leads to informational leakage between micro and macro models. And let me be clear, as long as there is there can be a difference in intervention distributions causal emergence is possible. So either a) reject the notion of fair comparison or b) accept the possibility causal emergence. My suspicion is that Scott would choose (a), and simply reject that one can compare scales at all. But at this point I am almost certain that instead of then saying issues of reduction/emergence cannot be resolved, he would still maintain universal reductionism. Yet the motivation for this position seems weak, since it is only taken to avoid the conclusion of what happens when you *do* compare scales under a measure like effective information.

So what is a valid criticism?

Plenty! There could be improvements and there might even be a fatal flaw somewhere. The biggest issue in my mind is that causal emergence shows up in various measures (like the EI, or the integrated information) in that they are greater at the macroscale. But how to convince someone that you are using the best, or even better, a unique measure to capture the phenomenon?

That’s actually a much stronger criticism what Scott did, and much more honest, since it doesn’t propose some thrown-together new comparison; rather, it is agnostic in just saying: “Why should I believe this is broadly true?” Scott could have taken his criticism of the uniform intervention in this direction, arguing that it indicates a kind of artificiality to the measure of EI itself and that there could be superior measures of causation that *don’t* show this effect. There are of course good reasons we use the EI, and good reasons to believe it is a successful measure of causation. But there could indeed be superior measures that contradict the results. Perhaps one could convince an extreme skeptic like him by proving a measure like effective information is unique and best-fitted for measuring causation, and therefore you shouldn’t use it only on one side of your comparison, but on both. But proving this with any mathematical measure of *anything* is both rare and hard, and neither I nor anyone else has done this. However, this criticism is not a good reason to disbelieve the conclusions of the theory. It’s, at best, a reason to go beyond the EI in this research. That would have been a much better blog post.

In this sense Scott’s proposal is that he wants a different, non-EI, but kind of like-EI, bit measure for causal relationship strength that fits with his intuitions (this is on par with his other ideas about these subjects, such as his criticism of Integrated Information Theory, where he ends up just saying he wants a measure that fits with his intuitions, with minimal to no justification of his intuitions).

References

Erik Hoel (n.d.). A primer on causal emergence: Reply to Scott Aaronson. www.theintrinsicperspective.com. https://www.theintrinsicperspective.com/p/a-primer-on-causal-emergence