AGENT ABOVE, ATOM BELOW

How agents causally emerge from their underlying microphysics

—

by Erik P Hoel*

**Department of Biological Sciences, Columbia University, New York, NY, USA*

Abstract

Some physical entities, which we often refer to as agents, can be described as having intentions and engaging in goal-oriented behavior. Yet agents can also be described in terms of low-level dynamics that are mindless, intention-less, and without goals or purpose. How we can reconcile these seemingly disparate levels of description? This is especially problematic because the lower scales at first appear more fundament in three ways: in terms of their causal work, in terms of the amount of information they contain, and their theoretical superiority in terms of model choice. However, recent research bringing information theory to bear on modeling systems at different scales significantly reframes the issue. I argue that agents, with their associated intentions and goal-oriented behavior, can actually causally emerge from their underlying microscopic physics. This is particularly true of agents because they are autopoietic and possess (apparent) teleological causal relationships.

1. Agents excluded

> "Marco Polo describes a bridge, stone by stone.
>
> 'But which is the stone that supports the bridge?' Kublai Khan asks.
>
> 'The bridge is not supported by one stone or another,' Marco answers, 'but by the line of the arch that they form.'
>
> Kublai Khan remains silent, reflecting. Then he adds: 'Why do you speak to me of the stones? It is only the arch that matters to me.'
>
> Polo answers: 'Without stones there is no arch.'"
>
> — Italo Calvino [1]

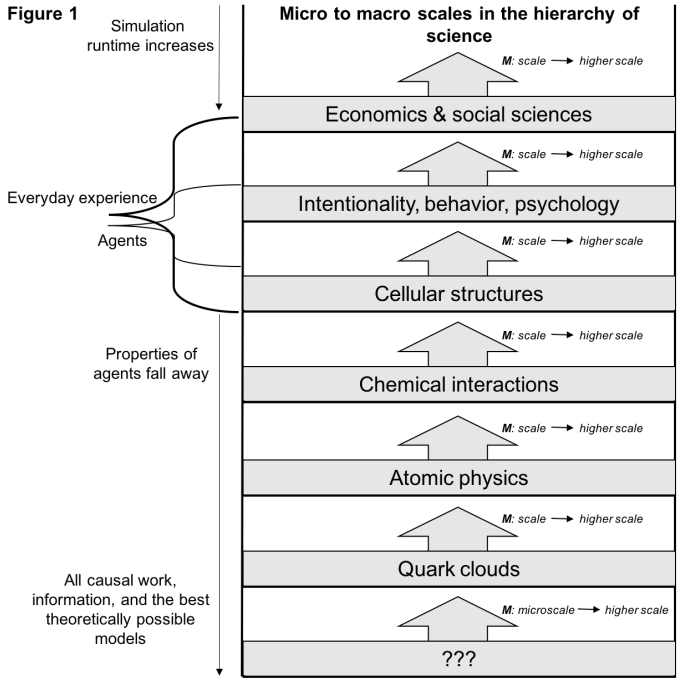
Agents are so central to everyday experience that they form part of the ontology of our lives. From our daily interactions with entities that evince goals, intentions, and purpose, we reasonably conclude that agents both exist and act as a causal force in the world. And yet there is a dissonance between our easy belief in the ubiquity of agents and what we know of physics. Down at the level of the Hamiltonian, the level of the quark, all of reality is describable as merely lawful changes in the state of incredibly small components. Even if these many-body systems are hopelessly unpredictable in practice because of their complex interactions, or exist in unique or interesting configurations, all the individual components still follow strict and purposeless laws concerning their evolution in time. Down here we see no sign of agents, their intentions, goal-oriented behavior, or their evolved pairing with their environments.

I'd like to begin by arguing that this dissonance between purposeful agent-like behavior at the top, and purposeless and automaton-like behavior at the bottom, speaks to a larger issue. While many systems lend themselves to descriptions at higher scales, there is always a base scale that

underlies, or fixes, all those above it. Given just the workings of the base, one could, at least in theory, derive all scales above it.

In the language of analytic philosophy this relationship is called *supervenience*: given the lower-level properties of a system, the higher-level properties necessarily follow [2]. The base scale is the microphysical scale (the finest possible representation of the system in space and time). Deriving the scales above it can be conceptualized as mappings from the base microscale to supervening macroscales, where each mapping is over the states, elements, functions, mechanisms, or laws. Broadly, a macroscale is any scale that contains fewer elements or states, or simpler laws or functions. A well-known example of a macroscale is a coarse-grain, such as between temperature and the individual motion of particles. The astronomical set of possible microstates constituting all combinations of particle kinetic energy can be coarse-grained into the single macrostate of its temperature. In this case the mapping from the states of the individual particles to the temperature of the whole is a function (the average) of the kinetic energy. Not all macroscales are coarse-grains; they might be a subset of the state-space, like a black box with inputs and outputs (see Appendix A for technical notes). One universal way to think of macroscales is to consider the runtime of a full simulation of the system: moving up in scale decreases runtime.

Explicitly considering these relationships across scales reveals a fundamental issue. Given that supervening macroscales can all be derived from the base microscale, it would seem natural that some form of Occam's razor applies. That is, the higher-scale descriptions aren't necessary. The most aggressive form of the argument is known as the 'exclusion argument.' Given that all systems are describable at the physical microscale, what possible extra causal work is there for any supervening macroscale [3]? And if those higher scales don't contribute causal work then they are by definition epiphenomenal: shadows of what's really going on. So it seems that the universe collapses like a house of cards down to just the purposeless events and causal relationships at the physical microscale.



While this may seem a rather abstract concern, I think it can be framed as a triad of issues not confined to philosophical ruminations: model choice (microscale models are always better, at least in principle), causation (all causal influence or work is at the base microscale, at least in principle), and also information (no new information can be gained by moving up in scale, at least in principle). In this broadened scope it affects all of science, especially the so-called "special sciences," like biology and psychology, which seem to comfortably float above physics as their own distinct domains. The special sciences are a reflection of one of the remarkable but often unremarked upon aspects of science: its large-scale structure.

Science forms a kind of ladder of ascending spatiotemporal scales (or levels), wherein each scale supervenes on those below it (Figure 1). The ladder is constructed of supervening relationships between scales. Yet even in this hierarchical structure all the information and causal

work seems to drain away down to the microscale [4, 5]. Not only that, but the exclusion of higher scales don't even require a final base microscale of physics. Rather, each rung could be causally reducible to the rung below it.

Agents are generally somewhere above biological mechanisms but below economics on the ladder. They are a major part of the slim section that corresponds to the scale of our everyday experiences. Agents themselves can be described, both mathematically and not, in agent-based modeling [6], in game theory [7], in behavioral analysis [8], biology [9], economics [10], and many other fields. While one might use evolutionary theory, or detailed mechanistic explanations, to show how to construct the mapping function from the scale of quarks up to the scale level of agent states or behaviors, this would beg the more perplexing question: how can something be both an agent at one level and not at agent at another? How can agents *really* exist when they are always definable in terms of goal-free and purposeless dynamics?

After all, as long as supervenience holds then agents appear to be just one more reducible representation of the underlying physics. At some point down the ladder all agent properties disappear. So while no one questions that some physical systems can be described in terms of their goals, intentional states, and behavior, the exclusion argument means that the purposelessness of microscale physics reigns supreme. It is these purposeless dynamics which theoretically make up the most informative model of the system and which are doing all the causal work in the system.

The reductionist, one might even say nihilistic, answer to this conceptual knot can be stated clearly using the terminology of information theory: compression. Macroscales are useful or elegant summarizations. They are, at best, lossless representations of the information and causal structure, but are far more likely to be lossy and thus missing crucial information or causal relationships. Their usefulness stems from the necessity of compression in communication, because all systems have a limited bandwidth. Macroscales make the world summarizable and thus understandable. In this reductionist view, a biologist studying a cell is really referring to some astronomically complex constellation of quarks. It is merely because of the limited data storage, cognitive ability, and transmission capabilities of the human brain and the institutions it creates that we naively believe that cells exist as something above and beyond their underlying microscale. Any abstractions or higher-level explanations assist our understanding only because our understanding is so limited.

This formulation of the problem in terms of higher and lower scales puts aside the details and complications of physics. This generality is important because agenthood seems likely to be a multiply-realizable property: different microscopic scales, laws, or mechanisms may lead to the same agent behavior. For instance, artificial agents created in simulations, such as Conway's the Game of Life [11] have basic rules that may differ drastically. If agents are multiply realizable in this manner, then any attempts to link the unique properties of agents to some unique property of our physics is doomed to underdetermination.

Recently, when asked, "What concept should be better known in science?" Max Tegmark answered: a form of multiple-realizability called substrate-independence. He gave the example of waves propagating in different mediums, and remarked that "we physicists can study the equations they obey without even needing to know what substance they are waves in" [12].

This property of multiple-realizability at first seems itself a good enough to not immediately exclude macroscales as mere summarizations [13]. But it's more suggestive than it is definitive. Unless, that is, the fact that something is multiply realizable means it is somehow doing extra causal work, or somehow contains more information? Recent research has argued exactly this [14, 15] by demonstrating the possibility of *causal emergence*: when a macroscale contains more information and does more causal work than its underlying microscale.

I'll first introduce causal emergence and then argue that agents not only causally emerge, but that significant aspects of their causal structure cannot be captured by any microphysical model.

2. Causal emergence

"We may regard the present state of the universe as the effect of its past and the cause of its future." — Pierre-Simon Laplace [16]

Demonstrating causal emergence requires causally analyzing systems across a multitude of scales. Luckily, causal analysis has recently gone through an efflorescence. There is now a causal calculus, primarily developed by Judea Pearl, built around interventions represented as an operator, $do(X = x)$ [17]. The operator sets a system into a particular state (at some time t) and then one can observe the effect at some time t_{+1}. This can be used to assess the causal structure of systems at different scales by applying either a macro or micro intervention. For instance, a micro intervention might be setting the individual kinetic energy of all the particles in a gas, while a macro intervention would be fixing the average kinetic energy (temperature) but not specifying what the individual particles are doing (see Appendix A for further details).

For the following proof-of-principle argument, let's just assume that all systems we're talking about are discrete, finite, and have a defined microscopic scale. To understand the causal structure of such a system, a set of interventions is applied to reveal the effect of each state. This is represented as an intervention distribution (I_D) of individual $do(X = x)$ operators, which will also generate a set of effects (future states), the effect distribution (E_D). The causal structure of a system can be inferred from the application of I_D (and a liberal helping of Bayes' rule) to derive the conditional probabilities $p(x|y)$ for all state transitions. Notably, the conditional probabilities can change significantly at the macroscale. For example, consider the transition matrix of a toy system (or causal model) with a microscale describable as a Markov chain:

$$S_{microscale} = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow{yields} S_{macroscale} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where a macroscale is constructed of a grouping (coarse-grain) of the first three microstates (it is multiply realizable). Notice how the conditional probabilities have changed at the macroscale: transitions are more deterministic and also less degenerate (less mutual overlap in transitions). This means that interventions at the macroscale are more sufficient in producing a specific effect and more necessary for those effects [14].

Quantifying this micro-to-macro change can be done using information theory. This is possible because causal structure is a matrix that transforms past states into future states; it thus can be conceived of as an information channel. In this view, individual states are like messages and interventions are like the act of sending them over the channel [15]. We can then ask how much each state (or intervention) reduces the uncertainty about the future of the system by measuring the mutual information between some intervention distribution and its resultant effect distribution: $I(I_D;E_D)$. It's an equation that quantifies Gregory Bateson's influential definition of information as "a difference that makes a difference" [18]. The method of intervening on system states and assessing the mutual information between past and future states has been called *effective information* [14, 19] and reflects the determinism and degeneracy of the system's causal structure (see Appendix B). In the above toy case, over the full set of possible interventions (all possible states) the microscale has 0.81 bits, while the macroscale has 1 bit.

Where does this extra information come from? If we consider the original Markov chain it's obviously not a perfect information channel. Rather there is noise and overlap in the state transitions. This might be irreducible to the channel, or it might be a function of the measurement device, or the channel might be an open system and the outside world provides the noise. Notably,

in 1948 Claude Shannon showed that even noisy channels could be used to reliably transmit information via channel coding [20]. His discovery was that channels have a particular capacity to transmit information, which in noisy channels can only be approached by using codes that correct for errors (such as using only a subset of the state-space): $C = max_{p(X)}I(X;Y)$.

I've argued there is an analogous causal capacity of systems to transform sets of states (such as input states, internal states, or interventions) into future states. Continuing the analogy, the input $p(X)$ is actually some intervention distribution $p(I_D)$. Under this view different scales operate exactly like channel codes, and the creation of macroscales is actually a form of encoding. (Appendix C). Furthermore, models at different scales capture the causal capacity of a system to different degrees. This means we can finally specify exactly what makes multiply-realizable entities (sometimes) do more causal work and generate more information: they provide error-correction for causal relationships and thus can be more informative than their underlying microscale, a phenomenon grounded in Shannon's discovery of error-correcting codes [15].

If this is true then causal emergence, whether through irreducible physical properties or because of measurement and observational tools, may explain why science has the hierarchal large-structure that it does. New rungs in the ladder of science causally emerge from those below them. Different scientific fields are literally encodings by which we improve our understanding of nature.

3. Agents and the scale of identity

"If it isn't literally true that my wanting is causally responsible for my reaching and my itching is causally responsible for my scratching… if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world."
— Jerry Fodor [21]

Many types of systems can demonstrate causal emergence, meaning that the full amount of information and causal work can only be captured at a macroscale. Some of these macroscales can be classified as agents that operate via purposes, goal-oriented behavior, intentional states, and exist in intelligent or paired relationships to their environment. It is actually these unique agential properties that prime agents to causally emerge from their underlying microscales.

For instance, agents are stable at higher spatiotemporal scales but not lower ones. They maintain their identity over time while continuously changing out their basic constituents. While the exact numbers and timescales are unknown, nearly every component part of you, from atoms to cellular proteins, are replaced over time. Like Theseus' ship, agents, such as Von Neumann machines or biological life, show a form of self-repair and self-replacement that is always ongoing, particularly at lower scales. This self-maintenance has been called autopoiesis and is thought to be one of the defining aspects of life and agents [22], a kind of homeostasis but for maintaining identity. As autopoietic systems, agents need to intake from the world for metabolic reasons, thus locally increasing the entropy around them [23]. Most importantly, during this process the supervening macroscale of an agent is much more likely to remain stable than its underlying physical microscale, precisely because it is multiply realizable. This is particularly true of the macroscale causal relationships, which may be stable even as their underlying microscale causal relationships are ephemeral.

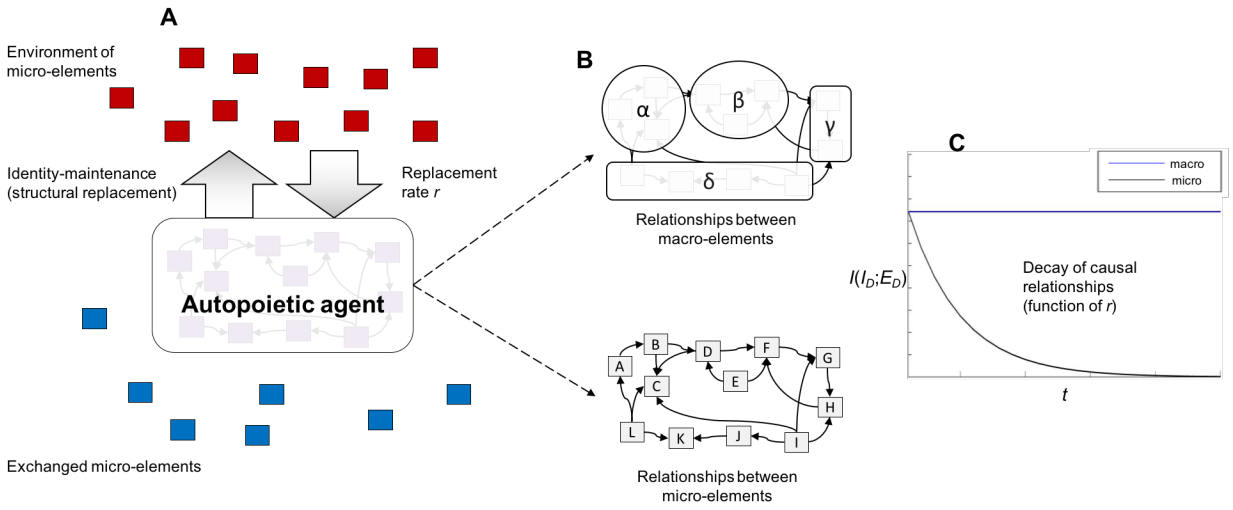


Figure 2. *A toy autopoietic agent*. **(A)** The agent engages in 'feeding' behavior in which it replaces its micro-elements with those from the surrounding environment. **(B)** The supervening macroscale causal relationships, compared to those of the underlying microscale. **(C)** Pairs of causal relationships either rapidly decay at the microscale as elements enter and leave the agent (after exiting, an intervening on one doesn't affect the other). However, they are stable at the macroscale. The response to sets of interventions would correspondingly decrease their informativeness over time, but only at the microscale.

Consider the simple system shown in Figure 2, where it is clear that only the supervening macroscale stays causally relevant over time because of autopoiesis. Rather than specifying the specific micro-element identities (A, B, C…) at each new timestep of the microscale, we can just specify the stable macro-elements (α, β, γ…) via some relational mapping. Defining the system in terms of its relations (rather than the identities of individual elements) seems to fit well the definition of a macroscale: if we again think about scale in terms of simulation time, relational descriptions don't need to be constantly updated as new components replace the identities of old components; thus their runtime is shorter than the microscale.

"Fine," a fervent reductionist might answer, biting the bullet. "While we must continuously update our microscale causal model of the system, there is still no information or causal work to be gained."

Of course, we already know this can be untrue because of causal emergence. However, there are even cases of causal emergence where there is a causal relationship at a macroscale (such as a relational definition) without any underlying microscale. The macro-elements still supervene on the micro-elements, but the macroscale causal relationships don't directly supervene on any underlying microscale causal relationships.

A thought experiment makes the point. Imagine an agent that's quite large (many microscopic components) and these components have a high replacement rate r. Like some sort of amoeba, it internally communicates by waves spreading across its body; what matters in this contrivance is that different parts of its body might have a long lag time in their causal relationships. Now imagine that the lag time (the propagation timescale t) is actually greater than the average turnover rate of the microscopic building blocks. At the microscale where identity is always being updated (as A is replaced by Z, and so on), the causal relationships for which $t > r$ wouldn't even exist, as the receiver would always drop out of the agent before the intervention on the sender could reach it. To give a suggestive example in plain English, a description like "intervening the far left element triggers a response in the far right element" wouldn't be possible

in the less-abstract microscale language of "intervening on element A always triggers a response in Z" because Z is always replaced before it can respond.

In general, we can refer to the reductive instinct that ignores this type of problem as *causal overfitting*: when a representation restricted to the least-abstract base microscale fails to capture the causal structure appropriately. This overfitting can even completely miss that causal work and information entirely: there may be causal relationships that exist solely at the level of agents. It points us to a good (if cloyingly tautological) definition of when something that seems like an agent *really* is an agent: it's when the only way of avoiding causal overfitting it to consider the system in terms of homeostatic, representational, or intentional states.

4. Teleology as breaks in the internal causal chain

"Romeo wants Juliet as the filings want the magnet; and if no obstacles intervene he moves towards her by as straight a line as they. But Romeo and Juliet, if a wall be built between them, do not remain idiotically pressing their faces against its opposite sides like the magnet and the filings... Romeo soon finds a circuitous way, by scaling the wall or otherwise, of touching Juliet's lips directly. With the filings the path is fixed; whether it reaches the end depends on accidents. With the lover it is the end which is fixed, the path may be modified indefinitely."
— William James [24]

The purposeful actions of agents are one of their defining characteristics, but are these intentions and goals actually causally relevant or are they just carried along by the causal work of the microscale? As William James is hinting at, the relationships between intentions and goals seem to have a unique property: their path can be modified indefinitely. Following the logic above, they are causally relevant because as causal relationships they provide for error-correction *above and beyond* their underlying microscales. How does this feed into the real, or perhaps merely apparent, teleology of agents?

One of the defining aspects of agents is that they are open systems. For example, the most reductive model of an agent would be a self-contained microphysical system, i.e., some set of microphysical states along with the associated possible state-space trajectories. But reductively restricting our view to just the internal structure of the agent can lead to *causal underfitting*. This in turn gives us a good definition of teleological causation.

Consider the case of Romeo's brain. While keeping in mind this is a drastic simplification, we can still represent it as some finite set of internal states, specifically as a Markov process: a finite stochastic model made up of states $\{s_a, s_b, s_c, \ldots\}$ and governed by some transition matrix $\boldsymbol{T}$, so it can be represented as a transition graph (Figure 3).

If Romeo's brain is cut off from its environment, let's also assume that as a Markov process it is reducible, so it's not always possible to get from one state to any other state. So if there is a state s_d, which is Romeo's desire to kiss, and a state s_k, which is kissing (or the feeling of a kiss), there may be no path between them. Which further means that some experimenter would conclude there's no possible causal relationship between s_d and s_k in Romeo's brain. So imagine the surprise of the same experimenter intervening to set Romeo's brain into s_d in a situation where Romeo is no longer isolated from his environment (and his Juliet). Struck by Cupid's arrow, Romeo will indefinitely pursue his goal of kissing Juliet, and to the experimenter's surprise s_d will inexorably, almost magically, always lead to s_k. It will appear as if teleology stepped in and ushered the system along a deterministic causal path that didn't exist given just the properties of the system itself.

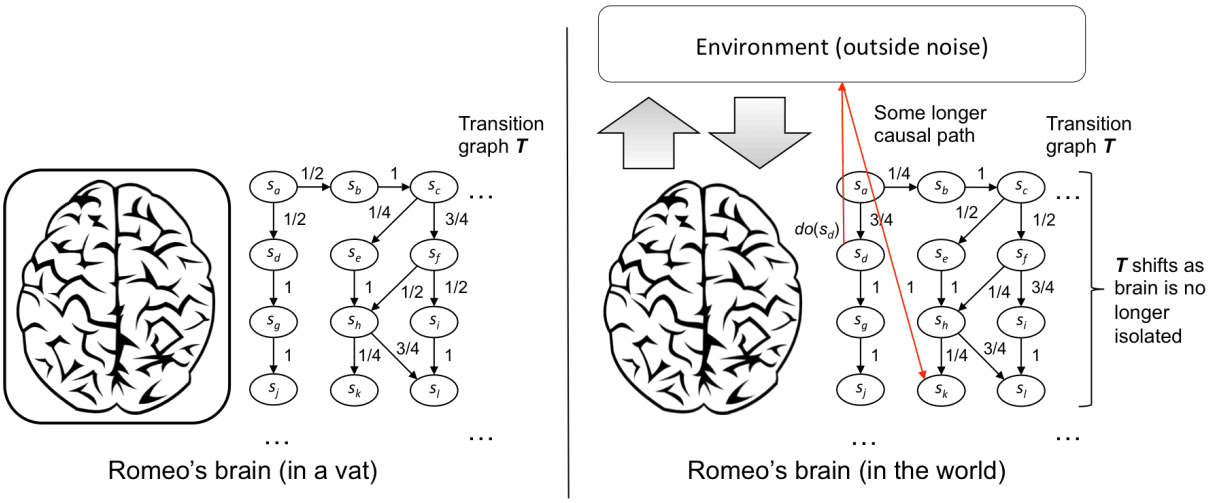


Figure 3. *Teleological causation*. **(Left)** Romeo's brain and its transition graph while separated from the environment. There is no possible path of any length from s_d to s_k. **(Right)** Shown in red is a longer transition which eventually connects s_d to s_k but only operates because the agent is an open system. It appears merely as if the noise transitions have changed (over some further down / future transitions) but really a deterministic causal relationship has been outsourced to the environment.

Considering just the internal architecture of Romeo's brain gives no hint of all the causal relationships that actually exist within it because it has outsourced so much of its causal structure to the environment. Note that for these causal paths to be reliable despite the vagaries of the environment they must be macro. Furthermore, we can identity what constitutes a teleological causal relationship: it's when causal relationships don't supervene locally on the system. The causal structure simply doesn't inhere to the system itself; it is only when the state of the entire environment is taken into account does it appear. Regardless of whether this is truly teleology, or rather just the local appearance of it, the source is the purposeful and goal-oriented behavior of Romeo as an agent.

5. Agents retained

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."
-Claude Shannon [20]

For any given system, the world is a noisy place (see Appendix D). If the system's causal structure is constructed down at the level of tokens, rather than encoded into types, it will be unpredictably buffeted by the slightest outside perturbation. As we've seen, changes in scale are a form of channel coding, and by encoding up at macroscales causal relationships can error-correct, do more work, generate more information, and offer a better model choice for experimenters. It may be that scientists *en masse* implicitly choose our scales for this reason, and our creation of a hierarchy of science is truly following Plato's suggestion to "carve nature at its joints."

Agents are exemplars of causal emergence; they are codes in nature that use to their advantage being open and autopoietic and having goal-oriented behavior, thus allowing for stable-over-time and teleological causal relationships. Ultimately, this means that attempting to describe an agent down at the level of atoms will always be a failure of causal model fitting. Agents really do

exist and function, despite admitting of goal-free and purposeless microscale descriptions. While these descriptions are technically true, in the framework developed here the purposeless microscale descriptions are like a low dimensional slice of a high dimensional object (Appendix E).

Acknowledgements

I thank Giulio Tononi, Larissa Albantakis, and William Marshall for our collaboration during my PhD. The original research demonstrating causal emergence was possible [14] was supported by Defense Advanced Research Planning Agency (DARPA) Grant HR 0011-10-C-0052 and the Paul G. Allen Family Foundation.

Technical Endnotes

A. Scales and interventions

To simplify only discrete systems with a finite number of states and/or elements are considered in all technical endnotes. The base microscopic scale of such a system is denoted S_m, which via supervenience fixes a set of possible macroscales {**S**} where each macroscale is some S_M. This is structured by some set of functions (or mappings) $\boldsymbol{M}: S_m \rightarrow S_M$ which can be of microstates and over space, time, or both.

These mappings often take the form of partitioning S_m into equivalence classes. Some such macroscales are coarse-grains: all macrostates are projected onto by one or more microstates [14]. Other macroscales are "black boxes" [25]: some microstates don't project onto the microstate so that only a subset of the state-space is represented [15, 26]. Endogenous elements in microscale (not projected onto the macroscale) can either be frozen (fixed in state during causal analysis) or allowed to vary freely.

To causally analyze at different scales requires separating micro-interventions from macro-interventions. A micro-intervention sets S_m into a particular microstate, $do(S_m = s_m)$. A macro-intervention sets S_M instead: $do(S_M = s_M)$. If the macrostate is multiply-realizable then a macro-intervention corresponds to:

$$do(S_M = s_M) = \frac{1}{n} \sum_{s_{m,i} \in s_M} do(S_m = s_{m,i})$$

where n is the number of microstates (s_i) mapped into S_M.

B. Effective information and causal properties

Effective information (EI) measures the result of applying some *Intervention Distribution* (I_D) itself comprised of probabilities $p(do(s_i))$ which each set some system S into a particular state s_i at some time t. Applying I_D leads to some probability distribution of effects (E_D) over all states in S. For systems with the Markov property each member of I_D is applied at t and E_D is the distribution of states transitioned into at t_{+1}. For such a system S then EI over all states is:

$$EI(S) = \frac{1}{n} \sum_{s_i \in S} D_{KL}\left(S_F | do(S = s_i) || E_D\right)$$

where n is the number of system states, D_{KL} is the Kullback-Leibler divergence [27], and (S_F |do(S = s_i)) is the transition probability distribution at t_{+1} from $do(S = s_i)$. Notably, if we are considering the system at the microscale S_m, EI would be calculated by applying I_D uniformly (H_{max}, maximum entropy), which means intervening with equal probability ($p(do(s_i) = 1/n)$ by setting S into all n possible initial microstates ($do(S = s_i) \forall_i \in 1 \ldots n$). However, at a macroscale I_D may not be a uniform distribution over microstates, as some microstates may be left out of the I_D (in the case of black boxing) or grouped together into a macrostate (coarse-graining).

Notably, EI reflects important causal properties. The first is the determinism of the transition matrix, or how reliable the state-transitions are, which for each state (or intervention) is:

$$D_{KL}(S_F | do(S = s_i) || H_{max}).$$

While the degeneracy of the entire set of states (or interventions) is: $D_{KL}(E_D||H_{max})$. Both determinism and degeneracy are [0, 1] values, and if one takes the average determinism, the degeneracy, and the size of the state-space, then: $EI = (determinism - degeneracy) * size$.

C. Scales as codes

The capacity of an information channel is: $C = max_{p(X)} I(X;Y)$, where $I(X;Y)$ is the mutual information $H(X) - H(X|Y)$ and p(X) is some probability distribution over the inputs (X). Shannon recognized that the encoding of information for transmission over the channel could change $p(X)$: therefore, some codes used the capacity of the information channel to a greater degree.

According to the theory of causal emergence there is an analogous causal capacity for any given system: $CC = max_{(I_D)} EI(S)$.

Notably, for the microscale S_m $I_D = H_{max}$ (each member of I_D has probability $1/n$ where n is the number of microstates). However, a mapping $\boldsymbol{M}$ (described in Appendix A) changes I_D (Appendix B) so that it is no longer flat. This means that EI can actually be higher at the macroscale than at the microscale, for the same reason that the mutual information $I(X;Y)$ can be higher after an encoding. Searching across all possible scales leads to EI_{max}, which reflects the full causal capacity. EI can be higher from both coarse-graining [14] and black-boxing [15].

D. What noise?

If the theory of causal emergence is based on thinking of systems as noisy information channels, one objection is that real systems aren't actually noisy. First it's worth noting that causal emergence can occur in deterministic systems that are degenerate [14]. Second, *in practice* nearly all systems in nature are noisy due to things like Brownian motion. Third, any open system receives some noise from the environment, like a cell bombarded by cosmic rays. If one can only eliminate noise by refusing to take any system as *that* system, this eliminates noise but at the price of eliminating all notions of boundaries or individuation. Fourth, how to derive a physical causal microscale is an ongoing research program [28], as is physics itself. However, it is worth noting that *if* the causal structure of the microscale of physics is entirely time-reversible, *and* the entire universe is taken as a single closed system, then it is provable that causal emergence for the universe as a whole is impossible. However, as Judea Pearl has pointed out, if the universe is taken as a single closed system then causal analysis itself breaks down, for there is no way to intervene on the system from outside of it [17]. Therefore, causal emergence is in good company with causation itself in this regard.

E. Top-down causation, supersedence, or layering?

To address similar issues, others have argued for *top-down causation*, which takes the form of contextual effects (like wheels rolling downhill [29]), or by groups of entities can have different properties than those entities (water is wet but individual H2O molecules aren't). Others have argued that causation has four different Aristotelian aspects and different scales fulfill the different aspects [30]. It's also been suggested that the setting of initial states or boundary conditions constitute evidence for top-down causation [31], although one might question this because those initial states or boundary conditions can themselves also be described at the microscale.

Comparatively, the theory of causal emergence has so far been relatively metaphysically neutral. Openly, its goal is to be intellectually useful first and metaphysical second. However, one ontological possibility is that causal emergence means the macroscale supersedes (or overrides) the causal work of the microscale, as argued originally in [14]. A different metaphysical option is that scales can be arranged like a layer cake, with different scales contributing more or less causal work (the amount irreducible to the scales below). Under this view, the true causal structure of physical systems is high dimensional and different scales are mere low dimensional slices.

References

[1] Calvino, I. (1978). *Invisible cities*. Houghton Mifflin Harcourt.
[2] Davidson, D. (1970). Mental Events. Reprinted in Essays on Actions and Events, 1980, 207-227.
[3] Kim, J. (2000). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT press.
[4] Bontly, T. D. (2002). The supervenience argument generalizes. *Philosophical Studies*, *109*(1), 75-96.
[5] Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research*, *67*(1), 133-150.
[6] Castiglione, F. (2006). Agent based modeling. *Scholarpedia*, *1*(10), 1562.
[7] Adami, C., Schossau, J., & Hintze, A. (2016). Evolutionary game theory using agent-based methods. *Physics of life reviews*, *19*, 1-26.
[8] Skinner, B. F. (1938). The behavior of organisms: an experimental analysis.
[9] Schlichting, C. D., & Pigliucci, M. (1998). *Phenotypic evolution: a reaction norm perspective*. Sinauer Associates Incorporated.
[10] Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American economic review*, *93*(5), 1449-1475.
[11] Conway, J. (1970). The game of life. *Scientific American*, *223*(4), 4.
[12] Max Tegmark's answer to the Annual Edge Question (2017) https://www.edge.org/annual-questions.
[13] Fodor, J. A. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, *28*(2), 97-115.
[14] Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, *110*(49), 19790-19795.
[15] Hoel, E. P. (2016). When the map is better than the territory. *arXiv preprint arXiv:1612.09592.*
[16] Laplace, P. S. (2012). *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator* (Vol. 13). Springer Science & Business Media.
[17] Pearl, J. (2009). *Causality*. Cambridge university press.
[18] Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press.
[19] Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC neuroscience*, *4*(1), 31.
[20] Shannon, Claude E. (1948). "A Mathematical Theory of Communication". *Bell System Technical Journal*. 27 (4): 623–666.
[21] Fodor, J. A. (1990). *A theory of content and other essays*. The MIT press.
[22] Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition—The Realization of the Living*, ser. Boston Studies on the Philosophy of Science. Dordrecht, Holland: D.
[23] England, J. L. (2013). Statistical physics of self-replication. *The Journal of chemical physics*, *139*(12), 09B623_1.
[24] James, W. (1890). The principles of psychology. New York. *Holt and company*.
[25] Ashby, William Ross. "An introduction to cybernetics." *An introduction to cybernetics (1956).*
[26] Marshall, W., Albantakis, L., & Tononi, G. (2016). Black-boxing and cause-effect power. *arXiv preprint arXiv:1608.03461*.
[27] Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.
[28] Frisch, M. (2014). *Causal reasoning in physics*. Cambridge University Press.
[29] Sperry, R. W. (1969). A modified concept of consciousness. *Psychological review*, *76*(6), 532-536.
[30] Ellis, G. (2016). *How can Physics Underlie the Mind*. Springer, Berlin, Heidelberg, New York.
[31] Noble, D. (2012). A theory of biological relativity: no privileged level of causation. *Interface focus*, *2*(1), 55-64.