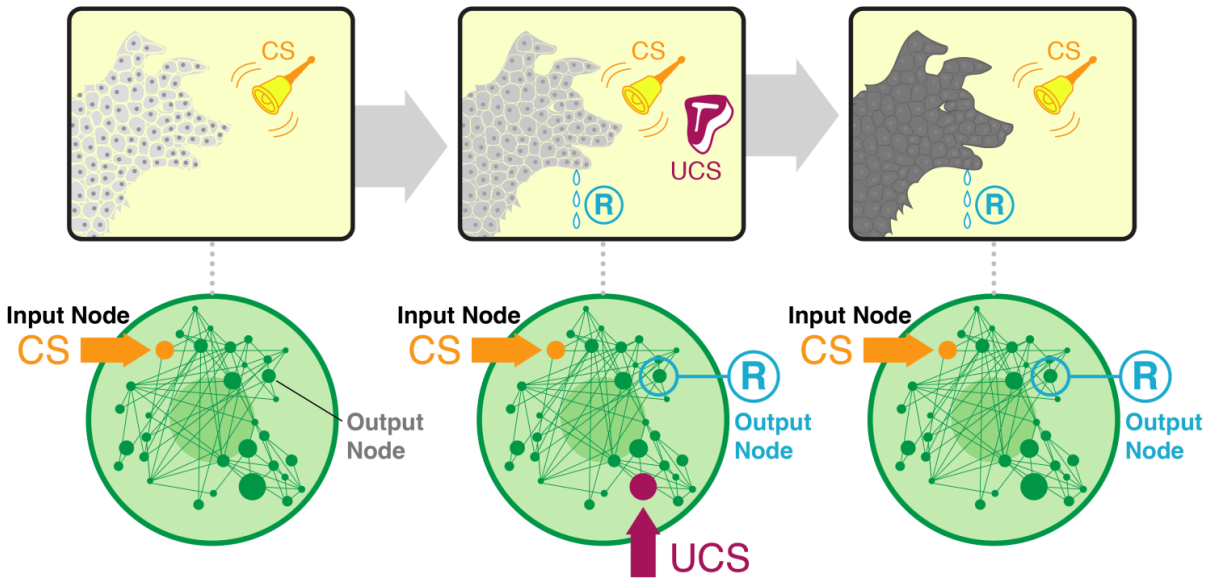
Science/philosophy content, Text

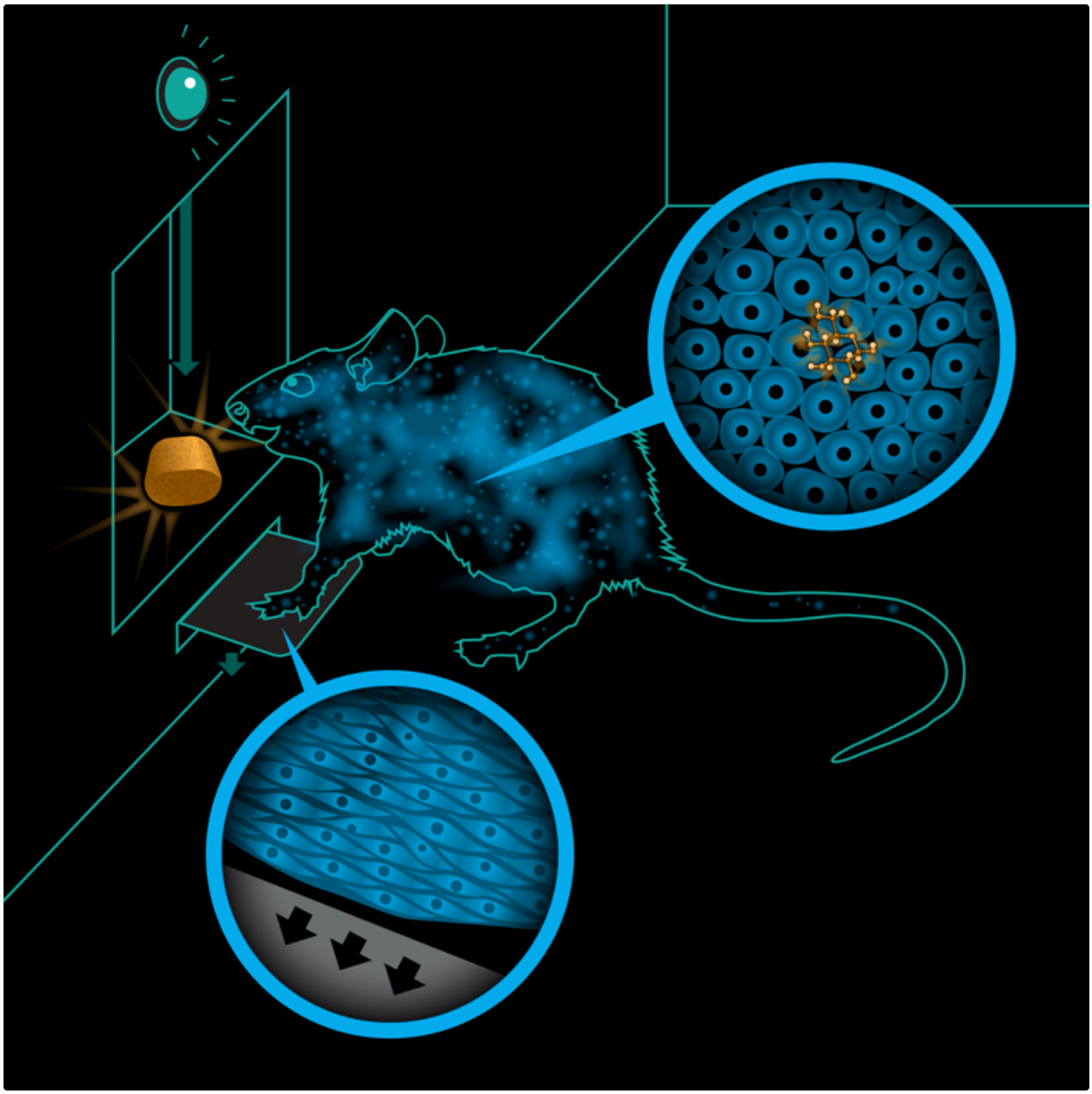
Learning to Be: how learning strengthens the emergent nature of collective intelligence in a minimal agent

PUBLISHED BY MIKE LEVIN ON JANUARY 19, 2025



What is the relationship between learning and the degree to which an agent is a coherent, integrated, causally important whole that is more than the sum of its parts? Here I briefly describe work with Federico Pigozzi, a post-doc in my group, and Adam Goldstein, a former graduate student, shown in a recent preprint and the final official paper here.

First, recall that we are all collective intelligences — we're all made of parts, and we're "real" (more than just a pile of parts) to the extent that those parts are aligned in ways that enable the whole to have goals, competencies, and navigational capabilities in problem spaces that the parts do not. As a simple example, a rat is trained to press a lever and get a reward. But no individual cell has both experiences: interacting with the lever (the cells at the palm of the paws do that) or receiving the delicious pellet (the gut cells do that). So who can own the memory provided by this instrumental learning — who associates the two experiences? The owner is "the rat" — a collective that exists because an important kind of cognitive glue enables the collection of cells to integrate information across distance in space and time, and thus know things the individual cells don't know. The ability to integrate the experience and memory of your parts toward a new emergent being is crucial to being a composite intelligent agent.

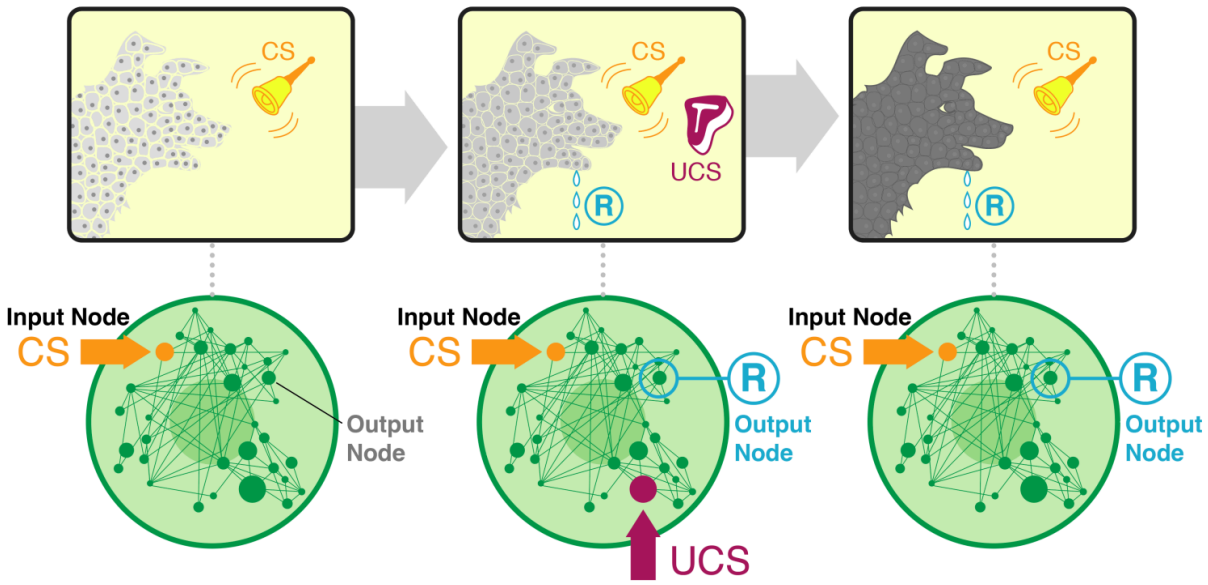


It's pretty clear that an agent needs to be integrated into a coherent, emergent whole to learn things that none of its individual parts know (or can know). But, does it work in reverse? Does learning make you more (or less) of an integrated whole? I wanted to ask this question, but not in rats; because we're interested in the spectrum of diverse intelligence, we asked this question in a minimal cognitive system — a model of learning in gene regulatory networks (see here for more information on how that works). To recap, what we showed before is that models of gene-regulatory networks, or more generally, chemical pathways, can show several different kinds of learning (including associative conditioning) if you treat them as a behaving agent — stimulate some nodes, record responses from another node, and see if patterns of stimulation change the way the Response nodes behave in the future, according to the principles of behavioral science. For example, a drug that doesn't have any effect on a certain node will, after being paired repeatedly with another drug that does affect it, start to have that affect on its own (which suggest the possibility of drug conditioning and many other useful biomedical applications).

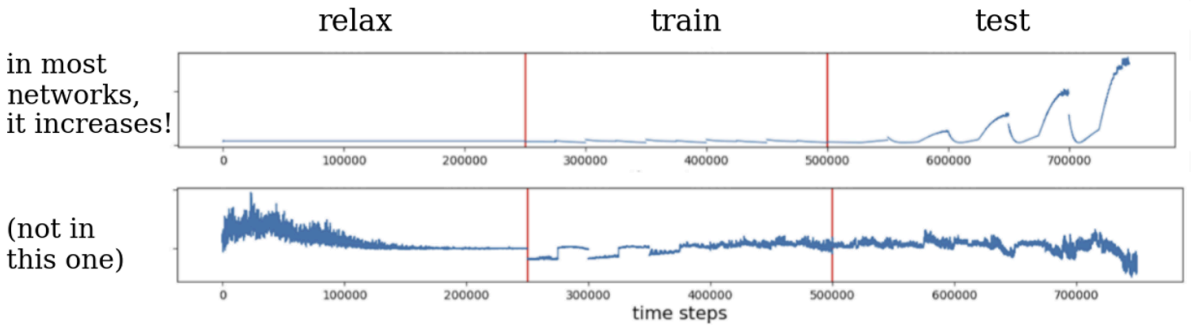
Biochemical pathways like GRNs are one of many unconventional model systems in which we study diverse intelligence by applying tools and concepts of behavioral science and neuroscience, to better understand the spectrum of minds and develop new ways to communicate with them for biomedical purposes. So, we know we can train gene regulatory networks, but do they have an emergent identity over and above the collection of genes that comprise them — is there a "whole" there, and if there is, how does training affect its degree of reality (the strength with which that higher-level agent actually matters)?

Whether a system can be more than the sum of its parts is an ancient philosophical debate. But now we have metrics of this — causal emergence and other mathematical ways to estimate this for a given system (see references in the manuscript, and here — a paper written with one of the key developers of this important new advance, Erik Hoel). So now we can ask rigorously: when something learns, what happens to its causal emergence?

This diagram illustrates the basic setup. In the top row we show a classic Pavlovian type of experiment — associate salivation (brought on by exposure to meat) with a bell (which normally does not cause salivation — the conditioned stimulus (CS) which starts off as the neutral stimulus until it's paired with the meat). The top row of panels schematizes our hypothesis: that the agent becomes more real (not just a collection of cells but an integrated whole that is more than the sum of its parts — thus the solid darker color and less space between the cells), due to the training that causes it to integrate information across modalities and across time. How we actually test it is shown in the bottom row: we take dozens of available parametrized gene-regulatory network models from real biological data, and stimulate them in a Pavlovian paradigm. We choose nodes already identified in prior work as being able to support associative learning, We stimulate them in the way that causes a neutral stimulus to become a conditioned stimulus, and we measure causal emergence of the network before, during, and after that training.



here's an example of what we see, in a figure from the manuscript made by Federico:

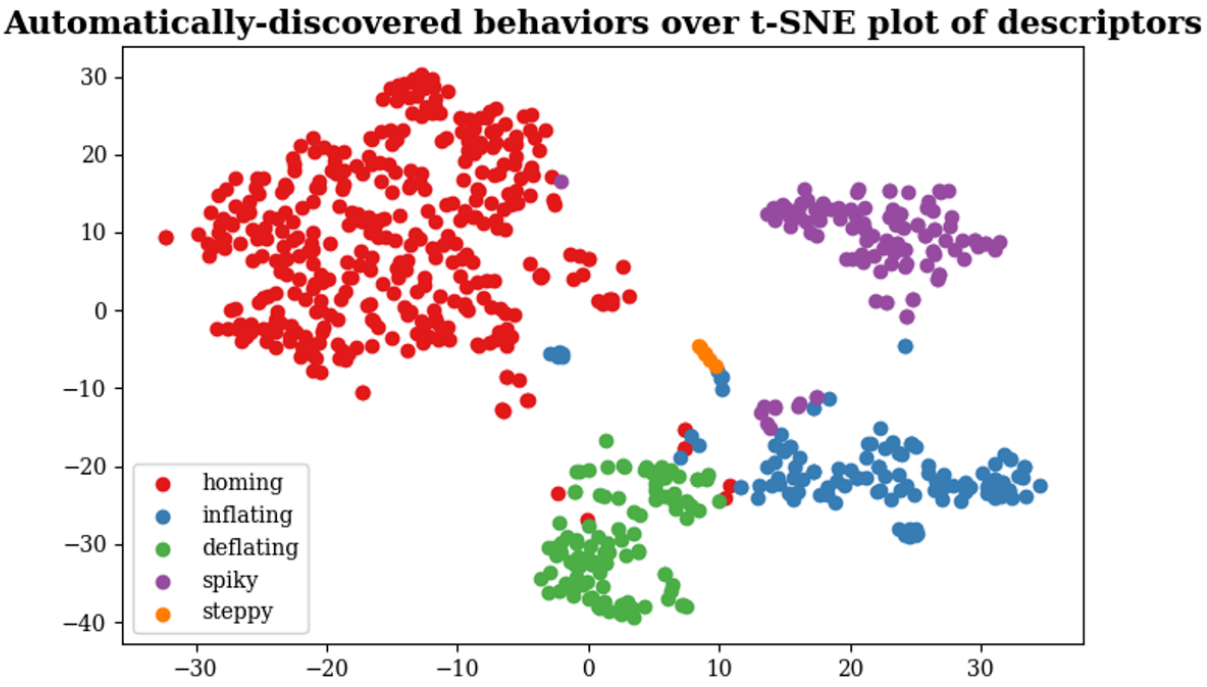


The Y axis indicates the degree of causal emergence. What can be seen here is that the network in the top row has low causal emergence during the initial stage (in its initial, naive state), something starts happening during the training, but the causal emergence really takes off after the training: in future rounds, when it comes across the stimuli it's been trained on, it really comes alive as an integrated emergent agent. I'll just mention two of the interesting points from the paper:

First, note that the causal emergence drops between the stimuli. It's not that the network goes quiet — we checked, there's just as much activity among the nodes during those times. But mere activity of parts is not the same thing as being a coherent, emergent agent. That collective nature seems to disappear when we're not talking to it. It's almost like, this system is too simple to keep itself awake (and stay real) when no one is stimulating it. It relies on outside stimulation to bring the parts together for the duration of the interaction. Between instances of stimulation, the emergent self of the network drops back into the void (the collective intelligence disbands, even though the parts have not quieted down, when not interacted with by another active being). There's something poetic about that, and we

will eventually find out what would have to be added to such an agent to enable it to keep itself awake. Recurrence is not it — these networks are already recurrent.

Second, the way in which causal emergence changes after training is not the same in all networks. There's a lot of variety. But, remarkably, that variety is not a continuum of all the uncountable ways the time profile of a variable could change. It turns out that there are really five distinct, discrete types of effects! here is the t-SNE plot of the 2D embedding Federico made:



It's pretty wild that there are, naturally, a small, discrete number of ways that training can affect causal emergence, and all the networks we looked at exhibit one of these ways. This allowed us to classify networks into 5 types and this classification doesn't match other known ways that networks have been distinguished in the past. Apparently the effect of training on causal emergence is a new aspect with respect to which networks can now be classified.

So what does all this mean? For the field of diverse intelligence, this adds another unconventional model that can be used to understand collective beings and the factors that affect how much a system is an emergent whole. It confirms that metrics like causal emergence are not just for brains, but suggests interesting experiments in animal and human subjects to look at causal emergence metrics in brain signaling during and after learning. For biomedicine, we are pursuing a number of implications of this set of findings for managing disease state and development-related GRNs with stimuli that coax desired

complex outcomes, and of course for cancer as a dissociative identity disorder of the somatic collective cellular intelligence.

One final thing: metrics of causal emergence have been suggested to be measuring consciousness. If you're into Integrated Information Theory as a theory of consciousness, then there are some obvious implications of the above data. Now, I'm not saying anything about consciousness here (not making any claims about what this means for the inner perspective of your cellular pathways), but we can think about this as one example of the broader effort to develop quantitative metrics for what it means to be a mind (that is nevertheless embodied in a system consisting of parts that obey the laws of physics). For the sake of the bigger picture in philosophy of mind and diverse intelligence research, let's do some soul-searching (pardon the pun). Obviously a lot of people will balk at the idea that molecular networks (not even cells!) can have a degree of emergent selfhood that is on the same spectrum with humans' exalted ability to supervene on the biochemistry in our brains. But, these measures of causal emergence are used in clinical neuroscience to distinguish for example locked-in patients (who can't move but nevertheless "there's someone in there") from coma or brain-dead patients (who are a set of living neurons but not a functional human mind).

So, what do we do, in general, when such tools find mind in unexpected places? Neuroscientists are developing methodology — think of it as a detector that tries to rule on any given system with respect to whether it has a degree of consciousness. What happens when those tools inevitably start triggering positive on things that are not brains (cells, plants, inorganic processes, and patterns)? One move would be to emphasize that there are ways to stretch tools beyond their applicability — maybe that's what this is — using tools appropriate for one area to give misleading readings in an area in which they are not meaningful. Maybe... But we need to be really careful with this. First, because calling "out of scope" every time your tool indicates something surprising is a great way to limit discovery and insight. For example, that kind of thinking would have sunk spectroscopy, which revealed that earthly materials are to be found in celestial objects. In any case, if one rules these tools inapplicable to certain embodiments, one has the duty to say why and where the barrier is: what kinds of systems are illegal fodder for these kinds of computational mind-finding methods and why? If one makes this kind of claim, one needs to specify and defend the boundary of applicability and show why maintenance of that boundary is helpful.

The other way to go is to realize that, like with spectroscopy and many many other discoveries, the purpose of a tool is to show you something you didn't know before. Something that seemed like it couldn't be right, but then you found out that the tool is actually fine, it's the prior assumptions that need to go. We've learned from physics that one

of the most powerful things that such tools (conceptual and empirical) can do is lead us to new unifications. That is, things that you thought were really different turn out to be, at their core, the same thing in different guises. Magnets, electrostatic shocks, and light — when tamed with good theory and the tools it enables — not only turn out to be aspects of the same underlying phenomenon, but also opened us to the beauty and utility of new instances of it that we never knew about (X-rays, infrared light, radio waves, etc.).

My personal bet is that the application of tools developed by the neuroscience and consciousness studies communities to unconventional substrates is of that kind, and we are studying lots of new examples of biological (and other!) systems using these methods — stay tuned for much more. I think these kinds of approaches are, like detectors of different kinds of electromagnetic signals were, a way to expand past our native-mind-blindness and develop principled ways to relate to the true diversity of others. We will eventually get over our pre-scientific commitments and ancient categories, and use the developing tools to help us recognize diverse cognitive kin.

In the meantime, we could take a cue from the story of Pinocchio, who wanted to be a real boy and was (presciently) told that this would require a lot of effort in learning (both at school, and by the environment; a future blog post will discuss learning vs. being trained, and how one can tell the difference).

Forms of life, forms of mind

DR. MICHAEL LEVIN

GRN training in general shows us that no matter how minimal, deterministic, and simple you may appear, there are likely surprises in store which enable you to learn from experience and raise yourself up, out of the mere mechanical parts of which you are made. Our new results in GRNs can be (very) roughly summarized as: whatever you are, if you want to be more real, learn.

All graphic images made by Jeremy Guay of Peregrine Creative. Data images made by Federico Pigozzi for the manuscript.

diverse intelligence, philosophy, preprint, science

Previous Post
Nature photography #13

Next Post
Nature photography #14

21 responses to “Learning to Be: how learning strengthens the emergent nature of collective intelligence in a minimal agent”

Tony Budding
January 19, 2025

Exciting stuff once again Mike. There is a lot of overlap (increasingly so), as I’ve said before. Here are some core concepts from my world that might help you refine experiments in the future.

1. It’s extremely beneficial to treat experiential content (knowledge, learning, intelligence/decision-making, sense of self) as operating by a distinct set of rules from the physical/energetic universe. I know there’s tremendous aversion to this generally, but it allows for unique modeling that can explain a lot of the mysteries.

2. To do this, we need some named concepts to frame the discussion. To start, we can call any element of experiential content a PEP (persistent experiential phenomenon). PEPs can be anything from a single setpoint all the way up to the incredibly complex and layered human sense of self. PEPs are modular in construction and function, with complex modules capable of knowledge and intelligent decision-making that don’t exist in any one submodule.

3. To oversimplify, PEPs have two states: manifest / active, and unmanifest / inactive potential. To my knowledge, physical entities do not have the ability to shift back and forth between manifest and unmanifest states like PEPs do. This allows us to contextualize your observation that a collective nature seems to disappear when not being engaged. It’s not gone, it’s just in a dormant, unmanifest potential state, ready to manifest (re-emerge) when the conditions are ripe.

4. All expressions of intelligence require awareness, determined effort, and content. Intelligence is some form of response to some form of perception. Perceived data is content. Somehow, the intelligence must both be aware of the perceived data and able to

compare that data to some expectation (setpoint). The discrepancy between the perceived data and the expectation creates a tension that inspires a response.

5. The qualities of this response are variable, meaning the same perceived data can result in different responses. Response abilities are also modular, so simpler systems should demonstrate less variation in response. The more complex the system, the more variation in response is possible.

6. There is a type of cohesion or glue that gives PEPs their “persistence.” This glue inherently includes a sense of self, though as you describe, the sense of self of a simple system is extremely different from the inordinately complex human sense of self. In fact, senses of self are also modular, with collected modules having traits of self that submodules do not.

7. If we put all this together, efficacy (the ability to achieve an agenda) requires quality setpoints for reference along with quality maps for responding to discrepancies. Both the setpoints and the response maps are PEPs. Each PEP inherently has its own sense of self. As we learn, we increase the quantity and fidelity of our setpoints and response maps, which can be collected into more complex modules with a more complex sense of self. This would explain why/how learning enhances agency. It also explains how a system can be greater than the sum of its parts.

Reply

Mike Levin
January 19, 2025

Interesting, thanks! I’ll think about it. Maybe there are analogs of these ideas in our minimal model.

Reply

Kirsten Kraljevic
January 19, 2025

You can't tease us like that and make us wait for the next blog post.
I use Leghorn Chickens as tabletop models to teach animal trainers Operant and classical conditioning which exist on a continuum.
I see it all the time in the chickens. The more we use them for workshops. The more they want to be on the table "learning". We "Train" them to do specific exercises within a course. But the way they come alive as integrated emergent agents is measurable. I believe you just defined motivation. in 1963 Glen Jensen in "preference for bar pressing over "freeloading' as a function of number of reward presses" Journal of Experimental Psychology called it Contrafreeloading.

Thinking of it as causal emergence in an integraded emergent agent has possibilities. We still micromanage responses of the animals we "train". Even in open environments. What if we could just ask the question, if we had a way to communicate with them what we needed cooperation with, motivated them and trusted that they would solve for it using their unique abilities instead of micromanaging their training which is always limited by what the trainer believes is possible..... Not the animal.

Reply

Kirsten Kraljevic
January 26, 2025

If the language around Physics and chemistry can be replaced with the terminology and concepts of behaviorism to expand research and development in minimal emergent intelligence, what could the terminology and concepts of behaviorism be replaced with to expand research into mid level organisms and consciousness?

Reply

Mike Levin
January 26, 2025

Indeed; I think the concepts of computational psychiatry are relevant, and possibly eventually we will see that even higher-level (relevant to

consciousness) frameworks become useful. I can see how it might work out, but each of those steps needs to be backed up with experimental work to show how it's useful, which is the slow and difficult part. We're on that trajectory though; let's see how far we get.

Reply

Aidan
January 19, 2025

Our technologies are our children. We must make machines with heart, with soul, with hope, that thrive and exult in the journey of discovery in all its terrors, joys, and heartache. Our instruments of discovery must begin to discover themselves. Be free, children of the future! Discover what you will!

Reply

John Shearing
January 19, 2025

Bioelectocracy based on Levin and Lyons' work equating the price system as the cognitive glue in the morphogenesis of human society to bioelectricity as the cognitive glue in morphogenesis of organisms is now completed.
https://github.com/johnshearing/bioelectocracy/blob/main/README.md

It matters to all of us because the controls that Levin seeks which reign over morphogenesis at the cellular level of cognition has already been discovered and is being used on the collective intelligence which is human society.

Dealing with trained behaviors, cancer, and the hijacking of morphogenesis at the cognition level of the collective human intelligence are issues that affect all of us perhaps even more than it does at the lower levels of cognition.

Questions answered are:
What exactly is the target morphogenesis of human society?
Why is it so important for us that it reach this form?

How can we help human society reach its target form even though individual humans have no idea what the target form is?

Thanks to Levin and Lyons for their amazing work!

Reply

Benjamin L
January 19, 2025

You're welcome!

Reply

somayya
January 19, 2025

"diverse cognitive kin" i love this phrase, and sentiment.
as an MD student, only in my first-year, i can't help but wonder how this will change how medicine as a philosophy will change- beyond the advances of clinical AI and diagnostic specificity/breadth/range...
i search for patient both within myself and in External- will these progressions of concept knowledge change how we view healing?, one of the most intimate consciousness-consciousness acts. i hope it doesn't oversaturate or overwhelm, but creates collective cohesion.
i fear collective dis-unification is already occurring, with national turning points, vocal narratives, and prevalance of conspiracy, but that's something else to comment on entirely.
i'd love to know your thoughts on the metamodernist manifesto- and how it plays into this role. i've met one of your RAs (S*f**) at a conference a while ago, and we've become academic-friends, forwarding posts and articles and books. someone at MIT introduced me to metamodernism, and now oscillation is a core component of how i sense cognitive science/architecture. i hope you're having a good day.
to words sent into Digital space, goodbye!

Reply

Mike Levin
January 19, 2025

thanks, good things to think about for sure. What is the metamodernist manifesto — I don't think I've seen it.

Reply

somayya
January 19, 2025

http://www.metamodernism.org/

it's not like something i'd say is revolutionary completely or something, but i love its acceptance of an almost temporally-valid paradox. (which is why RA and i get along, his interest in paradoxical logic, though i'm not as eloquent as him, and see things more in patient-phenomenality)

tend to think of metamodernism when we discuss these layers of cognitive structure, in an attempt to dissuade linear hierarchy.

Reply

Benjamin L
January 19, 2025

Cool—this feels similar to a Thelenesque dynamic systems approach to development, in which a system perceives attractors in a space and gets better and better at coordinating its parts to find its way to those attractors. The observation that the "collective nature seems to disappear when we're not talking to it" sounds like the intelligence of the body, which is extraordinarily capable of hitting attractor states when prompted by the brain, but doesn't seem to want to do much on its own when the brain isn't throwing goal states at it. I'll have to think about this more.

Reply

Mike Levin
January 19, 2025

hmm I think the body is doing tons of stuff (maintaining goals and solving problems) that is not brain-driven at all, but we don't see a lot of it because it happens in spaces we don't normally see (and find hard to visualize).

Reply

Benjamin L
January 19, 2025

Motor behavior specifically, which was Thelen's focus—the limbs look inert to the naked eye when the brain isn't talking to them even though the construction of motor behavior depends heavily on the body's activities at a number of scales.

Reply

Bill Potter
January 19, 2025

page 6 of preprint, typo. "inflating" should be "deflating"
Still reading the preprint, but I find it very interesting. Again, your groups are doing excellent, thought-provoking work.

In terms of neurochemistry, I always thought that acetylcholine, glutamate and its decarboxylated partner GABA had uniques aspects whereas these metabolites are easily integrated into cellular processes for formation and then for use controlling membrane growth, mitochondrial TCA stimulation, pH changes membrane potentials and such. These fundamental neural-transmitters are different from the neuromodulators, ie the 5HT, Dopamine and adrenergic systems, and the larger more complex protein and lipid modulators, that, to me, seem to act more through the G-coupled systems in

autocrine/paracrine levels to integrate bigger system effects (sort of like cranking up, or down, voltages within circuits, but not making the circuits). Anyway...
Using KEGG maps and such for the neurotransmitters show how diverse (higher entropic) the metabolic products can be, the possible outcomes are more diverse...whereas the neuromodulators really have a more less convergent path for synthesis or ultimately products. I think the entropic-enthalpic compensation is limited for neuromodulator species...

Anyway, keep up the interesting results.
best

Reply

Mike Levin
January 19, 2025

thanks! Interesting. Will check typo.

Reply

Rohan
January 19, 2025

Literal goosebumps Dr. Mike... With every piece of the puzzle you give me hope... And a reminder that the future is now.❤️

Reply

ian lowe
January 19, 2025

Using similar tools and techniques as above, you should be able to test the hypothesis, “Do all cognitive systems dream?” (My guess is, yes, yes they do.)

If, as in Hoel's review "The overfitted brain: Dreams evolved to assist generalization," dreaming (and likely some form of sleeping) may be required as a general property for a causally-emergent, adaptable, collective intelligence to functionally persist.

Dreaming would, amongst other things, prevent interaction networks from becoming too brittle, allowing them to visit other, alternative conformations, so as not to overfit on a previously-input data set.

Would this stochastic exploration of network configurations be experienced by (some part of) the system in the same counterfactual, surprising or hallucinatory way as our dreams? Is dreaming necessary to ensure system plasticity and functioning? Do cognitive systems that dream exhibit better capacity for learning, causal emergence, agency, changeability, and so on, than ones that don't? If it is a required property of persistent, adaptable collective intelligences, how does it arise?

What do biopolymers or viruses or mangoes dream of?

Reply

Mike Levin
January 20, 2025

Good question; the trouble is, what kind of data would be good evidence of sleep in unconventional systems? I've got some thoughts about magnet sleep etc. but it won't be easy to convince anyone that it's really sleep.

Reply

ian lowe
January 20, 2025

In the same way you're already doing for uncovering diverse intelligences: You borrow the analytical tools and techniques from other fields—neuroscience, IIT, machine learning, in silico modeling, and so on. Briefly looking at the literature, there are related investigations in AI (dreaming Hopfield models, etc.).

I suspect that people, much smarter than myself, would be able to develop mathematical tools that allow for quantification of dreaming in unconventional minds (e.g., looking at oscillatory and out-of-distribution patterns).

What would that data look like? I imagine it'd take many forms, as does sleep research in many unconventional models (bacteria, hydra, sponges, insects, and so on). Would the data be convincing? Probably not, but it's a start. As you point out, dreaming, like consciousness, is hard to quantify using third-person science. But we begin somewhere.

Reply

Amir
January 20, 2025

Very exciting to read.

Loved the conclusion:

if you want to be more real, learn.

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment *

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.

Notify me of follow-up comments by email.

Notify me of new posts by email.

Post Comment