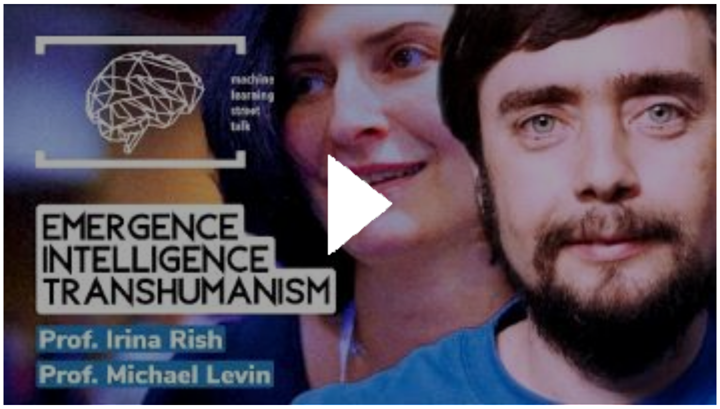
Transcript

Friday, March 3, 2023 1:18 PM

#102 - Prof. MICHAEL LEVIN, Prof. IRINA RISH - Emergence, Intelligence, Transhumanism

Machine Learning Street Talk



Michael Levin - Google Scholar

Michael, it's an absolute honor to meet you. You've got a very interesting background. Now I've discovered your work when we did a show on emergence and I was looking into graph cellular automata actually and I was looking into that the the CNN (Convolutional Neural Network) version which did this concept called Morphogenetic Engineering, which is this idea that you. Almost transgress. Uh, ladders. You know, like levels of the emergence ladder by describing something at the microscopic scale and then getting this kind of emergent global coherence. And in that particular case, it kind of emerged as the shape of a gecko. I'm not sure if you're familiar with that particular thing, but. So emergence is fascinating, but coming at it from your angle in biology, how do you see the interplay between what you do and artificial intelligence?

Well, thanks very much. I'm very happy to be here and have the chance to talk to you about these things. Other things, I think that intelligence is baked in basically at the very bottom of **the multiscale architecture** that we have in living organisms. And so what I think is very powerful is this interplay where we can of course use AI and machine learning to try to understand the biology better. And at the time we can take all these unconventional examples of intelligence and cognition that we find all the way down to molecular networks and use them as inspirations for things that that we build, right? We either hybrid systems or fully engineered systems. And I think one of the key things to say from my perspective (of molecular biologist) is that emergence is not the whole story.

So I think emergence is very powerful and emergence certainly happens and we have many, many, many scenarios in which highly parallel active implementation of local simple rules gives you some kind of complex emergent outcomes. So we certainly see that. But I think the magic of biology isn't just that I think the real magic of biology is that at every level these are not, uh, sort of feed forward emergent processes where you follow the rules, something comes out and you know there you go. Whatever comes out comes out. I think that the key to understanding the power of biology is that **all of these things** at all scales are **closed loop goal directed** in the cybernetic sense. Not in the magical sort of supernatural sense, but in the in the cybernetic sense. Goal directed agents that are able to detect error from specific set points and are doing their best in terms of using energy and sometimes actually very clever. Kinds of policies to achieve specific goals and specific problem spaces. And I think scaling that up gets us to gets us much farther than emergence alone.

Interesting. But I read a book by Douglas Hofstadter called The Strange Loop

```
@book{hofstadter2007strange,
  title={I am a strange loop},
  author={Hofstadter, Douglas R},
  year={2007},
  publisher={Basic books}
}
```

Abstract::

- Can thought arise out of matter? Can self, soul, consciousness, “I” arise out of mere matter? If it cannot, then how can you or I be here? I Am a Strange Loop argues that the key to understanding selves and consciousness is the “strange loop”—a special kind of abstract feedback loop inhabiting our brains. The most central and complex symbol in your brain is the one called “I.” The “I” is the nexus in our brain, one of many symbols seeming to have free will and to have gained the paradoxical ability to push particles around, rather than the reverse. How can a mysterious abstraction be real—or is our “I” merely a convenient fiction? Does an “I” exert genuine power over the particles in our brain, or is it helplessly pushed around by the laws of physics? These are the mysteries tackled in I Am a Strange Loop, Douglas Hofstadter's first book-length journey into philosophy since Gödel, Escher, Bach. Compulsively readable and endlessly thought-provoking, this is a moving and profound inquiry into the nature of mind.

and he has this idea of all sorts of interesting causal relations between the scales. And this is something that I find it a bit difficult to get my head around. So you know the mind, for example, is an emergent phenomenon. Then we have this thing called agency and I direct my hand and I tell my hand to move. But if everything does emerge from the lower level domain, how does that work? How do you get these feedback mechanisms?

Well, a couple of things. Uh, first of all, you know this this notion of reducibility to the lower level. So for a really long time for centuries this was a philosophical debate and that people could argue there were reductions and they would say well ultimately everything is reducible and people would say not out this strong emergence and things happened at the top. The amazing thing to me is that this debate has been actually moved from the area of philosophy to the area of mathematics and rigorous science by people like Giulio Tononi and Erik Hoel (found his name in bibliography of Tononi and ConnectedPapers, also [2003.13075] Causal Emergence in Discrete and Continuous Dynamical Systems (arxiv.org)), who have actually produced new advances in information theory where you can actually calculate which level of your system does the most work.

Grasso, M., Albantakis, L., Lang, J.P. et al. Causal reductionism and causal structures. Nat Neurosci 24, 1348–1355 (2021). https://doi.org/10.1038/s41593-021-00911-8

Abstract::

- Causal reductionism is the widespread assumption that there is no room for additional causes once we have accounted for all elementary mechanisms within a system. Due to its intuitive appeal, causal reductionism is prevalent in neuroscience: once all neurons have been caused to fire or not to fire, it seems that causally there is nothing left to be accounted for. Here, we argue that these reductionist intuitions are based on an implicit, unexamined notion of causation that conflates (merges) causation with prediction. By means of a simple model organism, we demonstrate that causal reductionism cannot provide a complete and coherent account of ‘what caused what’. To that end, we outline an explicit, operational approach to analyzing causal structures.

Hoel, E.P., Albantakis, L. and Tononi, G., 2013. Quantifying causal emergence shows that macro can beat micro. Proceedings of the National Academy of Sciences, 110(49), pp.19790-19795.

Abstract::

- Causal interactions within complex systems can be analyzed at multiple spatial and temporal scales. For example, the brain can be analyzed at the level of neurons, neuronal groups, and areas, over tens, hundreds, or thousands of milliseconds. It is widely assumed that, once a micro level is fixed, macro levels are fixed too, a relation called supervenience. It is also assumed that, although macro descriptions may be convenient, only the micro level is causally complete, because it includes every detail, thus leaving no room for causation at the macro level. However, this assumption can only be evaluated under a proper measure of causation. Here, we use a measure [effective information (EI)] that depends on both the effectiveness of a system's mechanisms and the size of its state space: EI is higher the more the mechanisms constrain the system's possible past and future states. By measuring EI at micro and macro levels in simple systems whose micro mechanisms are fixed, we show that for certain causal architectures EI can peak at a macro level in space and/or time. This happens when coarse-grained macro mechanisms are more effective (more deterministic and/or less degenerate) than the underlying micro mechanisms, to an extent that overcomes the smaller state space. Thus, although the macro level supervenes upon the micro, it can supersede it causally, leading to genuine causal emergence—the gain in EI when moving from a micro to a macro level of analysis.

It is no longer up for philosophical debate.
Links:: PPI-SyEN-114-July-2022.pdf (ppi-int.com)
Abstract:

- A deeper understanding of emergence is crucial to the field of systems engineering because systems are designed/created to achieve emergent system-level behaviors.
- Emergence, a 2500-year-old topic, has been a source of debate in philosophy, system science and complexity science, but extensive debate has not yielded a precise characterization of emergence that has general acceptance.
- The paper focused on practical implications of the open questions concerning emergence.
- The role of an explicit observer is essential for understanding and handling emergence.
- Emergence and complexity share a common trait, i.e., the amount of information required to describe a system.
- Axelsson raised four questions concerning emergence about which there are still significant philosophical controversies:
- What Phenomena Should Be Called Emergent?
- Are Emergent Phenomena Predictable?
- Can System-Level Phenomena Affect Element-Level Phenomena?
- Must There Be an Observer for an Emergent Phenomenon to Exist?

You can actually do the calculation. So there is literally a software toolkit that you can use and sometimes you will find out that yes, indeed it is reducible to the lower level. And if for other systems you find out that actually know the higher levels do more work, more causal work than the lower levels. And this has been obvious in biology. I think for a really long time, because we as scientists and engineers other organisms and in fact the evolutionary process itself uses higher level control knobs, exploits them very significantly, the kind of things that we for example, studies membrane voltage. So resting potential is a kind of aggregate coarse graining of the positions of the individual ions. But you get much further in terms of regenerative medicine and other applications. If you track the voltage, not the molecular details of the ions, it actually helps you do new experiments, produce therapeutics by tracking. This high level thing that that are reductionist might say well doesn't even exist in the first place.

So I think these higher levels absolutely have causal power and the way that I would put this strange loop idea and I think you know Hofstadter has many kind of ingenious concepts in these various books the way the way I would put this strange look. He is like this. Many of the things we are interested in memory, cognition. All these kinds of things in my framework I have a particular framework for thinking about these things and my framework are observer dependent. So when you have a particular system the claim that it had occupied some level along this this continuum of agency. And I do think it's a continuum. It's not binary. The claim that it occupies some particular place along that continuum is an observer relevant relative claim. In other words, it's not an objective fact about it. It's here's an observer. That observer can formulate some particular model of this thing as a cognitive agent. At some level of sophistication, there's a different observer that has a different model, and each of them can interact with that system well or poorly, to the extent that their model allows. And so where you get the strange loop is that the observer, which is necessary to define all of you know things like problem spaces and the goals. These are all observer dependent. The observer might be the system itself. So when you get the real boot up of agency comes when the system becomes its own observer and starts to make internal models of its own parts and models that help it control itself. And it's all in its own lower levels when the high levels start to control the lower levels.

And so that's somewhat reminiscent of Karl Friston. He has his free energy principle and the concept of the Markov blanket and being able to predict external states probabilistically. But yeah, there there's two things you said there. So first of all, this notion of things at different levels of the emergence ladder doing different amounts of work, and that suggests to me a kind of form of strong emergence which is an affront on physicalism. So I wondered if you could pick up on that and the other thing you said in terms of this observer relative thing that gets rather to this kind of ultra relativistic or even you know like in the Wittgenstein sense of or the Putnam sense that you can take any computation represented in any physical system and the kind of the meaning of that computation is in its use, so it doesn't really have any agency or balance in of itself.

Yeah, yeah. So, I think that the first question first, I guess it it's only in the front on physicalism if you assume from the beginning that physicalism has to be ultimately reductive. So I don't believe that to be the case. And so I'm OK with a certain kind of physicalism. That does real justice to the fact that I think you know Ian McGilchrist says this. He says that you know a lot of physical is under a really underestimate matter. Matter is amazing and it does some amazing things. And when you say wow, that can't be, you know matter that you know that that can't just be matter. You're not understanding what matter actually capable of and I think if we take seriously the fact that it can be matter and also higher levels of organization can do work that is not apparent at lower levels.

Regarding the second part. I think that's totally compatible. I think that's pretty necessary. I think that gets you very far actually because a lot of times we get stuck in these pseudo problems when you try to make objective claims about this area, you get into a real problems where that that resolve once you specify from whose vantage point are these things. Or false. And so you get into this observer relativity. But I think it's really, really critical to say that the system itself is also a valid observer. So it's not just that, well, you have no meaning of your own. It's up to us to interpret you. I mean, we can interpret you, but you can also interpret yourself. And you, you're not dependent on external observers to define these things for you. You are also a bonifide. Whatever. But I do think, and this is work that we've done with Chris Fields, the role of an observer in all of these things, I think in the end this is really critical.

```
@article{Fields2022TheFE,
  title={The free energy principle induces neuromorphic development},
  author={Chris A. Fields and Karl John Friston and James F. Glazebrook and Michael Levin and Antonino
Marcian{\`o}},
  journal={Neuromorphic Computing and Engineering},
  year={2022},
  volume={2}
}
```

Abstract:

- We show how any finite physical system with morphological, i.e. three-dimensional embedding or shape, degrees of freedom and locally limited free energy will, under the constraints of the free energy principle, evolve over time towards a neuromorphic morphology that supports hierarchical computations in which each ‘level’ of the hierarchy enacts a coarse-graining of its inputs, and

> dually, a fine-graining of its outputs. Such hierarchies occur throughout biology, from the architectures of intracellular signal transduction pathways to the large-scale organization of perception and action cycles in the mammalian brain. The close formal connections between cone-cocone diagrams (CCCD) as models of quantum reference frames on the one hand, and between CCCDs and topological quantum field theories on the other, allow the representation of such computations in the fully-general quantum-computational framework of topological quantum neural networks.

Wonderful Irina. Thank you so much for joining us today.

Thank you. And uh, yeah, really good to see both of you. I'm also looking forward to seeing Michael's talk at the workshop that we're running next week at the AAI conference. So, uh, yeah, I tried a bit late. So we're a specific questions you were kind of causing or I just called the last one. Anything?

Well, so we just did. But I'm really interested in the ranks of the emergence ladder, and how different ranks do different amounts of work. They have complex causal relationships between them. Is it in a fun an affront on physicalism and reductionism. And we're also talking about this very slippery notion of agency. And where do you draw the boundary? And if you can draw a boundary?

Yeah, I mean I yeah, I was saying that I don't even think that it's primarily an emergence continuum. I think that it's a well like I call it a persuadability continuum, but it's basically it, it really is a continuum of agency and I don't think there are any sharp boundaries. I think sharp categories are what we as observers bring to an underlying continuous phenomenon, and we can we can try to impose categorical boundaries on things that make it easier for us to do specific things, but the underlying phenomena, I think are completely continuous.

Do you agree with that? There is a notion of transitions in certain metrics of behavior. Uh, I'm talking purely from kind of practical examples saying behavior of large scale, say systems, neural networks. So we know and that's been observed recently so. That is, say scaling the amount of data or model, you might for a while still not be able to do certain task well, and then at some particular kind of critical size something happens and models suddenly get groups. So this grouping behavior have been observed over and over the past like year. So in the large scale machine learning models which is essentially kind of emergence of certain ability. Well, I mean it is still continuous, but there is a change in the slope and the change can be quite dramatic. So in a sense it also resonates with other examples where transitions were happening in AI but not necessarily in machine learning. Going all the way back to good old like satisfiability problems or constraints, distraction problems in general. Even other big hard problems like discrete problems where you might have a network of variables or nodes that can take different values and you're trying to basically increase their density in terms of number of constraints and see what happens is the property of being satisfiable problem to unsatisfiable changes continually, but that probability experiences sharp drop at certain critical value of say number of constraints and system divided by the number of variables.

So this constrictiveness plays a role of critical parameters similar to critical parameters such as temperature or pressure in physical systems which undergo state of matter transitions whatever the water to solid and vice versa. And similar like for magnetics when the temperature increases and the ferromagnetic basically stops being magnetic and that analogy was made in some cases quite clearly with those AI problems that their properties and those transitions. And then it's also kind of relates closely to like random graph theory. There is a famous old theorem by Friedrich Kalai that if you consider random graph and say peace probability that there is an edge in that graph and that probability keeps increasing. When it's zero, the graph is fully disconnected, no links at all. If it's one, everything is connected, it's complete graph in the middle. Say if you look at the connectedness of a graph as a property, what's the probability that from any point you can get to any other point and there is a pass? That's what means group is connected. That property is probability again sharply goes from zero to one around certain critical value of that parameter. And then any other property, what was the duty of that result? Any property doesn't have to be connectivity. That's monotonic. In addition, will have this transition, so though it's continuous change of the shape or with this highly nonlinear.

Super interesting. And of course there are many phenomena in physics and in math that show this kind of nonlinear change and that's fine. I'm interested in how sharp it really is and in particular in the cases that you talked about because what happens in biology is that people will sort of they look at the adult. And or let's say a human and they say, OK, this human has true cognition and, you know, real, you know, hopes and dreams and whatnot. Then they look at the at the one cell, you know, one cell egg at fertilization. They say, wow, that's just the chemical system. And so the theory is that at some point which people assume that at some point there's some sort of phase transition and a lot of people talk about this phase transition.

If you really dig into this and ask them to specify so what is that phase transition? So in biology it at least in that in that and the same thing happens on the evolutionary time scale, I have never heard a plausible story of where is it so, so at this moment boom. That thing moved from here there. Oh, now we have this like crazy, you know, large transition and it it's actually quite smooth, even though people tend to assume there must be some kind of phase transition. And so I wonder just out of curiosity, because I'm not super familiar with the examples that you're talking about, if you were to zoom in. Right. So let's just magnify it so, so clearly you sort of signature network. We're sort of zooming in. How sharp is it actually? Is there an atomic operation where I add one edge to this graph and bam, now I'm somewhere else. Like, how sharp is it really?

Just like phase transitions in statistical physics, uh, the sharpness increases with the size of the system. Say you have any finite size graph transition is while your sigmoid when you grow the size of this graph, say you grow that satisfiability problem or whatever. The transition becomes sharper and sharper. So in the limit when the system is incident that basically, the exponent diverges, so it is getting sharper. In physical results people were kind of doing in constraint selection problem satisfiability problems while ago and random graphs that was happening indeed. So as essentially in similar physical system sharpness depends from there. Size of the model and the number of units, particles, variables, nodes. While the critical parameter is another thing, and at the same value of the critical parameter, like for example ratio between constraints and variables, the transition happens and then all the pictures that did were that you see sharp or increasing number of variables in the that problem so in a sense, it behaved pretty much just like it was supposed to behave with certain physical phenomena. Phases of states of matter, for example, or paramedics, like Ising models.

So it's not always like that. Try to exercise caution when they see rapid improvement in performance of machine learning models, for example GPT3 on particular tasks like arithmetic. It was observed that while on other tasks and just on measuring how the loss function during training behaves. It's used to be quite common to observe power laws. And that was a famous Jared Kaplan (Googled through "Jared + openAI") and his colleagues from open AI paper work that basically the scaling laws, behavior in the language models, and then they found other modalities anyway, so it's a whole field.

```
@inproceedings{brown_language_2020,
	title = {Language {Models} are {Few}-{Shot} {Learners}},
	volume = {33},
	url = {https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-
	Paper.pdf},
	booktitle = {Advances in {Neural} {Information} {Processing} {Systems}},
	publisher = {Curran Associates, Inc.},
	author = {Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan,
	Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and
	Askell, Amanda and Agarwal, Sandhini and Herbert-Voss, Ariel and Krueger, Gretchen and
```

Henighan, Tom and Child, Rewon and Ramesh, Aditya and Ziegler, Daniel and Wu, Jeffrey and Winter, Clemens and Hesse, Chris and Chen, Mark and Sigler, Eric and Litwin, Mateusz and Gray, Scott and Chess, Benjamin and Clark, Jack and Berner, Christopher and McCandlish, Sam and Radford, Alec and Sutskever, Ilya and Amodei, Dario},
editor = {Larochelle, H. and Ranzato, M. and Hadsell, R. and Balcan, M. F. and Lin, H.},
year = {2020},
pages = {1877--1901},
}

Abstract::

- We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

arXiv:2001.08361 [cs.LG]

Abstract::

- We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.

Laws kind of ruled for a year or so and they had people started digging into more complex behaviors because for certain tasks like as I said arithmetic or any discrete iteration that can be described as a table. People discovered that, say GPT3 may not do that task well until certain size and after certain size of the model it rapidly improves its performance.

Another example was that so-called croaking paper. Also it was a nice thing I clear workshop paper I think was from openAI where they by mistake just left model running and training on the data multiple pokes, so they were not really scaling model or data, but just let it run. So compute was growing and it wasn't doing well. The accuracy was kind of low and at some point it's certainly later on day. So when they looked at their process, they forgot to kill. They looked at it, kind of grouped the problem and at some point it suddenly improved first training loss and then test loss.

Maybe this? But I'm not sure.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y. and Li, Y., 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Abstract::

- We present LaMDA: Language Models for Dialog Applications. LaMDA is a family of Transformer-based neural language models specialized for dialog, which have up to 137B parameters and are pre-trained on 1.56T words of public dialog data and web text. While model scaling alone can improve quality, it shows less improvements on safety and factual grounding. We demonstrate that fine-tuning with annotated data and enabling the model to consult external knowledge sources can lead to significant improvements towards the two key challenges of safety and factual grounding. The first challenge, safety, involves ensuring that the model's responses are consistent with a set of human values, such as preventing harmful suggestions and unfair bias. We quantify safety using a metric based on an illustrative set of human values, and we find that filtering candidate responses using a LaMDA classifier fine-tuned with a small amount of crowdworker-annotated data offers a promising approach to improving model safety. The second challenge, factual grounding, involves enabling the model to consult external knowledge sources, such as an information retrieval system, a language translator, and a calculator. We quantify factuality using a groundedness metric, and we find that our approach enables the model to generate responses grounded in known sources, rather than responses that merely sound plausible. Finally, we explore the use of LaMDA in the domains of education and content recommendations, and analyze their helpfulness and role consistency.

Now there are multiple groups very actively kind of studying this phenomenon. We got very interested back in, I guess September 2021 and submitted the paper just now.
arXiv:2302.01067 [cs.AI]
Abstract::

- The field of compositional generalization is currently experiencing a renaissance in AI, as novel problem settings and algorithms motivated by various practical applications are being introduced, building on top of the classical compositional generalization problem. This article aims to provide a comprehensive review of top recent developments in multiple real-life applications of the compositional generalization. Specifically, we introduce a taxonomy of common applications and summarize the state-of-the-art for each of those domains. Furthermore, we identify important current trends and provide new perspectives pertaining to the future of this burgeoning field.

Checkmarx Group worked on that and that they have some papers now.

arXiv:2004.03216 **[cs.CR]**

Abstract::

- The delivery of a framework in place for secure application development is of real value for application development teams to integrate security into their development life cycle, especially when a mobile or web application moves past the scanning stage and focuses increasingly on the remediation or mitigation phase based on static application security testing (SAST). For the first time, to the author's knowledge, the industry-standard Open Web Application Security Project (OWASP) top 10 vulnerabilities and CWE/SANS top 25 most dangerous software errors are synced up in a matrix with Checkmarx vulnerability queries, producing an application security framework that helps development teams review and address code vulnerabilities, minimise false positives discovered in static scans and penetration tests, targeting an increased accuracy of the findings. A case study is conducted for vulnerabilities scanning of a proof-of-concept mobile malware detection app. Mapping the OWASP/SANS with Checkmarx vulnerabilities queries, flaws and vulnerabilities are demonstrated to be mitigated with improved efficiency.

Jacob Steinhardt from Berkeley. They just had the accepted paper at the clear.

arXiv:2302.12349 **[cs.LG]**

Abstract::

Specifying reward functions for complex tasks like object manipulation or driving is challenging to do by hand. Reward learning seeks to address this by learning a reward model using human feedback on selected query policies. This shifts the burden of reward specification to the optimal design of the queries. We propose a theoretical framework for studying reward learning and the associated optimal experiment design problem. Our framework models rewards and policies as nonparametric functions belonging to subsets of Reproducing Kernel Hilbert Spaces (RKHSs). The

> learner receives (noisy) oracle access to a true reward and must output a policy that performs well under the true reward. For this setting, we first derive non-asymptotic excess risk bounds for a simple plug-in estimator based on ridge regression. We then solve the query design problem by optimizing these risk bounds with respect to the choice of query set and obtain a finite sample statistical rate, which depends primarily on the eigenvalue spectrum of a certain linear operator on the RKHSs. Despite the generality of these results, our bounds are stronger than previous bounds developed for more specialized problems. We specifically show that the well-studied problem of Gaussian process (GP) bandit optimization is a special case of our framework, and that our bounds either improve or are competitive with known regret guarantees for the Matérn kernel.

So anyway, people very much got interested in growing behavior. Not necessarily this specific example, but in general. When parameter of interest, which might be critical increases on the X axis. It could be a computed data size, model size, perhaps something else, and then you have whatever performance metric on Y axis. It could be accuracy, could be loss, could be something else and trying to understand when those sharp changes may happen and how sharp they are. My collaborators with the first was the recital,
Ethan Caballero and 2nd order basically trying to feed all these strange behaviors with one functional form. It's so-called broken neural scales paper because it's much richer behavior than previously assumed, just power loss.

arXiv:2210.14891 [cs.LG]
Abstract::

> We present a smoothly broken power law functional form that accurately models and extrapolates the scaling behaviors of deep neural networks (i.e. how the evaluation metric of interest varies as the amount of compute used for training, number of model parameters, training dataset size, or upstream performance varies) for various architectures and for each of various tasks within a large and diverse set of upstream and downstream tasks, in zero-shot, prompted, and fine-tuned settings. This set includes large-scale vision, language, audio, video, diffusion generative modeling, multimodal learning, contrastive learning, AI alignment, robotics, out-of-distribution generalization, continual learning, arithmetic, unsupervised/self-supervised learning, and reinforcement learning (single agent and multi-agent). When compared to other functional forms for neural scaling behavior, this functional form yields extrapolations of scaling behavior that are considerably more accurate on this set. Moreover, this functional form accurately models and extrapolates scaling behavior that other functional forms are incapable of expressing such as the non-monotonic transitions present in the scaling behavior of phenomena such as double descent and the delayed, sharp inflection points present in the scaling behavior of tasks such as arithmetic. Lastly, we use this functional form to glean insights about the limit of the predictability of scaling behavior. Code is available at this https URL

And those sharp transitions are not so sharp transitions, but transitions may happen anyway, so it's a bit maybe off tangent, it's about is there a one functional form to capture them all? But what I'm trying to say, those behaviors keep happening. There is a paper I guess from Stanford like emergent behaviors and language models, and they show many, many empirical examples of when this type of behavioral change happens.

arXiv:2206.07682 **[cs.CL]**

Abstract::

> Scaling up language models has been shown to predictably improve performance and sample efficiency on a wide range of downstream tasks. This paper instead discusses an unpredictable phenomenon that we refer to as emergent abilities of large language models. We consider an ability to be emergent if it is not present in smaller models but is present in larger models. Thus, emergent abilities cannot be predicted simply by extrapolating the performance of smaller

> models. The existence of such emergence implies that additional scaling could further expand the range of capabilities of language models.

And Needless to say, people are very interested in, well, you don't have to call that emergent behavior, but people interested in, unexpected changes. In performance or other metrics which may relate to like a safety alignment like truthfulness of the system, robustness of the system or any other property of interest. So it would be really interesting to understand better. When and why behavior may change, so this broken neural scaling loss was just trying to do it as a black box, looking at the system from outside and trying to statistically predict where it happens. But many people also look inside trying to understand what they dynamics of solutions can possibly lead to and can allow you to predict what's about to happen? So actually our studies it seems to be related to certain oscillation frequencies in the training laws. But you might be able to tell from the beginning of the run forgiven initialization of system if it's going to grow up or not, which is good to know.

I think it's super interesting and of course I mean they have I'm not doubting that these things exist. What I don't know is to what extent these are good models for what goes on in the biology because the biology seems to be quite continuous, right?

Do I understand their positions correctly and the argument of Irina Rish goes that we can observe an emergent behavior for engineered systems, but we cannot observe it for the biological systems?

So these may be two different ways to look at this and also sort of big. The reason I emphasize that the continuity is not that I want to minimize the importance of great transitions, because I do think of course there are great transitions in behavior. What I want to do is emphasize the continuity of substrate because a lot of to the discourse you know the human does this and the human does that. And say OK so which kind of human like how far back and you can do this in evolutionary scale you can do it in mental scale, right.

But eventually you follow the human back and you end up with a single cell and we just have to deal with the fact that there is no magic bright line that somebody can draw. OK, boom at this, at this point in the embryonic development. Like that's it now and is and so what we have to understand, of course there can be a non-monotonic, nonlinear changes and all that. But it's a transformation of the exact same substrate. It isn't something magical that gets added at some particular point that and then everything else.

So I'm sort of all about that aspect of continuity. Doesn't have to be linear per se, but we have to come to grips with people who think that it's crazy to attribute, you know, they say it's anthropomorphic to attribute certain kinds of cognitive claims to a single cell. There always a very specific story of how you get from a single cell to a human. You can't simply say that, well, you know, there's some magic that happens in the middle. You actually have to lay out of what? What is it? What's the, you know, what's the phase transition actually? And people sort of assume, yeah, but I've never heard of one. I've never heard of a good one. And so it's quite interesting to me that these two come up in the machine learning world. That's interesting.

Come up, but there is still continue, because, I mean you essentially, well to some extent the X axis is well not exactly, maybe absolutely continually growing the size, but for all practical purposes. So the course are usually always continuous, there's no discontinuous jumps, but continue to see it was kind of different indeed. Speed of development, though those things change. And I basically agree with your point completely that there is no magic sauce or ingredient. Well, unless you believe in divine intervention or something like that, but if you kind of keep this hypothesis as more complicated than prefer simpler based on just things you observe, then indeed there is. There is continuum of intelligence from human and back to the cell. But then of course a good question is like if you start decomposing cell and go deep dive inside. Yeah, then what?

We and other people have looked at this. And in fact, if you look at something as ostensibly mechanical as a gene regulatory network you just look at a model of genes into turning each other on and off. We and others have shown that kind of thing is capable of six different kinds of learning, including associative conditioning.

Biswas, S., Manicka, S., Hoel, E. and Levin, M., 2021. Gene regulatory networks exhibit several kinds of memory: Quantification of memory in biological and random transcriptional networks. *Iscience*, *24*(3), p.102131.
Abstract::

- Gene regulatory networks (GRNs) process important information in developmental biology and biomedicine. A key knowledge gap concerns how their responses change over time. Hypothesizing long-term changes of dynamics induced by transient prior events, we created a computational framework for defining and identifying diverse types of memory in candidate GRNs. We show that GRNs from a wide range of model systems are predicted to possess several types of memory, including Pavlovian conditioning. Associative memory offers an alternative strategy for the biomedical use of powerful drugs with undesirable side effects, and a novel approach to understanding the variability and time-dependent changes of drug action. We find evidence of natural selection favoring GRN memory. Vertebrate GRNs overall exhibit more memory than invertebrate GRNs, and memory is most prevalent in differentiated metazoan cell networks compared with undifferentiated cells. Timed stimuli are a powerful alternative for biomedical control of complex in vivo dynamics without genomic editing or transgenes.

Just from the dynamical systems properties of the gene regulation, nothing else. I was just looking today, there is apparently a literature on chemotaxis in individual molecules and right, I mean some of these individual molecules, right and so and then you can go below that and the kind of stuff that we've done with Karl Friston and Chris Fields and really start looking at very fundamental particle interactions as having sort of extremely with primitive levels of non-zero competencies. So that's a really good question if you have a continuum so the question is there anything and Chris and I can't tell.

That's a great question. So, Chris and I have talked about this and because my claim is, if you think about what would the absolute minimum of intelligence look like, right, like what's the absolute minimum. And so to me, I would, if I had to come up with the absolute minimum, I would lean on two things.

I would say the first thing is some level of goal directedness, some ability to pursue a particular outcome out of a set of other outcomes. And we have this in the least action principle and even in particles and then some level of indeterminacy such that local conditions don't fully determine what this thing is going to do. And we have that in particles. And so I would claim that it doesn't actually bottom out at zero even at the atomic scale. And so I said to Chris one day - could we have a world with no least action laws that where it literally would be zero and what he said was - yes, but it would have to be a universe in which nothing ever happened. So it would have to be a completely static universe, that one. One in which things happen you're going to be able to derive some sort of least action kind of thing, which looks like the basement of cognition to me.

Yeah, very interesting. So it was always kind of an open question how to link this to what people are say doing right now developing like artificial neural networks and those systems. Basically how that would translate an AI domain? What would it mean for this modeling ?Like, can we indeed, I mean people did try to define intelligence and its units, although it's not something that people commonly use, but I wouldn't claim that there were no such attempts, but nevertheless. I mean, we still not really doing that? And keep it there, people say, well, probably should be measuring more formally what we want from our model in terms of intelligence and like how those things compare and can you say that it's as intelligent as like that type of the organism. Basically making this whole engineering field a bit more of a science like physics or even biology in order to make claim that something is intelligent would be at least good to define it more formally. OK, I'm going back to physics and have some motion of units so you can say that in that units this is larger than that. Picture factories. Your models can be a proxy for that.

Chris and I talked. Pulled this in a recent paper is starting off with the idea that making claims about something of the intelligence of something is really taking an IQ test ourselves. Because you have to be smart enough to pick a problem space and identify the degree of cleverness of the policy that the system is following in that space. And you may not see it right, it's not obvious that we can always detect it. And so in biology we are very used to seeing mid sized objects moving at medium speeds in three-dimensional space and saying oh look what you know, look at what that crow is doing that's intelligent, but there are other spaces, right.

Physiological space, metabolic space, transcriptional space, anatomical morphospace, which my lab studies. Cells and tissues do amazing things in those spaces that are absolutely if they navigate those spaces in a way that if we had a robot or an autonomous vehicle doing that would this would be off the charts performance and it's not in the fee, it's not in the domain of classical 3 dimensional behavior.

It's behavior in these other spaces and learning to recognize that, it's very hard for us. Because as humans, all of our sense organs sort of tend to point outwards. If you had an innate feeling of your blood chemistry, you would know without a doubt that your liver and your kidneys were intelligent, because you would see how they navigate that space. You would directly feel it the way that we currently see each other navigate and say, oh, that's pretty clever.

So what we try to do is to lay out and this was this was it goes back to like the 40s, you know Wiener and Rosen we've had this scale right of from all the way from passive materials all the way up to second order like you know metacognition and whatnot and everything in between the kinds of competencies that different systems have been navigating these spaces, and one way to do it is, and I repeat this quote a lot, and I wish I could remember who said it, but somebody said this. It's the continuum between the way the two magnets try to get together and how Romeo and Julia try to get together, right? What's the difference? The difference is in the degree of ability to overcome obstacles along the way.

So you can be extremely simple and all you know how to do is go down an energy gradient and magnets can do that or you can be extremely complex and be able to (overcome anything). But in the middle there's all this stuff, right. And so you can kind of start and this is what we do experimentally. We put these systems in ways where we start to challenge their normal behavior and we find very surprising things that people always thought were like hardwired and just sort of you know this well of course this is you know it's been switched to all it knows how to do these things rolling downhill and that's it.

Once you start to challenge it with various perturbations then you get to find out what it actually knows how to do and then you can form specific hypothesis, does it you know how far ahead can it plan? Does it see, does it have a memory does it you know you can put the way you find out is by challenging all these things and you know like the frog face example and many other things you find out that these things are actually way more clever than they seem at first.

I definitely remember how you gave the talk in December 2013 that new reps word bodies think about and the message to people studying neural networks that intelligence was there way before the 1st neuron appeared, so maybe we should look. But still the common part was networks. They doesn't have to be neural. They don't have to be on your own, but like essentially yeah, I do remember all these experiments were essentially but changing the network communication so that the organism is a pool or the way changes its form to whatever you kind of wanted it to be, but I think even if we scratch neural from that cheap essence of kind of phenomenon will be always network.

Yes, but I would extend that to. Say that it can't be a single level of organization, right? So in biology, the trick is that every network is a network of networks and each one has goal directedness. So every level is trying to accomplish specific things in specific spaces and they all cooperate or compete with each other. So you know this is another talk that it gives sometimes why it's called why robots don't get cancer. Why don't why don't our robotics get cancer and it's because they're not made of parts that have the capacity to go off on their own and do their own thing. In biology every single part is basically being held against its will by the level above. So everything, if you were to just like release the top levels goal directness, the parts would go off and do other things and sometimes that they do and then we get phenotypes like cancer.

I remember really well this five hour discussion in IBM research cafeteria in Yorktown few years ago and the others were being very insightful, making a knowledge between kind of the cell going cancerous or coming back to its senses and essentially the notion of at what scale their objective of the cell to survive and thrive as applied. Because if you are single cell and the environment you treat as a source of food, it's environment, right? But if you become part of the organism, you're under pressure to behave in the best interest of organism. If you forget about that due to changes in your communication with other members of organism society, then you start acting selfishly and too bad. That's cancer.

Michael, because of you I'm very interested in the definition of intelligence and you're saying words like goal-directedness, agents, planning, and these terms are a little bit anthropomorphic, but you could argue that they are universal primitives and you're also talking along the lines of intelligence. Being almost like some people describe it as capability or with an underlying principle but it's almost like it's a continuum and there are people who believe that humans are special. We think you know the behaviorists thought that it was just a very simple stimulus response, that monkeys don't have a flash of inspiration, they don't plan how to get the banana. We have something different. What do you think?

Well, I'm certainly not saying that we don't think or that there is no material difference between us and let's say a paramecium so that the differences are of course real. I insist on one thing and then I have hypothesis about other stuff that the thing that I insist on. Is that when we make a claim about the level of cognition of some system, real artifact, you know, it evolved, artificial, whatever. You cannot make those claims from an armchair. You have to do experiments and you cannot have feelings about these things that people say to me all the time. Oh well, you know when sentence - well for it is a frog skin that can't do XYZ, you name it, whatever, you can't just say, (because I would answer) - how did you arrive at this conclusion? You cannot just have feelings about this stuff, you have to do experiments.

And when you do experiments as we found and other people have found, you get surprised because it is not obvious what the capabilities of something is until you start to perturb it and see what happens. And the thing and to find out. So in order to make an intelligence claim, you need a couple of things. You need to pick a space within which you think it's solving problems, right? So you as the observer have to say, I think this is the space in which it works. Then you have to say here is the goal I think this thing is pursuing in a in a cybernetic sense. And then you have to say - and now I have a hypothesis about how clever this is, what can it do? Can it go around obstacles as a trap by local minima? Does it have history? Does it have this that you make some sort of claim?

Interesting idea to assess the intelligence of professionals we hire and for self-evaluation as well. What is a space, performance in which we evaluate?

Having made those three claims, we can then empirically see how well does that work out for you? What does that enable you to do? And so then then multiple observers, so someone can come along and say, you're crazy, this thing doesn't have any of that. Somebody else can say, oh no, it's actually much more clever than that. And then we all interact with the system and we see who wins. It's an empirical question that has to be put. So the only thing that I insist on is that this charge of anthropomorphism. It has to be quite specific. We have to make a specific claim. What do we think this thing is able to do? And then we get to find out is that a useful lens onto the system or not?

OK. Just a quick follow up. What is it? Because a goal feels like it's a wicked form of reductionism along the lines of we what we were talking about before and even if because you know, as Shane Legg has this idea that the definition of intelligence is an agent being able to solve a variety of tasks in different environments and so on and even if there were such a goal why would it be intelligible to us?

Legg, S. and Hutter, M., 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*, pp.391-444.
Abstract::

- A fundamental problem in artificial intelligence is that nobody really knows what intelligence is. The problem is especially acute when we need to consider artificial systems which are significantly different to humans. In this paper we approach this problem in the following way: We take a number of well-known informal definitions of human intelligence that have been given by experts, and extract their essential features. These are then mathematically formalized to produce a general measure of intelligence for arbitrary machines. We believe that this equation formally captures the concept of machine intelligence in the broadest reasonable sense. We then show how this formal definition is related to the theory of universal optimal learning agents. Finally, we survey the many other tests and definitions of intelligence that have been proposed for machines.

I don't believe that there is an objective goal that is intelligible in the sense that we get the that it discovered once and for all and then we're done. I think that when somebody says that this system has a particular goal, what they're saying is they have discovered a lens of perspective on the behavior such that positing that it has a goal and some set of competencies to reach that goal gives you increased ability to predict and control that system.

So we come across the thermostat and you look at this thing and you know you're a reductionist. Then you say, OK, I can sort of calculate what all the particles doing and that's great. And I look at it and say, hmm, I see this and this, and I think this is the set point. And you say this is such thing as a set point. There's just atoms. And the thing I say, well here's the thing though. I will exert much less energy and effort in changing the temperature of the room while you're running around pushing on all the individual atoms of this thing. I'm just going to change the set point and let the thing take care of itself.

So the level of what it affords you is a more powerful level of relationship to the system. For simple systems, that's prediction and control. For complex systems, that may be everything from training to friendship to who knows what, right? For complex systems what we're looking at is what level of agency gives me the best interaction with the system. And I can guess too high right? If I'm standing there arguing we could pleading with my thermostat. You know here are all the reasons you should change. Like that's not going to work. And at the same time if I have a system that's a learning agent but you know, a horse or a dog, humans thousands of years ago figured out that you don't need to know the neuroscience of what's between their ears. You could train the thing.

So you found an interface that that that allows you to do amazing things because you understood that this is not the same as a complex this instead of bowling balls that you would have to engineer bottom up so that's all. These things are just meant to be. I don't believe there's any such thing as anthropomorphism. Anthropomorphism sounds to me like we're still working with the Garden of Eden story that you've got the human. The human has this magic of glow about them and trying to sort of sort of get into other things. It is a major transgression. I don't buy it. I think that what you have to do is you have to make very specific claims. This human or this robot or this is cell can do XYZ and then we find out it can or it can't and then we you know we sort of empirically we know did we guess correctly, did we overdo it, did we underappreciated. In my experience, we tend to underappreciate intelligence way more than we overappreciate it. That's just what I see in the Biosciences.