A Conversation with Collin Lysford

by Suspended Reason

March 29, 2026.

Contents

On form lightness

Reason: One of the approaches you advocated at TIS is "form lightness." I understand this partly in Piagetian terms: rather than immediately assimilating new encounters into an existing classification system, you should stay open to accommodating their novelty and refining your worldview.

Lysford: When I'm thinking about form lightness I'm thinking about the Kegan developmental psychology model, which David Chapman's a big fan of. I'm not the biggest Kegan guy, but I like the idea that there's a pre-rational stage, and then you reach technical rationality, which gives you a ton of leverage. The reason we can feed more people than non-rational societies is because these findings about the yields get formalized well enough that we can lever them. But they can still kind of go bankrupt, these concepts, right?

I think a lot about concepts as "loans of meaning"—where I need to really understand something if I want to lever up on it, but I'm also willing to invest in riskier assets if the dividend is worthwhile and my losses are bounded. To me, Kegan Stage 5 is about treating technical rationality as leverage that you apply differently to different portfolios. Because there will always be factors from without your rational system, that make it stop working, and that aren't under your control. And when I think about form lightness, I think about allowing magical thinking, but only for spots that are dark enough that magic could plausibly live there.

Reason: I remember when we watched a film in your new home studio, and you'd done so much research to get a truly HD viewing experience, down to the cables. The so-called closed system had been perfected. And then the angle of the sun that day came through the

window and caused screen-glare and the closed system was shown to be quite open.

Lysford: Another example of form lightness: every nurse practitioner will tell you, “Okay, it’s a full moon, the ER is gonna get nuts.” You want to be able to say, “All right, I’m willing to entertain this possibility. If a lot of nurses report this, maybe it’s true.” Whereas a wet-blanket rationalist skeptic says, “I don’t understand how the moon could influence that, ergo it’s false.” That’s the cardinal sin of Stage Four.

What you can’t do, though, is say the ER uptick is due to the moon making people’s werewolf glands expand. Because then you’ve got to show where the werewolf gland is. If you’re trying to cash it out so precisely, you’re doing Stage Four, and you’ve got to play by Stage Four rules. Instead, you’re allowed to just say, hey, I think there’s something about the full moon. That’s form lightness.

There’s a related anecdote I like, in this book about one of the original big chip fabs. The pressures these chips are sensitive to, during assembly, are so tiny that you even care about the phase of the moon, because it changes groundwater levels, and therefore the moisture levels within the labs. Or if a woman was ovulating, her hands had a little more oil on them, and that would stray outside the error tolerances.

As you start going into the fractal-like nature of reality, it gives you all these new error tolerances. And no matter how precise your Stage Four gets, that just enables you to do things that are exposed to further levels of error tolerance. This, to me, is the idea of form lightness: understanding “Here’s where I’ve got a concept that I’ve pinned well enough that the error bars matter, and here’s the stuff

outside the concept where I'm willing to just observe and let the world do the thinking for me."

Reason: And this is related to your notion of formal empiricism as well. To use the meaning-debt, cash-value metaphor of meaning, you say: such an approach is formless, because you don't need to pay in a conventionally understood way, but it's empiricist because you do have to pay all the same.

On strip-mining language for meaning

Reason: In your essay "Representation & Uncertainty," you write:

> ...any intelligent AI will need to be an embodied AI that can pick novel details out of its environment, not a correlation machine munching on pre-chewed data. Hopefully once you've internalized "You can't diagnose patients from behind the door," it's easy to see why "Okay, but what if the computer behind the door is like, really super duper fast" is not an especially serious argument.

To play devil's advocate, doesn't a lot of medical diagnosis already happen functionally behind a door, based on patient self-reports of symptoms? Or else through imaging, which is another form of representation?

Lysford: I still think interactivity is necessary to reach dramatically post-human reasoning, but I have a new appreciation for how you can strip-mine insight out of language. Wittgenstein's point, when he talks about language games, is that "slab" is a meaningful word insofar as it's a chip that can be reliably used in an interaction. And it was an appreciation of this that first turned me against AIs behind the door, static AIs.

What we've learned is that, the fact that 'slab' was a useful concept is enough to learn about the nature of slabs. Other hypotheses were reasonable before LLMs, and they're now I think pretty definitively disproven. LLMs are effectively our interface to talk to language,

and language happens to have oodles and oodles of fossilized meaning inside of it.

Getting better and better at strip-mining the meaning in language isn't the same as making new language. It's like being really good at drilling for oil. You don't eventually start putting oil back in the ground. That's a totally different thing. And it's the kind of thing that takes a lot of time and pressure, that you can't necessarily speed up.

On social deduction games

Reason: You helped me get into the world of social deduction games. I'm curious what you learned from playing Mafia online.

Lysford: I talked about this for Adam Mastroianni's Mad Science Talks, at the Montauk Club in Brooklyn.

I'm in a weird spot where I'm trying to formally study tacit knowledge. There's a logjam between tacit, super-involved domain experts, and the nerds who can exert leverage in a lossy way that eliminates detail. The way to break this logjam is to be a leverage nerd whose subject is tacit knowledge.

My understanding of talent distributions is strongly influenced by playing Mafia. Because what you notice is that, in the beginning, there are some easy wins from thinking about the game in an at-all structured, formal way. There's a newbie queue for players new to the site, and they always do the same things, and they always get picked apart. Trial by fire.

But down the road, people's self-reflections aren't that well-codified, which makes lots of sense, because if you think about the universe of codifiable things—where chess games have perfectly describable states—social deduction games are on the opposite side of the spectrum; the representations are actively fighting against you. They are anti-inductive: recognizing a pattern can lessen its frequency; naming a theory can make it less true. They're the slipperiest representations out there, so a lot of new players hit a wall and stop improving.

To improve, you have to start chunking and coming up with concepts on your own, because you can't join a community of empiricists trying to develop better chunks, if everyone's trying to defect from the shared understanding to win a game.

And yet you also see these skilled practitioners who hit way over replacement. In face-to-face social deduction games, some people are just good readers of human beings, and that makes sense, but people who can get over 80% accuracy rate over text-only games—that's insane.

When you notice this, then you begin to understand that there's an unarticulated mass, demanding to be found, and that there are an infinite number of ways you could try to dredge it from the seas of understanding. And some are going to be better than others.

I also learned a lot about how to be a good truth teller. We tend to think, "Oh, I'm telling the truth, so it's incumbent on the reader." But if you think of communication as being a lot more two-way, then you understand there are ways of acting that make it harder to be a liar. And I'm going to demand the people I talk to act this way.

Reason: The importance of keeping counterfactuality in mind seems like a significant takeaway from social deduction strategy. Have you noticed yourself maintaining counterfactual models in the world?

Lysford: Running counterfactuals is really, really hard. It's expensive, because in the world, you let the world do the thinking for you, right? You throw the ball; it lands there. You don't have to calculate air resistance and gravity. You just look, and looking is so much easier than building counterfactuals. And that's why one of the easiest ways to find the townie is: It's the guy who's working the hardest.

Theoretically, you can be mafia and work really hard, in order to trick people. Some people get off on it. But most of the time, if you're working really hard, it's because you're trying to run counterfactuals, because you actively need them. The mafiosos already know the right answers.

When a mafioso needs to run a counterfactual—when he needs to pretend to be a misguided townie, and build his fake townie suit—he doesn't want to tweak his counterfactual structure once he's built it. Whereas early-game townies, a lot of the time? Look almost drunk. They're shouting, "You're a mafia! No, you are!" Because that lightness of foot is already priced in for townies.

In the real world, I think a lot of the time, the best you can hope for is an understanding that a counterfactual has or hasn't been run.

On GeoGuessr and overfitting

Reason: In your blog post "Mongolian Meta," you talk about GeoGuessr, and distinguish between robust and flimsy sources of information. Some user signals are easy to fake, or change on their own. Others are reliable as a basis for inference. Pro GeoGuessr players can make snap-judgments based on the shadowy outline of the StreetView car, but that correlation will change as soon as Google re-surveys the region. Whereas the location of the mountains in Mongolia won't change for millions of years.

Lysford: If some people go out in the rain, and some people go out when it's dry, everybody knows to take their coat off, if they want to disguise which person they are. But the question is: Are your fingertips pruney?

That's a second-order effect of rain that rained-on people get for free without simulating, but fakers need to spend energy on. Pretending means compiling dynamic processes to an arbitrary level of detail, whereas living is just living.

Reason: What are the entailments for AI safety, given that texts and symbols are in some sense the cheapest and flimsiest correlations? It takes active work to maintain the relationship or binding between symbols and reality. Is there some text analog of the Mongolian mountains, something as robust as counterfactual coherence, or is it an inherently flimsy medium?

Lysford: I think it all depends on how the text is generated. Normally, playing mafia is like reading back the court record. This is

a text-only mafia game, mind you. You go back to post #575, and say, "Hey, here you claimed X, but in post #590, you claimed Y."

But if the heat is on you? If you're under suspicion? You should just go to a new text window, not read the thread, and just state your views on everybody in the game, top to bottom, without ever looking back on the record. Because what you're doing is showing everyone how you've encoded the game state in your brain.

The townspeople and the mafia are encoding things in their brain for very different reasons. As a mafioso, to encode things as a townsperson, you need to simulate a townsperson, which is much, much harder than being a townsperson. This idea comes up constantly in my research. If we lived in a Minecraft world, where all of the primitives were given to us like chess, then plausibly simulation could become as cheap or cheaper than observation, and machines would rule the world.

Machines don't rule the real world because the real world's hard to simulate.

And because the real world is hard to simulate, different collections of text can be more or less likely to be true, because they come from different generative processes, and faking the generating process is much harder than merely enacting the generating process.

If I ask an LLM a general knowledge question, it might just make shit up. But if I say, where in our database do I find all the variables, and all the files that touch those variables, and it prints out a list, that list is more likely to be true. If you are forced to tightly interlock text with certain elements of the world, it makes the text more resilient.

Reason: In your writing on GeoGuessr, you talk a bit about your contentious relationship with the concept of overfitting. It seems like your perspective, perhaps I'm not doing it justice, is that describing a model as "overfitted" is a value statement. It isn't a neutral kind of statement of fact.

Lysford: A phrase I like to use is "aesthetic of the real." Some people, and I'm one of them, really don't want to learn the car-shadow meta. Devoting space in my brain to these photographs, that I know are going to change, feels offensive to me. But you couldn't enforce this aesthetic in a GeoGuessr tournament, nor would you want to. Knowing the car-shadow meta makes you a better player up and until they take new photos.

Learning a set of photographs is just a new patch to learn. You can accept this, and still have an understanding that it's possible to predict in advance which attributes of the new photos will change and which ones won't, because they're generated by real world constraints. It takes different amounts of literal thermodynamic energy to perform different kinds of changes.

It's not necessarily wrong to call it overfitting, but it's also not the most productive frame to think within. When people say "overfitting," what they mean is, they know they're going to get some out-of-distribution stuff, and they want the model to still work there. But what "work" means is value-laden; you need to make judgment calls about whether a certain kind of error is costly or not. So you can't have a values-free or objective "overfitting," because it matters what you're doing with the correlations afterwards.

On predator models

Reason: What is a predator model? You describe it as being very different in structure than its prey. Do you think we're going to end up with an emergent ecology of models, complete with trophic layers and complex symbiosis?

Lysford: Back in the early 2000s, I'd play this game *Backyard Football*. And I figured out this play that I called "the hook," with my wide receivers, and because the AI code was so simple, no one would intercept it.

I always won, because I'd exploited a regular pattern in the AI, but that doesn't mean I was good at *Backyard Football*. If I played online, against people, they would change their strategy in response to the hook, and even people who couldn't beat the same AI as I could beat? Could easily beat me.

That's a predator model. In an adversarial game, if you can beat your opponent, it doesn't matter if you can win their game or not. In the same way that a cheetah doesn't defeat zebras by eating grass faster than them.

To the extent that people are going to have different reasons why they run LLMs, some sort of ecology will emerge. But it's just too arbitrarily easy to spin one up and shut one down. So you won't have the levels of trophic layers that you get in the real world. Because energy is just too cheap.

Reason: In "Adversarial Asymmetry," you say that LLMs can always be easily preyed upon by predator models, since:

> ...doing so necessarily diminishes their predictive power over the data that exists today, because it entails throwing away the parts of the meta that won't survive change. So an AI that is trying to optimize score does so exactly by considering every single bit of the meta, thus becoming more susceptible to predator models. They don't need to survive change and so they're not trying to. The strategy to build a living thing that will endure is altogether different.

Will AI need to care about surviving change, like a Darwinian agent, in order to threaten human existence?

Lysford: Yes. The AGI models that scare me involve agents interfaced with the world, and the richness that comes out of having thousands of sensors measuring the world. Because that richness is what dictates your ability.

You need to use AI to understand how weird it is, in terms of, say, leading-the-witness problems. If you're asking a factual question, and you're genuinely unsure, you cannot tell it what you think is true. Because then you hand it a frame. It's very hard for it to fight your frame.

And the test for survival is surviving, because something I genuinely believe is that simulating the world is always going to be infinitely harder than observing it, and if you want to have a computer that predicts the wave, eventually it's gonna have to be so big and so wet that your hard drive has to be the ocean. Or you won't have the space to encode everything that's happening.

On excluded middles

Reason: What is the excluded middle, and how does it relate to your thinking?

Lysford: People tend to think a thing is either X or Y. I'm always looking for whether they're X or X-prime, but nobody knows the difference between the two yet.

Sometimes you can make a bajillion dollars by figuring out how to sort X and X-prime into two different buckets. Insider trading scandals are a real laboratory for this. Traders are looking for a dimension that'll make the difference in a hostile bid. Somebody's always trying to find the differential razor first, then trade on it.

A lot of times, I think the sciences could stand to have this level of scrutiny applied. But then again, you would impose that through a Polymarketesque regime. And Polymarketesque regimes are exactly what I'm rallying against, in the sense that they are trying to decompose the world into binary yes-or-no outcomes.

On object-labeled memes

Reason: It's been four years since your infamous post on the Jar Jar Binks x Catwoman meme, and AI's inability to understand novel jokes. How has your thinking on LLM capabilities evolved since then?

Lysford: I think LLMs are weird. And that's the one cardinal sin in the discourse, I think: Not finding them weird. The jaggedness of the frontier of what can and can't be done.

The idea that we first got photorealistic replicas based on a prompt, and then we got the number of R's in "strawberry"—that's strange, right? And I think that when you're trying to build a model of reasoning, which is what I've been working on, certainly LLMs are the most novel thing—the most new evidence—in that sphere in years.

Reason: Phil Agre, David Chapman's research collaborator in the 80s, thought of AI as applied philosophy. That it's a way of testing and developing our theories about how minds work.

Lysford: When Wittgenstein is writing his stuff, he's reflecting on the human activity he sees, and that's been the same for millennia. And suddenly we have this new type of activity to investigate.

And I do think my predictions partially were wrong. I thought that getting novel context was going to be a bigger barrier than it turned out to be—the way web search and an MCP push context back at you.

Reason: Before this call, I checked whether current models can figure out the Catwoman x Binks meme, and they can now. They actually pointed out something I hadn't realized, but that was obvious in retrospect, which is that it's a variation on the distracted boyfriend meme.

The Distracted Boyfriend meme is an example of what gets called "object-labeled memes," which have some base relational structure whose instantiations vary. So in Distracted Boyfriend, you have these three base characters, and you can metaphorically re-label them to explain some novel situation.

I'm curious if you've thought at all about how memes and object-labeling might drive the formation of new concepts.

Lysford: Yeah, like the Rage Guy extended universe. And with the F7u12 team memes, that literally happened, right? It was like Avengers Assemble. Wojaks today are exactly the same.

Concept formation is interesting because what I focus on in my research is the idea that you've got a pattern in the world that you can work with, but can't articulate yet. And then these articulations are inherently lossy. And the question is whether the concepts bind enough that you can spend them in other contexts. I think of this as a form of leverage.

Pro-AI people get mad because AI critics are constantly moving the goalposts: First they say AI can't do this or that, and then AI manages it, and the critics come back and say, Well actually. And pro-AI people say, You're not a credible critic anymore. But I don't mind the moving of the goalposts, as long as your theory explains why you were wrong before. That's what I mean when I say that

LLMs are the most interesting thing to happen for our theories of reasoning in a long while.

Reason: What do you mean by leverage?

Lysford: Skilled practitioners have metis, they have tacit knowledge. They can do things in the world which go beyond anyone's ability to articulate or generalize. Rationalist systemizers, on the other hand, are good at abstracting and articulating.

On an individual level, people usually prefer the skilled craftsman, but the craftsman isn't levered the way the rationalist is levered.

Leverage means that no matter how mediocre the solution, it can be multiplied and it just doesn't matter. I'm sure there are many cases where the standard practices of farming, going by the manual, will perform worse than the farmer who's lived on the land for a hundred years and knows its particulars inside and out. But you can scale up the manual, and interrogate it and prove it, in a way that the tacit craftsman can't.

When I talk about ontological development, I'm gesturing at our ability to conceptualize the tacit better.

On LLMs and science

Reason: How much can LLMs contribute to science if they're not able to do ontological remodeling?

Lysford: I think of ontology and computation as an electromagnetic wave: a changing electric field generates a magnetic field, and a changing magnetic field generates an electric field. They feed off each other.

Nowadays, ontology and computation have been thrown out of balance. Computation is getting quite close to free. I think pure mathematics is really the place to look out for this. The results in pure mathematics... In five years, ten years, computation is going to be so close to free, it will be almost indescribable.

All of the roadblocks will be ontological. Historically, we find that it's often ontological findings that are highly levered. You make a new finding, you have a new theory, and suddenly all this new information comes to you.

Free computation frees people up, in theory, to get better at ontology. But most people who get good at ontology get good by doing a lot of computation by hand, and subconsciously noticing the higher-order patterns.

I don't know how people get good at ontology without doing computation. That's one of my research aims, to find a principled way to get better at ontology beyond just suffering through it.

Thanks to Zak Hap for his Not Nothing bookmaker, which was used to compile this PDF. For more of Collin's writing, visit tis.so/archives.