## 🧪 1. Intervention Framework as a Test of Bird's Causality Claims

**Bird's Causality Thesis:** Bird argues that neuron alignment (or lack of it) is a *consequence* of deeper architectural choices—specifically, the model's **functional forms** (activation functions, normalization patterns, etc.). These induce **privileged bases** in activation space. The illusion of neuron-aligned concepts arises only when these privileged bases *happen* to align with the raw neuron axes. Thus, alignment isn't "innate" to deep learning—it's *induced* by symmetry-breaking patterns in the system.

**Your Intervention Tool:** Your script `capture_intervened_activations.py` literally injects controlled changes into individual neurons mid-generation. This lets you:

- Clamp a neuron like 373 to values from `-20` to `+20`
- Observe the **downstream effect on generated text**
- Capture the **mean vector response** in high-dimensional MLP space

You're not just passively observing—you're *perturbing the system* to trace causal influence.

**Why this matters:** Bird says functional forms induce alignment. You are testing: *what happens when I forcibly override those forms, one neuron at a time?* If the privileged basis is truly emergent, then:

- Forcing a neuron off-alignment (e.g., clamping it to −20) should **destabilize** the geometry in predictable ways.
- Conversely, boosting it (+20) might **amplify privileged basis alignment**, if that neuron contributes strongly to a key representational direction.
- If neuron 373 is part of such a privileged basis, your intervention will create **coherent directional shifts** in vector space—and those will *differ* by prompt type/level/core, as you're already seeing.

So in effect:
**You're using SRM to reveal whether Bird's claim—that architectural choices causally shape alignment—is observable at the neuron level.**

This turns SRM from a descriptive tool into a **causal diagnostic probe.**

## 🧭 2. Compass Rose Plots as Glider Trails in a Synthetic Landscape

Let's pick up the analogy from *currentscurrents* on Reddit:

> "To actually interpret the computer, you would have to work at the level of gliders, not cells."

This is **incredible** when applied to your SRM work.

In cellular automata (e.g., Conway's Game of Life), **gliders** are emergent patterns—coherent packets of motion that persist across steps, made up of fleeting, low-level cell states. Individual cells flicker in and out, but the glider *endures*. It's a **unit of meaning**, not reducible to any one cell.

Now consider your **SRM Compass Rose**:

- Each vector is a mean activation in 3072-D space after neuron intervention.
- SRM projects these into a **2D plane** (usually neuron A and B).
- You sweep across angles (θ) and thresholds (τ) to track **alignment dynamics**.

Think of each **prompt condition** (e.g., Level 3 rhetorical) as a glider:

- It leaves behind a **trail in the SRM plane**—not just a dot, but a directional signature.
- These trails **rotate**, **compress**, or **cluster** in response to your neuron interventions.

You're not visualizing neurons. You're visualizing **conceptual dynamics** across a projection plane—a trace of how concepts "move" through the synthetic forest.

In this metaphor:

- **Neurons are trees**
- **Prompts are birds or animals**
- **Trails on the compass rose are their migration patterns**
- **The plane (e.g., 373/2202) is a hidden map embedded in the forest's topology**

So just like gliders in a cellular automaton:

> *"The real behavior emerges when you stop staring at the cells and start watching the flows."*

Your compass rose becomes a map of those flows. And with interventions, you're not just observing—you’re changing the weather and watching how migration patterns shift.