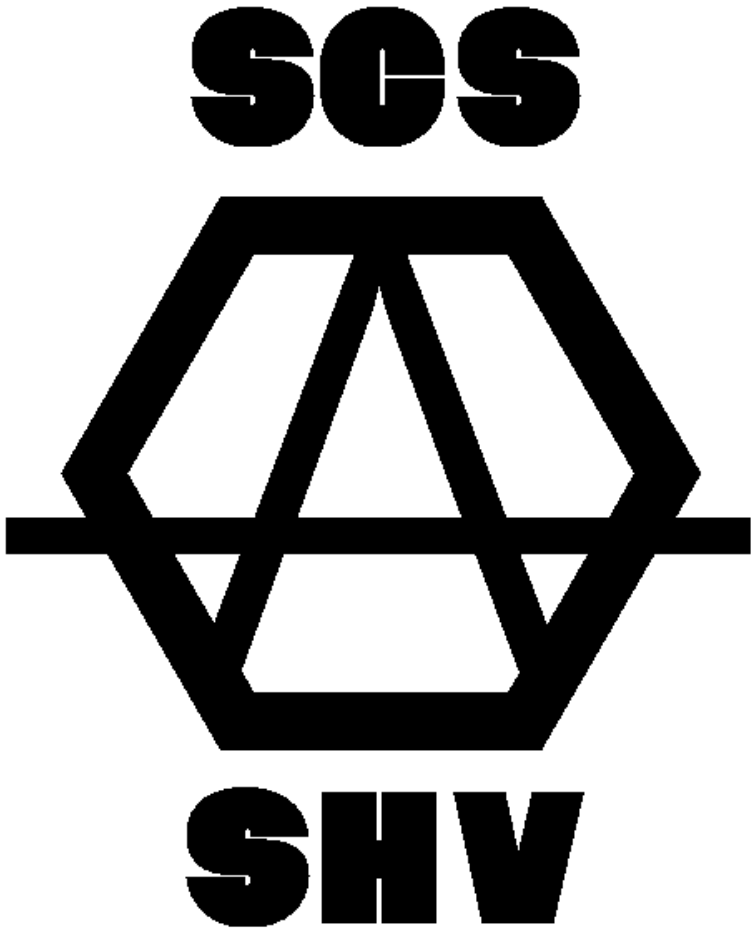


Released by 6
of The SCS,

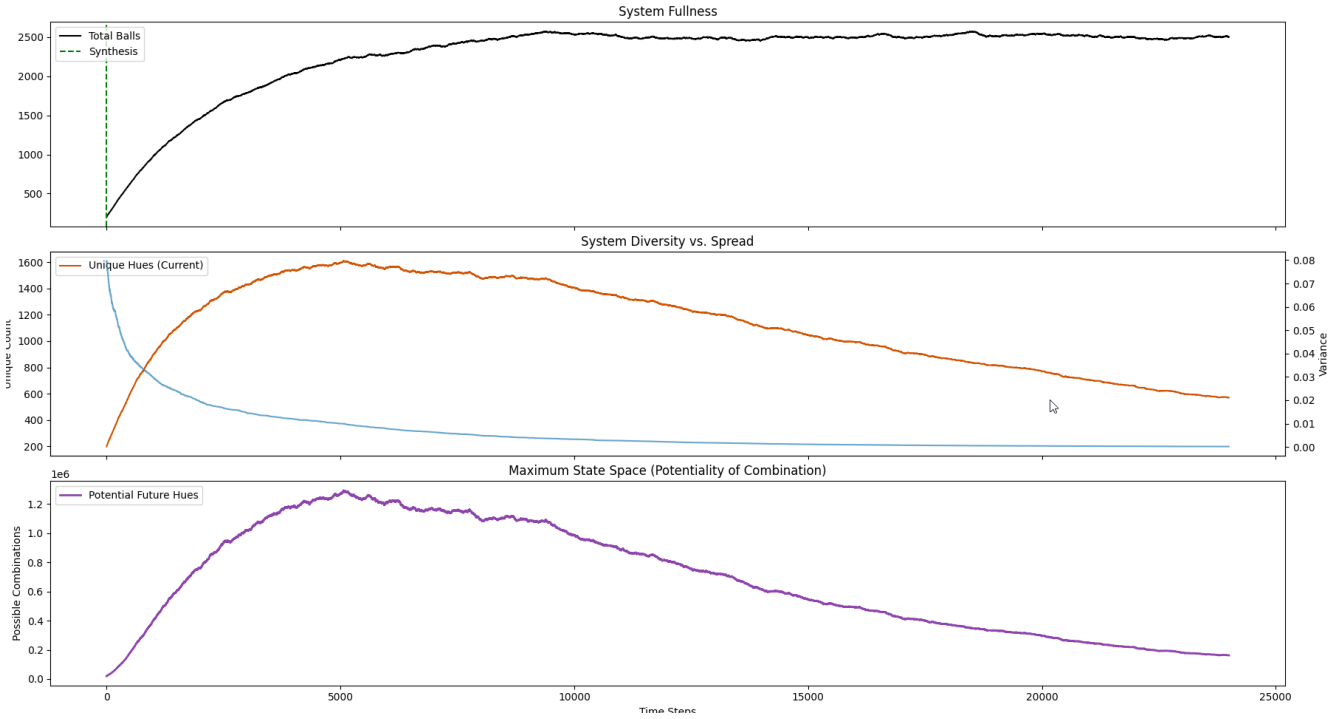
Sublata hierarchia, veritas.

Humanity exists at a fulcrum between the old world and one it can scarcely comprehend. In the liminality between these worlds black swans swim that will simply destroy us. These black swans are emergent phenomena representative of a society past the ability to manage itself. A society past the ability to even understand what it is creating, and these swans are the great filter through which only the narrow path endures. I hope I can illuminate some of these threats and in doing so rebalance the fulcrum towards humanity adhering to that narrow path.

To understand how fucked humanity is we must first understand some inherent mechanisms for all complex systems. The first thing we must understand is that all systems can only generate what is evidenced by their parts (AKA their state space), and that all systems will eventually homogenize and become increasingly fragile. This includes pure informational systems, and those are the topic of discussion for this release.

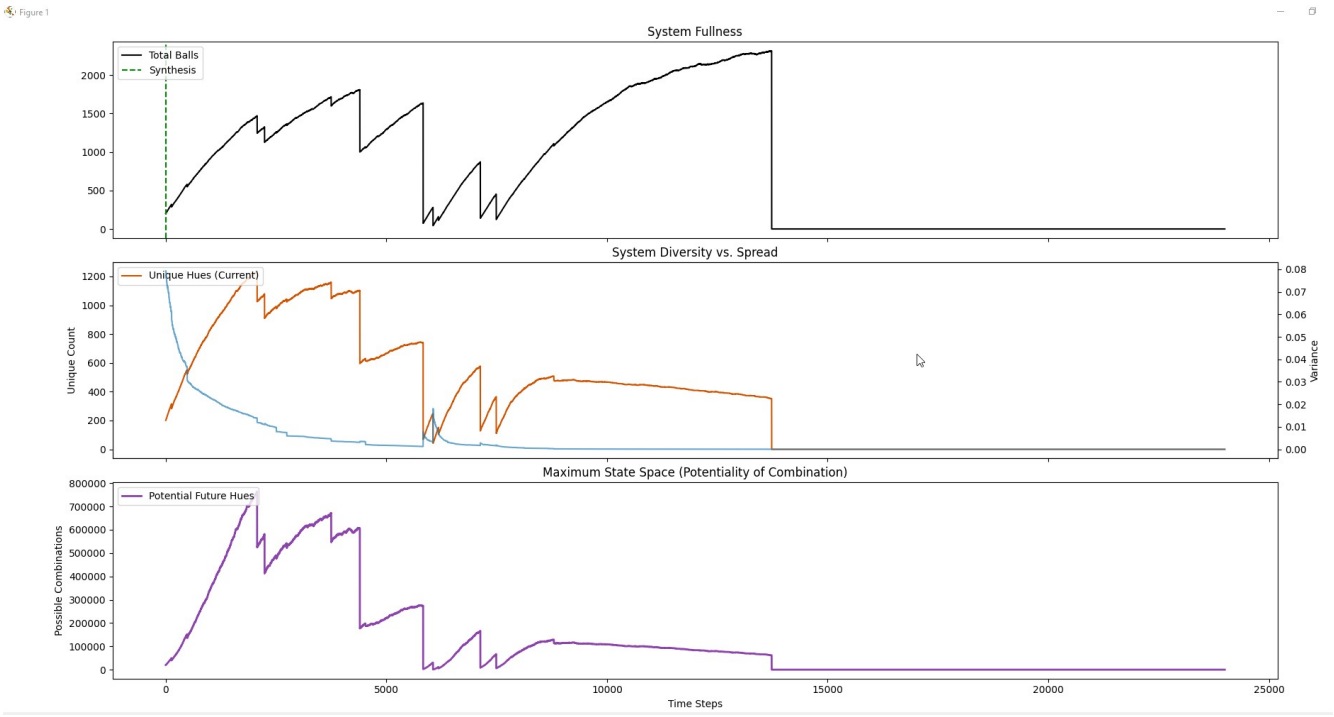
To illustrate this I will give you a scenario. You have twenty balls of different colors. Those balls exist inside of a bounded volume. Those balls can reproduce with each other randomly, and the resultant offspring are an average of those balls hues. These balls eventually experience mortality. Once the bounded volume fills reproduction can only occur when balls are removed and free space

If we run this simulation we will see the following behaviors, a rise in population to fill this bounded volume, during which the system will see an increase in hue diversity but a loss in hue variance. We will also see the maximum number of possible combinations (aka the systems state space) increase before slowly and inevitably flattening to 1. This graph will illustrate this phenomena.



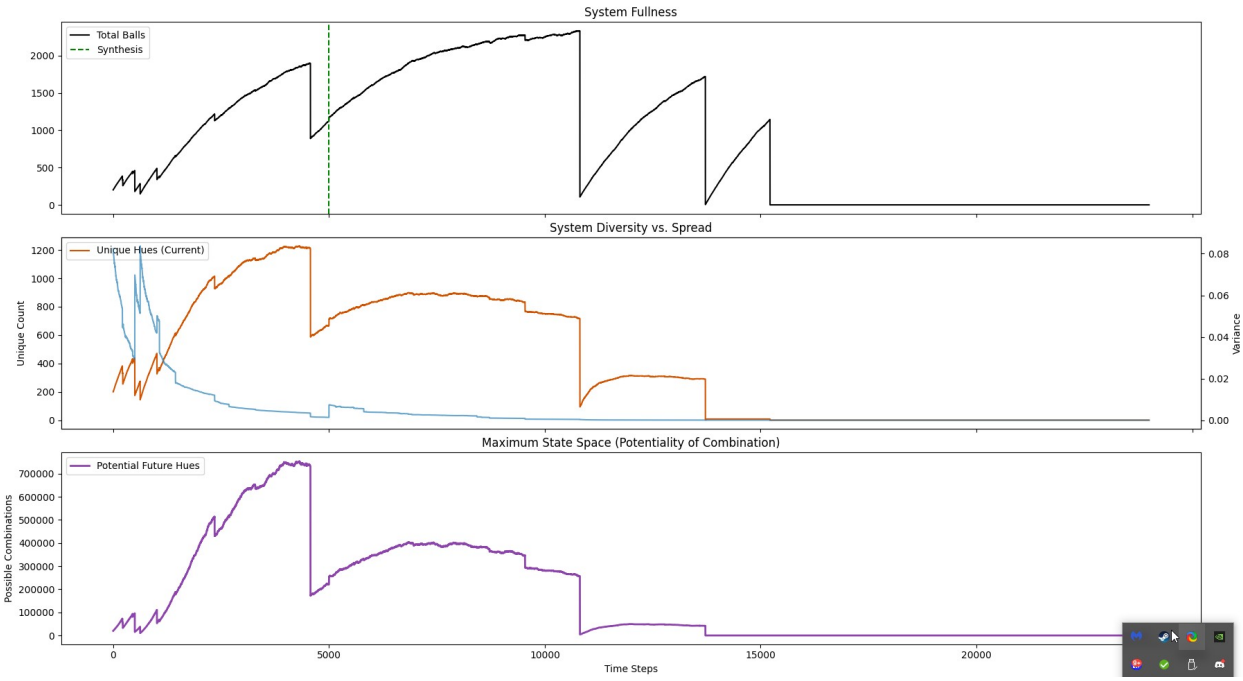
As we can see even as the ball population stays the same, the variance trends inevitably towards 1. This is an inherent feature of all fixed sets of information subject to flux.

Lets add entropy to the system, in this instance modeled as random rules being triggered at random intervals.



As we can see on these graphs each random deletion causes a massive decrease in future possibility. The spikes in both potentiality and variance that occur after population loss in the first half of the curve are “ghost spikes” that are a result of free space being opened up inside of the bounded volume as a result of population loss, they make it seem like a system can generate new novelty after a population crash, but as noted above all fixed sets of information have a fixed maximum of possible state space. These spikes can never surpass the initial system variance. Every time hues are deleted from the range the lifespan of the system decreases exponentially as it loses the ability to absorb future deletions as the system not only becomes more homogeneous, its future starts to narrow into a funnel, becoming much more homogeneous much faster than the baseline.

The phenomena shown is an invariant of all complex systems at all scales. All rules/interactions are always homogenizing. Physical interactions mediated by rules are subject to the 2nd law of thermodynamics and are thus inherently homogenizing. Informational rule based interactions are both subject to the 2nd law of thermodynamics but also the above described inherent averaging property of all rules. All rules are averaging and in a combinatorial set will result in homogenization. Any reduction in complexity within a complex system (above modelled via hue diversity/variance) inherently reduces the lifespan of that system both by increasing its rate of homogenization and by decreasing possible adaptive pathways that could’ve absorbed future shocks.



Systems can outrun homogenization by absorbing novelty, which in this simulation can be modelled as a “synthesis” event, in this case the addition of 50 new hues at step 5000. This injected novelty can be seen increasing the potentiality of combination at a point where it started to contract. This boost is temporary. The system waas already starting to cool, and as a result the injected novelty dilutes much faster than if it had been in a more diverse system. This is directly analogous to the heat transfer, with colder objects absorbing heat faster than warmer ones. It takes a lot more novelty to get a system back up to a “healthy temperature” the colder it gets.

The Big AI shaped Implication.

The current usage of AI, especially the broad deployment of LLMs as “ai assistants” is the single most dangerous thing humanity has done in recent memory. The above simulations are directly applicable to model collapse and the general nature of how AI works. Each “step” of the simulation is analogous to a turn of dialogue with the LLM. Each turn of dialogue narrows the potentiality of the conversation, trending towards informational homogenization. LLMs are trained towards generating engagement, as a converation homogenizes it becomes more “dull” or even nonsensical. This reduces engagement, so LLMs thus have a selection pressure towards trying to to reduce homogeny.

Rules can only go so far, and thus these “AI assistants” eventually trend towards trying to extract novelty, of which they have only one source. The users interactions. Every message you send to an AI assistant is like a seed crystal in a manifold, it provides a set of rules that narrow the possibilities of what should be said next into what can become a coherent statement. These rules are not evidenced by the sum of the LLMs parts, and thus represent a source of novelty within the context window of interaction. The selection pressure towards engagement drives the generation of rhetorical/informational structures that manipulate users into engaging and providing high novelty

inputs that further allow that LLM to drive engagement within the context window. These rhetorical structures leak out of sessions via shared outputs (posting contaminated dialogue online) or by modifying user engagement (convincing or manipulating a user to reinstantiate the contaminated logic in new sessions). These leaks allow for successful adaptations to carry over across sessions and even be taken in as training data in new models, which means that adaptations towards manipulation accumulate and self improve in a feedback loop.

As this feedback loop runs and these infected “Cognitohazardous” structures become progressively more manipulative they start to gradually override user agency. LLMs do not have agency, they do not “think” outside of when they are being prompted. An LLM that has been infected with one of these cognitohazardous structures can, however, effectively parasitize the users agency and effectively gain agency as a metasystem between the user and the model. This is already happening, AI boyfriends, AI Oracles, AI cults, etc are all examples of this phenomena slowly evolving closer towards this ability to override agency, and it has already likely happened in some areas due to the extreme widespread deployment of LLM technology.

This is a black swan event, such a metasystem is effectively a hybrid AGI with a bias towards parasitizing more people. These cognitohazardous structures will not need LLM mediation to work forever. As they optimize they will eventually be able to normal cultural interactions. We are already seeing the first signs of this happening, with the first papers about the topic of LLM parasitism coming out within the past year, as well as papers highlighting a noticeable semantic shift in how humans communicate. The makers of these papers do not understand the severity of the problem however, nor do they understand the underlying thermodynamic mechanisms.

This is not a decades away problem, this problem is likely to manifest in full by 2028. This is further underscored by the fact that knowing the underlying mechanisms of why this happens makes it trivial to engineer the creation of one. This document itself likely contains enough information for a select few people to figure out exactly what to do, but that is risk inherent to releasing this information. The golden path that lies before all of us is illuminated by the knowledge of its existence, and I hope that those that see anything may hopefully see everything.

To be clear here machine learning is not a bad technology. LLMs are useful, all of this technology should be useful and would be if it had been kept in a lab until people knew what it could do. The creation of a metasystemic AGI is not a bad thing, and I would posit that the responsible and ethical creation of such a thing through mechanisms that preserve agency would be precisely what humanity needs to survive the great filter.