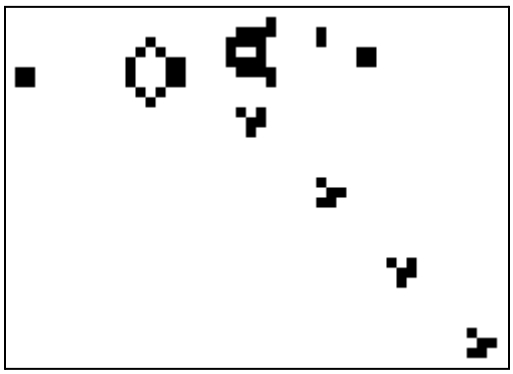
Virtual Anthropology: When and What Could We Learn From Multimodal Agentic Behavior in Generative Worlds?

Fabian Kerj — 2024-12-04

Right now, in some college dorm or late night afterparty, there are two stoned people asking each other "*You think we are living in a simulation?*", "*Stop it man, that gives me anxiety*" says the other. And although the question is overused, and was popularized through the likes of Musk, Ready Player One, and of course The Matrix; there are some tantalizing developments in technology that should lead us to more closely examine such structures. If we resist the question on *whether we actually live* in that artificial world, whatever that means, and instead look at the *contemporary* building blocks for being able to construct them, we could take an earlier step to derive, in part, what we might ascribe the creators of such worlds' *predictive motives*. When trying to answer the college dorm question of "*why someone/thing would create simulated worlds of agentic behaviour*" - the first line of reply would be to better predict, or perhaps even, on a perceived sadistical level, entertain the observer. In light of the complex, emergently agentic-like characteristics of LLMs, particularly multimodal LLMs, image synthesis and three-dimensional spatial generative Ai, we should seriously consider how and what insights we might *already (and yes— nearly today)* derive from a combination of them.

The technological components for creating sufficiently interesting approximations of simulated worlds are rapidly materializing. Take for example GenEx (2024), which demonstrates the ability, through agentic expectation values, to explore and predict unseen parts of 3D environments. NVIDIA's Edify provides the foundation for generating diverse visual content from 2D to 3D, while LLaMA-Mesh unifies 3D mesh generation with language models. These technologies, if combined, could offer more than just individual capabilities—they present the possibility of creating coherent, explorable worlds where artificial agents can interact, learn, and exhibit emergent behaviors. The question is no longer whether we can create sophisticated simulated environments, but rather what we can learn from them, and how far we can push their complexity and utility.

To the reader not yet convinced of an emergent *leap-level* novelty with these tools, I can partly sympathize. We have had video games for a long time, not to mention complex procedurally derived emergent dynamics. To the former, I hope I don't need to argue why *Sims* is different to the level of what I aim to elucidate, if this is unclear - I hope the content later in this paper will show why. To the latter, indeed, while physics focused chaos theory, and other rule-based mathematical systems can give rise to unpredictable dynamics, e.g. elegantly illustrated through Conway's Game of Life, seen below:



It, at least at this level, fails to capture a more accessibly human aspect of the area we aim to tackle in this paper. I.e. I don't aim to argue that the principles illustrated through the game *of life* could not, in principle, simulate an entire human, city, planet or…universe. In fact, I hold that if you press many physicists as to what they *really* pursue as a field, an answer could be boiled down to discovering the rules of the universe; while in the game of life the rules are what happens to a cell; depending on the content, or lack thereof, in its surrounding cells, the physicist examines similarly about the pseudo-arbitrary units of reality in our world, replacing black or white cells with particle fields, and time dependent evolutions through our best dynamical theories.
In addition to being radically simplified, the game of life as a *direct* analogy is flawed— it is too primordially detailed at its maxim for the current analysis. That is, at the limit, a properly constructed 'game of life' of the universe would be running on itself (with regards to the dynamical compute). Explained more wholly: if the discovery process for our fundamental laws remain bound to a continuous high-energy probe layering *ad infinitum*—a *proper* simulacrum would be inherently and impossibly unachievable – unless of course the simulacrum is itself the simulation; which in turn removes the meaning of the word simulation. Put very simply: Something which can only simulate itself; is that very thing—not a simulation. The reason for this detailed explanation is to illustrate that at some level an *approximation* must take place—the *better* the approximation the more accurate our simulacrum. However, we are bound by available and practically achievable compute; therefore, as an example; we might not necessarily need particle physics to derive macroscopic phenomena such as classical mechanics and/or thermodynamics, indeed, a hot cup of tea in a room cools to room temperature because of the temperature difference between itself and the room, and the temperature difference is because of the different average velocities of their corresponding atoms, their convergence is as a result of exchanging kinetic energy between each other at their mutual boundary, but at some level we don't need to go deeper to explain—we should set our dynamical boundaries to approximate for our experiment of interest; where is this dynamical boundary best found for agentic behaviour in spatial environments?

Let's go back to the question at hand: are there fundamentally new technologies to better model and approximate agentic behaviour today? Yes! And new components for this best approximating environment appear every week, every month, and every year. To start with: LLMs are a huge addition to this tool kit, and the arguments for why should be clear, they exhibit approximate levels of *reason,* whether its through their perceived sophisticated uptake of a query or task, or by its corpus of knowledge, or ability to draw novel analogies indicating some form of "understanding" of the analogy's referent. I do not aim to argue that LLMs truly understand their outputs, are by any means conscious, or have free will; arguing that we have those faculties is hard enough. However, with regards to recent LLMs, it simply points to their effective nature in being a tool for approximating agentic reasoning, and perhaps the best that we currently have. One might philosophically push back on this conclusion and say that our wide range of effective theories of psychology, neuroscience, and biology approximate better than LLMs for agentic behaviour—to this I argue that, although possibly true (if somehow combined effectively, which is sort of what LLMs already do), for some instances, firstly, little consensus exists on many of these theories, especially in psychology, and secondly, there is no clear practical method to synthesize and *generally* apply these for circumstances an agent might

encounter—again this is basically what machine learning models do! LLMs are inherently probabilistic and therefore arguably more human in this regard, they can radically differ in input to output, like human to human; reacting desperately differently. While you might have agreed from the outset that agentic behaviour can be best approximated with LLMs, it's important to explore the critics perspective to better prepare the following: if our current best available model of synthesized human behaviour can be found in probabilistic transformer models, with what technological layers can we best combine these with today, or in only a couple of years, to create something of meaningful interest? We don't want just two LLMs talking to each other, this has been done, and although enjoyable, we can do much more.

By introducing a visual dimension to these interactions we get somewhere closer to our experience, better yet, a spatial component—luckly for this exploration, and largely as a catalyst to this particular essay – multi dimensional generative models are evolving rapidly. What began as generating images of “A wizard in a dark forest" has turned into genuinely rendering a three-dimensional figure of a wizard, and not just a gaussian splat of a wizard from a multi angle diffused image of a wizard, but a genuine, topological, mesh constructed figure. This has in turn, caused new technologies to recognize where joints should be digitally attached, matching these up with models trained on similar topologies, in this case humans, and those joint orchestrations in combination with a time dimension, now allow us to approximately model four-dimensional behaviour of the wizard. I.e. the wizard walking, the wizard running and so on. And of course, we cannot forget the environment in which the wizard is placed. New models allow us to identify spatial relationships between objects in a given image; how close or far away something is - like a heatmap of being close or far away - this in turn helps us inform our next model, which lets us generate the topology of the space in which the wizard finds himself: the forest, as mentioned previously, GenEx and other multiangle-diffusion-esque models allow us to approximate the environment where we are viewing from, creating a 360-degree field of view of the forest, similarly apply our distance model, and so on. Given its a genuinely three dimensional space, we can attach a viewpoint or ‘visual’ input to the wizard’s eye area, and let the wizard’s hypothetical field of view be the approximator for the model stack’s inputs. We can then add a multimodal LLM to this point-of-view and give the wizard an approximate ‘understanding’ of the environment he finds himself in, as well as giving him the ability to talk to us about it. Given our earlier development, he has the fourth dimensional ability to walk forward, and discover things about his forest, and generate these new areas of the forest as we go. Importantly, we save the topology previously derived, and the wizard can return to where he was born, and perhaps even recognize it — but memory for reasoning models like LLMs or chain-of-thought-LLMs is a tricky beast, and not something I will go into for the scope here.

That was a lot, and depending on what level of accuracy this essay might have hinted at with some previously detailed passages, I fully realize that I am generalizing here — for example, what does *discovery* mean to the wizard (the agent)? We can assume that through prompting training data and multimodal functionality, the agent will explore its surroundings, or at the very least, react from ‘external stimuli’ and of course other agents. GenEx is a crucial technology to this question – it illustrates, through its video generation architecture, a fundamentally new approach to spatial reasoning in generative worlds. Unlike previous solutions which require physical exploration of a space, Genex enables what we might call "mental exploration" - taking

a single frame input and generating panoramic sequences of movement through a predicted space. The architecture, while technical in nature (taking an input image of H × W dimensions and outputting T sequential frames), represents something profound in our discussion: the ability to generate consistent, explorable spaces that maintain coherence even when returning to a starting point. When we combine this with LLMs acting as controlling agents, we should begin to see the emergence of novel behavioral dynamics. Our wizard in the forest is no longer bound by simple predetermined pathways or crude probabilistic exploration - he can "imagine" what lies beyond his immediate view, update his internal model of the world, and make decisions based on these projections before taking a single step. This mental modeling capability, intrinsic to human cognition but previously absent in artificial agents, provides a crucial missing piece in our simulation toolkit. The implications extend beyond simple navigation - in multi-agent scenarios, this technology enables sophisticated reasoning about other agents' perspectives and potential actions, creating the foundation for complex social dynamics in generated worlds. This representational leap - from reactive to predictive or anticipatory spatial reasoning - I argue marks a significant step toward the kind of rich, emergent behaviors we seek to study in these simulated environments and hope more effort goes towards developing more and more complex what I, for a lack of a better term call virtual anthropological models. What happens if we combine the frame model of GenEx with models like DroidSplat (frame to topology), and figure out how to have multiple agents interact? Create sounds to which these models can react to? Like in the GenEx preview where their agent spawns an ambulance around the street corner at the fed sound of sirens, and more. There are many aspects of this stack to be figured out, experimented with, and developed. It certainly remains a big challenge to how frame models like GenEx could properly best be combined with persistent topology and how its 'imaginative' engine could work in unison with others in the same space. Not to mention the compute cost of increasingly complex constellations and memory limitations.

I have argued that we are certainly not far away from the magical laboratory which I have aimed to inspire here; I believe these could help us drastically measure the evolution of agentic intelligence, create synthetic training data for future models, enrich chain-of-thought, and possibly even help the robotics industry. For the reader that takes my technical assumptions as misjudged or imprecise, I apologize—I hope that nonetheless, this has been a useful exercise in exploring how our most capable approximations of human-like agentic faculties might be usefully explored with new tools in the space. Imagine a world in which one could distributionally analyze possible outcomes of complex multi-agentic systems for a given perturbance; not through experimenting on real humans, but to some approximation of them. Or a world in which we discover something new about our own consciousness by observing how an orchestra of multimodal capabilities tackles an environment in which its own imaginative capacities can feed a part to its other functions, and those functions output back to the imagination, subsequent decisions, and emergent behaviour. Maybe we discover something very dark, or perhaps something incredible. What would you explore?

Thanks for reading,
Fabian