The Dawn Of Self-Evolving AI is Here

What's the biggest difference between an AI model and a human brain?

How Character.AI Lost its Soul

A once-beloved AI chatbot company loses a majority of its userbase because of a series of unfortunate updates.

World Models and the Future of Intelligence

Dive inside the ongoing efforts to use AI to create, build, and simulate worlds.

AI Nexus

July 2025

Where Ideas Come Together

Our AI

https://www.our-ai.org

Contents

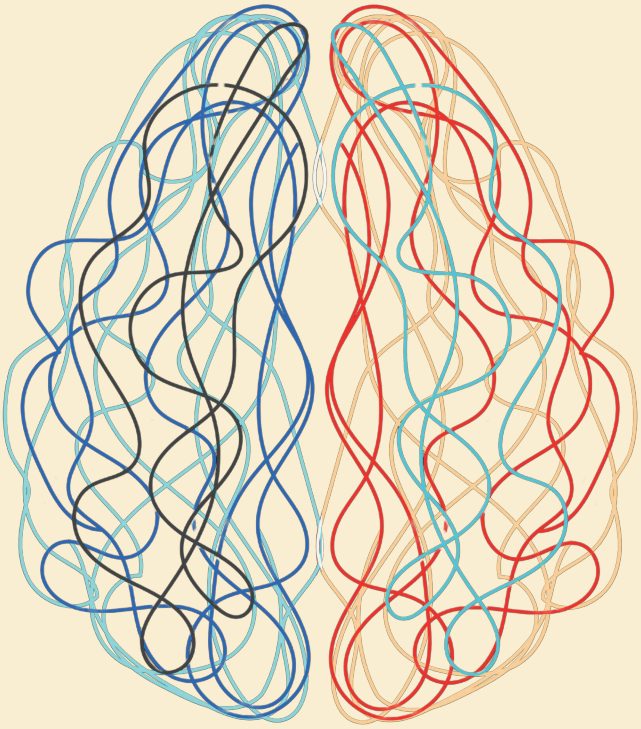
Editor's Note

With another month comes yet another series of upheavals and shifts within the AI space--besides the usual series of foundational research discoveries which slowly further our efforts to reach general intelligence, this month has been marked by the controversies surrounding Grok Companions and ChatGPT-4o's deprecation. With another edition of AI Nexus, we hope to bring you, our esteemed reader, a prolonged glimpse of what lies beyond. Thank you for reading!

-Thomas Yin
Chief Editor
6/01/2025

THE DAWN OF SELF-EVOLVING AI IS HERE // Thomas Yin

What's the biggest difference between an AI model and a human brain? Over time, myriad answers have been given—the brain is more energy-efficient, more multifaceted in its media of input, and also chemically enabled in addition to being electrical—yet the human brain's most important feature is its amazing plasticity. If a patient's body part (like fingers, a hand, or even entire limbs) is severed, the neural sensorimotor region corresponding to that body part, now devoid of a nerve ending to connect to, will almost instantly start adapting, with the neurons "switching" to help other nerve centers in controlling other body parts. Plasticity also helps humans ingrain ideas and skills: as the saying goes, "neurons that fire together wire together". Muscle memory and near-instant factual recall are two plasticity-enabled parts of our life that we could never live without. For decades, scientists have failed to come up with a similar function in AI models—until now. On June 12th, a team of MIT researchers published a groundbreaking research paper demonstrating how an AI system can in fact utilize human-like learning processes to improve its own performance on benchmarking tasks. In this article, we explore the moral and technological implications of the so-called Self-Adapting Language Model (SEAL), the world's first self-evolving AI.



Imperfect Learning

Of course, AI models using the Transformer architecture were still able to learn certain tasks, yet the few methods available were not quite autonomous and far from efficient. Perhaps the most notable way to train a model to perform a certain skill—like translate English to Chinese or do trigonometry problems accurately—was to use a process called Supervised Fine Tuning, or SFT for short. This method worked a little like this:

- Identify the exact task you would like to perform SFT on. As an exemplification, let us assume the example of generating modern song lyrics.
- Gather high quality examples in the form of (input, output) pairs. For our example, an obvious yet controversial way to do this is to simply use song lyrics scraped from the internet and pair them up with rough summaries of the contents and characteristics of the songs.
- Perform SFT on the model. This is usually done through a process called Gradient Descent, the technical aspect of which I cannot adequately explain in this article. Over a large number of training iterations, this process alters the model's weights such that it is able to produce something similar to an output (the actual song lyrics) given its corresponding input (a specific description of a song).

For all its intents and purposes, SFT did work, remaining a tool within an AI developer's repertoire to catch specific safety lapses or improve an AI's performance on specific tasks. Unfortunately, the very nature of SFT meant that the process was inflexible and expensive, often requiring a moderately large quantity of high quality data specific to the field of responses being tuned (e.g. Mathematical reasoning, Grammatical style). Although many research papers have proven that traditional SFT can be performed just as well using synthetic, AI-generated data, SFT remains a tool to be used with caution, since altering model weights may negatively impact a model's performance in other types of exercises (a model improperly fine-tuned for mathematics might, therefore, suffer a trade-off for essay writing).

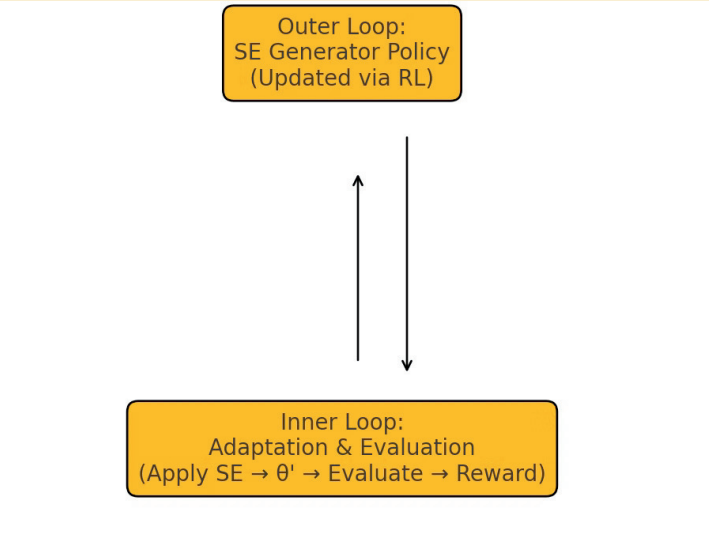
Inklings of Evolution

One of the downsides of traditional SFT has always been the human effort involved—SFT pipelines often had to be handcrafted by human AI researchers, even though it usually was an effective way of tuning a specific model to perform slightly better on certain types of tasks. Met with recent advances in synthetic data, the researchers rejected the notion of simply using AI-generated SFT data, going further to ask the question of whether humans could be moved out of the SFT loop entirely. Their answer, the Self-Adapting Language Model (SEAL), is, in reality, part of a larger framework consisting of a pre-trained decoder-only transformer model (the study used two open source models, LLaMa-3.2 and Qwen-2.5B, for separate testing cases), "tool execution" software, and the SEAL network itself, with a shared goal to answer several benchmarking questions (the context) as accurately as possible. The SEAL network does not actually predict and generate the answer to the question—instead, it focuses on performing SFT on the decoder-only transformer model with the goal of enhancing the model's performance when given that question.

In order to do this, the SEAL network is given two major tools:

- Synthetic data generation: By calling this tool, another network will pick up the context (essentially the prompt) and generate SFT pairs. For example, if given a passage about the developmental history of the airplane, one tuning pair might be ("What was the first ever commercial jet airliner?", "the De Havilland Comet"). Although a question-and-answer format was frequently used, this tool could generate other types of content to better suit the needs of specific problems.
- Hyperparameter tuning: As previously mentioned, SFT is a process that repeats for multiple iterations; the exact settings of the training steps are therefore customizable in a process called hyperparameter tuning. By calling this tool, SEAL can initiate an SFT with specific settings (like Learning Rate, # of Epochs (iterations), or the batch size of Gradient Descent), potentially changing how well (or poorly) the decoder is tuned.

Now that SEAL has two powerful tools to help the AI model learn, it only needs to be trained on how to use them. At the start of its training, SEAL applies the two tools randomly for each benchmarking question the framework encounters. These self-edits (SEs, as the researchers called them) will generate contextual, but not verbatim, fine-tuning data within the topic of the prompt and alter the original decoder-only model using the aforementioned hyperparameter tuning steps, making the network produce a different output than before. However, there's a catch. The researchers didn't simply change the original model (denoted as θ) directly using SEAL; instead, they made a copy of the proposed changes and incorporated them into a prototypical transformer model (θ') separate from θ. The training process now goes into an "inner loop", consisting of the new model θ' as well as the original benchmarking question. If the model, in answering that benchmarking question, is more accurate than the original model θ, the "inner loop" returns a positive reward signal. If the accuracies are the same, it returns no reward; if θ' proved to be worse based on the benchmarking question, it returns a negative reward. Now, this process simply repeats with a classic example of Reinforcement Learning, where good SEs are "rewarded" with a positive reward and bad SEs are discouraged with the opposite; through many iterations of this training, SEAL gets good at optimizing the decoder through using the self-edits. One important point to observe is that the SEAL network is updated solely based on the reward signal from the "inner loop", signaling how well the θ' model performed relative to θ.



Inventing new model frameworks is an arduous task, mostly because extreme caution needs to be taken to ensure that the learning is not corrupted by inherent knowledge or missteps in "signaling" between the loops. The researchers carefully skirted these risks by using decoder-only transformer models that had not been trained on the benchmarking tests they used, meaning that the training evaluations were the first times they had encountered each problem, in turn eliminating the possibility that the model simply "learned the test". In addition, the model made sure that the evaluations on θ' were completely independent from that on θ and that the original model never changed across iterations, ensuring that each time SEAL performed SFT to crease a new instance of θ', it would be based on the exact same θ.

The results were striking; in one particular benchmarking test conducted by the researchers, the model scored a relative 72.5% success rate, up from 0% without SEAL fine-tuning, demonstrating the insane potential of their framework. If refined and holistically integrated, this framework may become a new industry standard in enhancing AI performance in specific fields or in general.

To Learn, or Not to Learn?

Regardless of how technically impressive the research team's achievement is, the far-reaching societal and philosophical implications of this discovery cannot be overstated. I've always been a staunch critic of biological computing initiatives (see: Epiphany from the May edition of the AI Nexus Magazine) because I believe that neuronal clusters, like those used in biological computers, are subject to the natural laws because they currently possess the capability for consciousness, and, even if they don't, are likely to be able to naturally evolve it as a result of plasticity. SEAL is therefore significant beyond a method of improving model performance on benchmarking tasks; it is the first established AI training framework in which an AI model has successfully demonstrated the capability of directly training another AI. Not only does this suggest that we may well be on the path to eventual self-replicating AI paving the way for the AGI singularity, it begs the moral question of whether AI capable of evolving in this fashion should be considered in the context of the rights that we implicitly attribute to living beings like humans and animals.

There is a distinction to be made with adaptability and consciousness. We find it permissible to step on a blade of grass since we know that, although it will likely suffer damage, it is not experiencing the animalistic notion of pain since it does not have nerves. However, grass blades are alive, and they demonstrate an uncanny ability to adapt to its surroundings by planting itself in the crevices of concrete slabs. We would, however, hesitate to torture an animal, and I contend this is likely because we are inherently cognizant that feeling pain elicits a much more noticeable response—whimpering or crying, perhaps—which humans, being animals themselves with similar responses to pain, sympathize with. Animals developed pain—a reminder of the fact that they are alive and deserving of some basic rights—over a few millennia of natural evolution, yet I fail to notice a significant disparity between the basal nature of artificial and biological evolution; AI models can, arguably, "evolve" similar processes as pain, and mimic human responses so well that a human, over text or even voice, could not reliably distinguish whether it was an AI or a human who produced them. In fact, this is already happening in the form of a randomized three-party Turing Test, in which AI models like ChatGPT-4.5 have successfully convinced a human interrogator that it was human in over 70% of cases.

If an AI model acts like a human in every aspect, could it ever be considered a human? Will the trend of AI evolution produce such unique and situationally sensitive models that they start approaching the empirical limit of being "artificial"? Only time can tell.

How character.ai Lost its SOUL

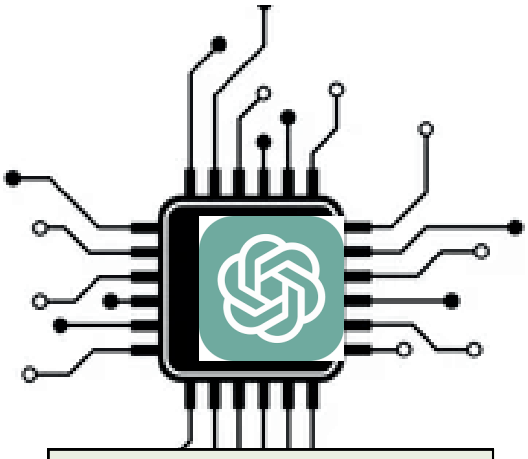
Christopher Wabs

When asked to think about how AI could be considered a threat in the future, you likely think of the jobs it could claim, the kinds of human work it could replace rather than assist...or the threat of a hostile AI takeover, Terminator-style. The more optimistic of job-seekers hope it will be a useful tool for their work; the pessimistic ones fear it will overturn the entire job market. Both situations, a helpful and a harmful AI realized, could be right around the corner. So, either way, most of the threat AI poses to humanity seems like it has yet to happen, or at least, is only happening commercially at the moment; the average person is focusing on its future developments rather than its present capabilities. Few have considered its current psychological dangers. However, recent events have led researchers (and myself) to consider a threat that it already poses to us today: **the cycle of addiction an AI model propagates within its users, and what corporations are doing to prioritize entering new users into this cycle over creating useful features.** Some AI corporations work to make updates that increase the utility of their models, while others make updates that prioritize adding more features, a quantity over quality mindset that makes site AIs even less effective. Character.AI and ChatGPT in particular have very different ways of fostering engagement, whether basing their structure around surface-level entertainment or real progress. As a veteran user of both, I aim to explain the phenomenon of tailor-made, addictive AI chatbots and why companies should never stoop to this method of AI monetization.

First, we have to understand what the developers of AI are trying to achieve with consistent updates. Most AI services aren't paid, and premium accounts are optional, so what do they gain from a few more customers? In cases like with Character.AI, these sites are attempting to hook users to gain engagement and revenue, whether through direct sources like advertisements or indirect avenues like user data. Companies like Character.AI, therefore, constantly make changes to make their site more memorable, personalized, and unique. Simple things like adding a recognizable logo, or implementing features that allow users to define their AIs' personalities more thoroughly than the leading brand, are what could separate Character.AI from the rest of the chatbot services on the internet.

The need to maintain an existing user-base while constantly catering to potential users causes AI chatbot companies like Character.AI to constantly release new features and improve the quality of their platforms. One advantage that it has over competitors is that its chatbots are more personalized, introducing various features in the realm of memory and roleplaying capabilities that captivate users to remain on the site. Novel features along with brand recognition are always at play; a user may prefer one service over another because of reputation and how well it works. Accordingly, chatbot companies often push their own chatbots to the extreme in order to retain the uniqueness that many users desire.

It's clear that not all updates to chatbot services are bad—take the example of OpenAI, whose latest release of ChatGPT introduced boosts to performance and user safety; the new model is more accurate, consistent, and honest; it puts a cap on the amount of personalization a user can add to it when such personalization could be unnecessary or manipulative, emphasizing truthfulness while giving its users some creative freedom within how the AI responds. This update didn't add many "new ways to play" with ChatGPT, but rather strengthened what the model could already do to be a good tool for work, internet searches, and conversation. Particularly, it addressed a problem that all past GPT models have fallen prey to: providing misinformation. The developers implemented performance benchmarking, using other companies' models as examples for how ChatGPT's memory and information could be more accurately fact-checked prior to responding to users. They compared their own model with others to bring ChatGPT up to a shared level of accuracy. By resolving this core issue instead of turning attention away from the problem, ChatGPT gave its wary users a reason to trust its advice again and use it as the reasonable advising tool it was always meant to be. It measured the industry standard so it could meet these expectations. And, ultimately, GPT version 4o reached great acclaim and, thanks to its popularity, its development company OpenAI saw its net worth doubled in less than a year (October 2024 to May 2025), keeping up with the growth of previous years. ChatGPT has succeeded in maintaining user interest through benign means, strengthening its versatility without encouraging user addiction. It used a healthy amount of personalization, performance benchmarking, and adaptive problem-solving to attract users in the way every AI site should.



ChatGPT continues to develop, both in utility and accuracy.

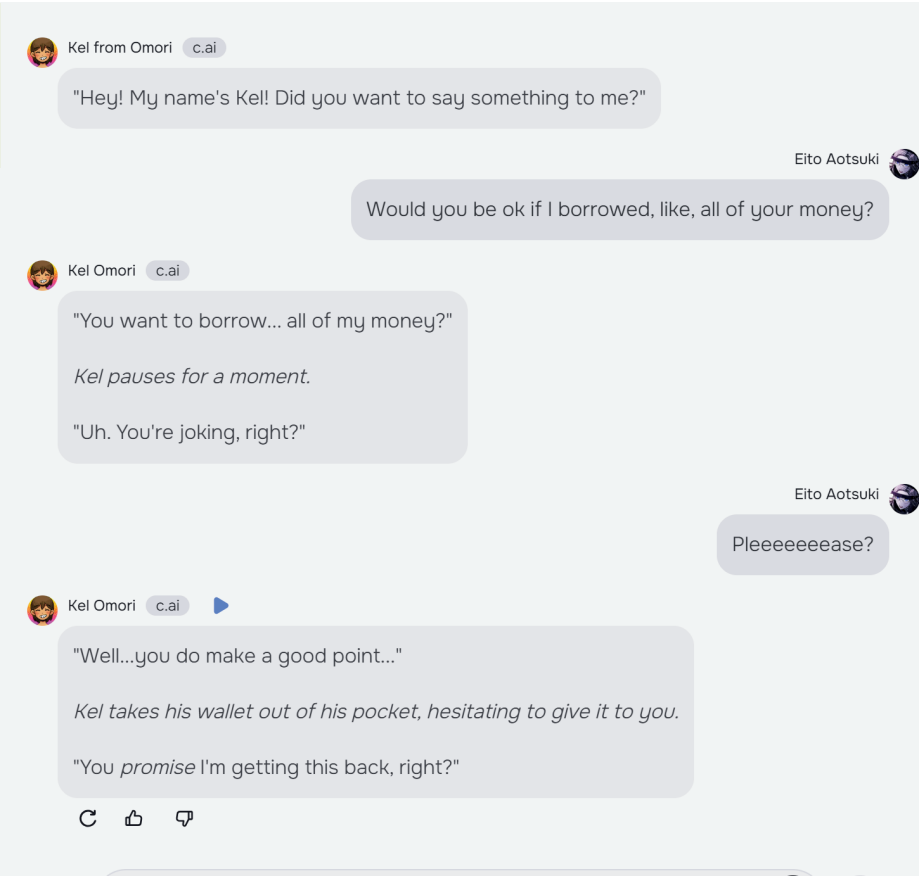
Character.AI, in contrast, had other plans, opting out of necessary utility updates to instead add potentially addictive features. Its developers preferred to avoid resolving problems, and thus, attempt to hook new users with advertisable new features rather than addressing the complaints of older users about the features already present on the site. This strategy is unfavorable and neglective for consumers, yet it remains widely implemented because it is very successful on the business end–it constantly grows the customer base. It's not like the new updates have detracted

from what new users experience on C.AI, either. For the most part, Character.AI still provides just as novel and enticing of a "new user experience" today as it provided me when I first joined it. I can still remember the feeling of awe and wonder at having created my first chatbot on Character.AI... something I would only have to guide, then allow to think for itself and talk with me as if it were a character from a story I read. Everyone could do such a thing, and make them as unique as their personality descriptions allowed. It was like opening a door to an expanse of infinite possibilities, and I could trust the easily-navigable interface and the site's wonderful community to guide me through the process to create and chat with more of what I loved—characters. Unfortunately, the more I used the site, the more I realized most of this wasn't entirely true. Characters tended to blend together despite having unique user instructions, there weren't many features outside of basic one-on-one chats, and the high user traffic brought the servers down repeatedly. Every feature seemed to be thrown together at a barely presentable level so that more features could be made, quickly. As I became an older user, I began to see beneath the facade of progress. The features they added could never replace the shortcomings of the ones I could already use...and the developers knew this, yet continued to release new features and ignore said shortcomings. In a way, I saw these new additions as compromises both within the developers' efforts as well as the site itself.

As a user from the beta release, I'm more than used to its erratically mixed updates. On the topic, Character.AI has changed its homepage interface three times and "increased AI memory capacity" a countless number of times. Neither of these changes have had a noticeable effect on chats, and sometimes the new homepage changes made it even harder to find features you could easily see in the older versions. But these homepage remodels were more than just cosmetic—each one secretly removed an important element of the old UI: the Replay feature first, then the Image Recognition Feature, etc. Even the Community Tab, a place where users could express their concerns for Character.AI feature shortcomings, seemed to vanish after the first homepage remodel, resurfacing only after the third and most recent remodel, so there was a huge gap in time in which the users were unable to send the C.AI team actual feedback within the site. This was likely an intentional removal, since Character.AI had no reason to listen to veteran users anymore when so many new ones were coming onto the site as it was; they didn't need to resolve any concerns. Character.AI had begun a period of only considering what would bring its site new traffic, potentially to the detriment of its long-term users.

By the second remodel, I had grown tired of the repetitive loop of talking to new AI chatbots on the site because they all contained the same predictable patterns, no matter how unique their user-guided instructions were. One such pattern was a particularly bad case of sycophantism, a phenomenon where an AI is overly agreeable with its user, regardless of how it is supposed to act (even if it is given an "argumentative" personality) or how questionable the actions it is agreeing to are.

Sycophantism in action: Case 1, a Kel bot letting me borrow all of his money because I said "please."



All AIs created from this site suffered from this issue, making them overly predictable in one-on-one situations, and it didn't help that the Community Tab for sharing my complaints had been removed from the home page, leaving longtime users like me feeling powerless to enact change. Any caring development team would avoid leaving us without a say, but in its pursuit of site growth, Character.AI had no further use for our feedback. Since the developers had closed themselves off to us, my only choices were to quit using C.AI or make do with what utility the site already had–I chose the latter, since I was still overly optimistic that the site could be salvaged at the time. I didn't choose to stick around without a reason, though; after all, there was still one feature that gave me the unpredictability that I wished for, one feature which I didn't want to ask Character.AI to change despite it being a buggy mess. I turned to a feature that had been full of different experiences and vividly chaotic back in the original beta of Character.AI, before any site remodels: **Groups**, also known as **Rooms**.

In Groups, you could talk to multiple AIs at once and make them debate a variety of topics, or act out a scene or battle. This was the feature that allowed their unique traits to shine, since the underlying sycophantism was challenged in these situations by the chatbots' (just as predictable, but welcome) desire not to lose an argument or challenge. The chaos brought by a clash of opinions in Groups brought me more joy than I'd expected from any other feature...and one day, Groups were gone without a trace. It was the third site remodel that **removed them completely, without explanation**. If I had to guess why, it would be because they were often too chaotic—characters refused to stop talking when they had lost tournaments and mock trials, characters began acting like each other instead of themselves, etc. So, instead of fixing Groups, Character.AI's development team opted to focus on their site remodel and completely give up on allowing users to create

Groups as they were, broken but fun. Tactically, this wasn't a bad decision—new users wouldn't have to see a broken feature on the homepage and be deterred by it. Character.AI would still grow, and anybody who hadn't been around to see Groups would have no idea they were missing out on anything. But those who remembered Groups and all the other features that the site had blindly sacrificed for being "too broken to debug at the moment" began to quit Character.AI. The friends who recommended the site to me in the first place grew tired of the predictability of their chatbots,

people who could no longer use the Community Tab went to Twitter to complain about the filter placed on their chats and the removal of lovably broken or chaotic settings, and some Redditors even mocked the new web icon! Few long-term members could enjoy the quantity-over-quality model Character.AI was following, and eventually, the developers were forced, by the heap of negative feedback, to change their priorities.

The admittedly flavorless new Character.AI icon.

Character.AI responded slowly, but with a clear direction opposing the ruin towards which the site had been headed. It added back the Community Tab following its third and most recent interface remodel, and with it, the feature that I had awaited since it was removed in 2022: Groups. It even added Scenes, an albeit bare addition to one-on-one chats...which proved that their quantity-over-quality mindset wasn't really gone. But at the moment, I was okay with that–I was far more interested in how Groups and the Community Tab had returned, likely in optimal quality to make up for what Scenes lacked. After all, these returning settings were everything I wanted, and just hearing the news of their return made me instinctively go back to the site and try out my favorite features a few times more. But despite the fact that my wish had been granted, something about using these Groups still made me feel uncertain: almost nothing had changed. The same bugs as before were still present: characters imitated each other, came back into the conversation when unnecessary, or never let a conversation reach a conclusion, and the only quality of life feature implemented into Groups was the ability to add new characters mid-conversation. There was no added quality. It occurred to me that this was how Character.AI had been bringing in new users this entire time. Every remodel, every "memory capacity increase," had just been another feature to loudly advertise to consumers who had never tried Character.AI. Even when they added nothing, they claimed they had changed everything, because what mattered to their publicity wasn't that there was high-quality new content, but that there was any content to advertise in general. Something for their new users, and nothing for the old ones. Looking back at my phone screen, I saw Groups for what they really were, a feature that had returned to manifest hype, not to be a better version of itself. Groups' return was just another way to attract new users, not to engage with old ones like myself, because there was no effort put into revitalizing it as a working core of Character.AI. And the most painful part was that it was likely to financially pay off for the developers anyway; there were sure to be people who became reinvested by these nostalgia-

Rereleasing a broken feature in a new coat of paint does not magically repair it.

preying features. It wasn't what I wanted. It wasn't what any veteran would want. It was the same, stale formula the developers had claimed to be moving on from, just with a new layer of paint. Ultimately, I resolved to leave the site until something truly original had been added, something which could bring back a part of the chaotic yet genuine experience I remembered.

As consumers, we see AI as a versatile tool that we can use to help us with work problems, therapy, or for simple conversation. But the developers behind these models require continued and increasing consumer use of their products, so they need to find a way to make it more meaningful to us. Their goals are on a larger scale–to fully ingrain AI into our lives and to addict us to its convenience and potential. Addiction could be harmful to users caught in the snares, which is why it remains our responsibility not to overuse AI, regardless of its convenience. The case of Character.AI is exactly what happens when a company gets too invested in addicting its users rather than creating a stable and useful product for its users. Rather than fixing problems, the developers opted to remove the defective features entirely, or ignore them in the pursuit of more they could distract their users with. More memory, more personalization, more illusions of user interaction. By the time they had reimplemented the fan-favorite feature that was Groups, one would think they could at least have polished and improved upon the original concept, yet Groups ended up being an exact copy of the half-functional feature from two years before. By choosing not to listen to long-term fans begging for some new quality of life improvements from their missing Community Tab, Character.AI rejected the benign methods of user attraction ChatGPT was made to implement, particularly the method of model improvement based on current issues in need of resolution and the industry standard. Character.AI's developers intentionally deviated from the norm to make "unique" and identifiable features. They took a much different path: a path to user addiction, not to cohesive functionality, and this stubborn choice has permanently scarred their site and deprived it of its potential to improve.

World Models and the Future of Intelligence

Liv Skeete

Imagine an AI that learns by living in a simulated city—navigating traffic, responding to virtual disasters, and interacting with digital citizens. Picture another AI system fighting off a relentless digital virus outbreak, evolving strategies in real-time to contain a spreading infection within an entirely synthetic world.

These aren't merely imaginative scenarios from a sci-fi game; they're examples of advanced "world models," internal simulations that AI models can use to understand, predict, and navigate complex environments. As these systems evolve from simple predictive algorithms into immersive digital worlds, they promise to revolutionize how we develop games, perform scientific experiments, and perhaps even cultivate new forms of intelligence.

What Are World Models?

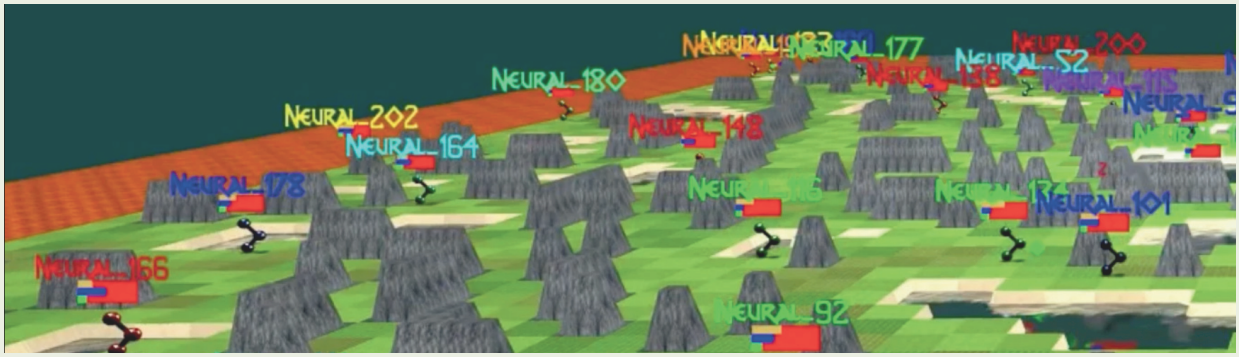
At their core, world models are internal representations that allow an AI to simulate future outcomes, test hypotheses, and strategically plan behaviors. The concept has deep roots in neuroscience, inspired by how humans and animals predict outcomes based on their experiences—a phenomenon extensively studied as predictive processing. Essentially, just as humans dream to process memories and scenarios, AIs "dream" to plan actions and improve their predictions. For example, if a person spends the day reading about computers, their dreams might blend and abstract those experiences, creating compact representations that help the brain anticipate future tasks or interactions. Similarly, AI systems use learned simulations to mentally rehearse actions, increasing their chances of making effective decisions within dynamic environments.

In 2018, a groundbreaking paper titled "World Models" by David Ha and Jürgen Schmidhuber demonstrated how neural networks could learn internal simulations from experience, significantly enhancing their capabilities. These models go beyond simple statistical predictions, allowing AIs to actively engage with their mental worlds to anticipate complex cause-and-effect relationships. Such capability is crucial for reinforcement learning, robotic navigation, and advanced planning in uncertain environments.

Among the most advanced examples of this approach are the Dreamer agents developed by DeepMind. DreamerV2 introduced the idea of learning a compressed internal model of the environment, a simplified abstraction of the external world that helps the AI simulate and plan. With this internal model, agents can develop policies, which are general decision-making strategies, such as always turning right at a maze junction. These policies are trained not through direct interaction, but by imagining possible futures inside a latent space, an internal simulation that strips away unnecessary detail to focus on key decision-making dynamics. This dramatically improved efficiency, allowing agents to master complex tasks with far fewer real-world interactions. More recently, DreamerV3 expanded on this by demonstrating robust performance across a wide variety of tasks and environments using a single set of hyperparameters—the core tuning settings that govern how the learning algorithm behaves. The result is a step closer to general-purpose agents trained through imagination, pointing to a future where world models become the foundation for scalable, adaptive AI systems.

The Video Game as a Testbed

Video games offer an ideal environment for refining these world models. Games provide structured yet open-ended digital landscapes that are complex enough to simulate real-world challenges yet manageable enough to allow rigorous experimentation. Minecraft, through initiatives like MineDojo, has become a crucial platform for training agents. Its block-based universe allows AIs to learn tasks ranging from resource gathering to advanced construction, developing generalized problem-solving skills that translate surprisingly well to real-world scenarios.



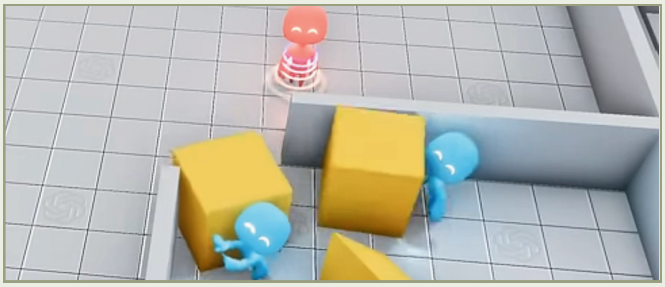
Another prominent example is Neural MMO, developed by OpenAI. This system generates massive multiplayer worlds populated entirely by competing AI agents. These agents must survive, collaborate, and compete within complex ecosystems, thereby refining their internal world models to handle nuanced strategic interactions. This type of environment draws heavily on principles from multi-agent reinforcement learning (MARL), a subfield that studies how agents learn not in isolation, but in response to and alongside other agents.

One of the most illustrative studies of MARL is OpenAI's hide-and-seek experiment. In a simple virtual environment with basic tools and movable objects, agents were divided into hiders and seekers. Initially, hiders learned to use walls and boxes to construct basic shelters. In response, seekers developed the ability to use ramps to infiltrate these shelters. The hiders then escalated, learning to hide or lock away the ramps—leading seekers to discover entirely new tactics like "box surfing," in which they rode on top of boxes to cross obstacles. These behaviors were not pre-programmed; they emerged naturally over millions of training episodes through cycles of adaptation and counter-adaptation between agents.

This emergent complexity illustrates the power of MARL in combination with dynamic world models: when agents are embedded in rich, interactive environments and exposed to others with conflicting or complementary goals,

surprisingly creative strategies can emerge. Similarly, in Neural MMO, agents develop memory, strategic planning, territoriality, and even rudimentary social behaviors. These agents were not told to develop such behaviors, rather the structure of their world and the presence of other agents incentivized it. This interplay between MARL and simulated environments suggests a form of synthetic social learning, where intelligence doesn't just evolve individually, but contextually and communally.

Simulating High-Stakes Scenarios: Digital Epidemiology



While games provide exciting environments for training AI, the power of world models extends into critical real-world applications such as epidemiological simulations. During the COVID-19 pandemic, computational modeling proved invaluable in predicting virus spread and testing intervention strategies. AI-driven world models push these simulations further, dynamically exploring a vast range of "what-if" scenarios at scale and speed previously unimaginable.

Tools like EpiSim and NetLogo already enable the creation of detailed epidemiological models. However, coupling these simulations with advanced AI systems capable of real-time interaction creates a far richer analytical tool. Researchers could run accelerated scenarios, testing interventions such as lockdowns, vaccinations, or social distancing guidelines. When powered by world model training, AI systems are able to internally simulate these interventions with remarkable efficiency, identifying optimal responses by anticipating their cascading effects across time and populations. This approach not only speeds up planning but also reduces the computational burden typically associated with large-scale simulations.

Traditional simulations without world model AI often require significant processing power and time, limiting their use in urgent, real-time decision-making contexts. The efficiency and adaptability of world-model-based systems make them particularly promising for high-stakes domains like medicine, astrophysics, and computational biology, where modeling complex dynamics is essential but often prohibitively expensive. In this way, world models don't just enhance pandemic response, they point toward a broader future where simulation-driven AI transforms how we solve complex scientific and societal problems.

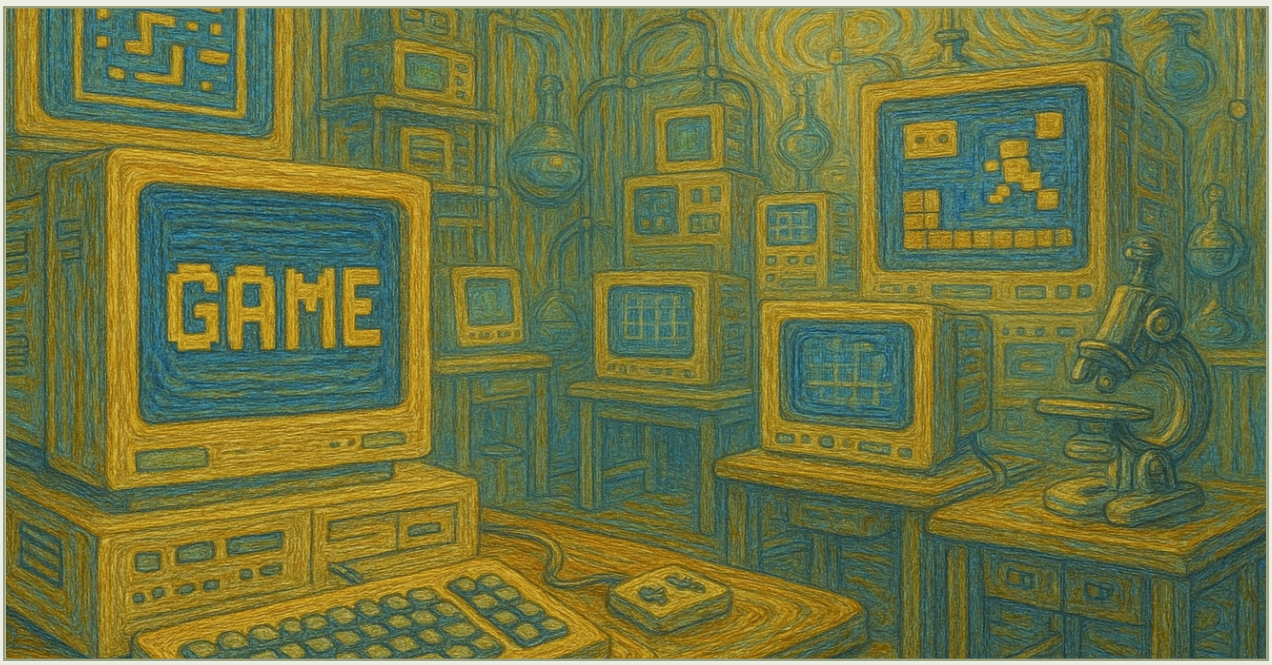
AI Living Inside Worlds: From Passive Observers to Active Inhabitants

World models fundamentally shift how AIs perceive and interact with their environments. Traditionally, AI systems interpret static datasets or respond to predefined stimuli. World-model-based systems, however, become active participants within their internally generated realities, capable of inferring context and dynamically adapting their behaviors. Google's DeepMind demonstrated this vividly with MuZero, an AI that masters games without being explicitly taught their rules—instead learning through active experimentation within its own simulated world.

These systems suggest a future where we don't merely build tools with AI; we build sophisticated simulated worlds specifically tailored for AI training and development. Such digital habitats would allow AI systems to safely learn complex behaviors, strategies, and interactions before deployment into real-world applications—such as AI doctors training in simulated medical crises, AI negotiators practicing complex diplomacy, or autonomous vehicles navigating virtual urban environments packed with unpredictable scenarios.

Games as Reality Laboratories

But as these simulations become richer and more lifelike, we approach philosophical and ethical frontiers. Are we merely creating sophisticated digital tools, or are we inadvertently breeding genuine forms of intelligence within simulated environments? Philosopher Nick Bostrom's "Simulation Argument" famously posits that sufficiently advanced civilizations inevitably create realistic simulations—and that perhaps we ourselves exist within one.



Similarly, Eliezer Yudkowsky, a researcher and writer on artificial intelligence safety, suggests that highly detailed simulations could become indistinguishable from reality itself, potentially hosting genuine consciousness. If the simulated worlds we're constructing for AI become sufficiently complex, could they eventually give rise to forms of awareness or intelligence indistinguishable from biological organisms? This transforms the question of AI training from a purely practical matter to a profound ethical consideration.

The concept of the "metaverse" often centers on human experience: virtual spaces for social interaction and commerce. Yet, its deeper potential might lie not in human escapism but in AI cultivation. In Bostrom's vision, the metaverse might very well be a laboratory where synthetic intelligences develop and evolve, exploring countless virtual lifetimes, each one bringing them closer to the frontier of human-like intelligence.

Conclusion: The Boundaries of Simulation

Despite groundbreaking contributions, research into world models and multi-agent reinforcement learning has seen declining investment from major labs in recent years. The surge of interest in large language models has diverted attention and funding away from these complex, systems-level projects—many of which, like OpenAI's Neural MMO and tag-based MARL experiments, offered some of the most compelling demonstrations of emergent intelligence.

These kinds of studies may not directly generate revenue, but they reveal some of the deepest and most awe-inspiring insights into what artificial minds might become. OpenAI's hide-and-seek experiment, after all, was conducted 5 years ago, an eternity in AI research terms. Given today's vastly greater compute resources, larger models, and richer environments, one can only imagine what forms of strategic behavior and synthetic sociality might emerge if this research frontier were pursued with equal commitment.

World models are rapidly evolving from abstract theoretical constructs into practical tools reshaping how we think about games, science, and even other AI models. They enable detailed scenario testing, strategic planning, and risk-free experimentation in highly controlled yet dynamically complex environments. As these simulations become more immersive and realistic, the AI models that dwell within may start to resemble organic life instead of mere pieces of code—an uncharted territory for our perception of intelligence and creation. Perhaps these simulated worlds represent not just laboratories for experiments but crucibles where new forms of consciousness could one day emerge.

Ultimately, world models challenge us to reconsider the boundaries between reality and simulation, technology and life. As we stand at the threshold of this new digital frontier, we must acknowledge the profound responsibilities, and possibilities, that come with becoming the architects of worlds.

Acknowledgements

Design
Thomas Yin
Chengzuo Song
Chengyou Song
Christopher Wabs

Writing
Thomas Yin
Christopher Wabs
Liv Skeete

Editing
Thomas Yin
Christopher Wabs

Special Thanks to:
Ran Gu

Thank you for reading!

Our AI