The Complex Universe Theory of AI Psychology

Andrew Tomazos - andrewtomazos@gmail.com - www.tomazos.com

1. Introduction

We describe a theory that explains and predicts the behaviour of contemporary artificial intelligence systems, such as ChatGPT, Grok, DeepSeek, Gemini and Claude - and illuminate the macroscopic mechanics that give rise to that behavior. We will describe this theory by (1) defining the complex universe as the union of the real universe and the imaginary universe; (2) show why all non-random data describes aspects of this complex universe; (3) claim that fitting large parametric mathematical models to sufficiently large and diverse corpuses of data creates a simulator of the complex universe; and (4) explain that by using the standard technique of a so-called "system message" that refers to an "AI Assistant", we are summoning a fictional character inside this complex universe simulator. Armed with this allegedly better perspective and explanation of what is going on, we can better understand and predict the behavior of AI, better inform safety and alignment concerns and foresee new research and development directions.

2. Foundations

2.1 The Mind

The dictionary definition of the word *mind* is "the part of a person that makes them able to be aware of things, to think and to feel". In 1637, René Descartes famously declared "Cogito, ergo sum" - "I think, therefore I am". He made the observation that the only absolute fact one cannot doubt is that one's own mind exists, as it would be a logical contradiction for it not to. How could one even consider the question of whether one's mind exists, if it doesn't exist? One can't, so one is left in a unique and surprising situation where the very act of considering the question provides and proves an answer to that same question: yes, one's mind definitely exists.

However, everything else one experiences, including other people, other minds and the whole universe, could conceivably be some kind of dream. One's own mind's existence is the only absolute certainty - everything else is an educated guess to varying degrees of confidence.

2.2 The Real Universe

With that philosophical caveat acknowledged, let us define the *real universe* as the things that exist independent of the mind. If we somehow isolated and destroyed all the minds in existence (and only the minds), the real universe is exactly that which would survive such a calamity. This is sometimes called the physical universe, the observable universe, nature or the natural world. Some commonly believed examples of its components include the Sun, the Moon, Mount Everest, the Pacific Ocean, a proton, an electron, the force of gravity, light, horses and trees. For most people, it's what they are referring to when they say something exists in reality, or something that is real, objective or concrete. It's what the laws of physics are attempting to explain. It's what the sciences seek to test and predict.

Per Descartes' observation, the real universe, as we have defined it, may be an empty set - but this will render no impediment to how we will use the term.

2.3 The Imaginary Universe

In contrast, let us define the *imaginary universe* as the things that are dependent on the mind. The set of things that exist only within the mind. The content of the mind. If all minds were destroyed, the imaginary universe is that which would be destroyed along with them. It consists of products of the imagination and thought: abstract concepts, feelings, patterns, ideas, emotions, dreams, mental models, memes, fictional stories, the various components of the human condition and even consciousness itself. It ranges from perceptual things such as subjective experience to things that are completely divorced from reality such as a wild dream about an alternate dimension.

2.4 The Complex Universe

Borrowing some terms from mathematics, we now define the *complex universe* as the union of the real universe and imaginary universe. If the complex universe is a book store then the real universe is the non-fiction section and the imaginary universe is the fiction section.

For certain things, whether they are part of the real universe or imaginary universe is hotly debated. In a way, that is what Descartes was getting at: What's real and what's imaginary? For example, it is often asked whether mathematics is invented or discovered. We can reframe this as asking whether mathematics is part of the real universe or the imaginary universe. Another common example is God: Theists generally hold that God is part of the real universe and atheists generally suspect God is part of the imaginary universe. Another more obvious example is an accusation or a hypothesis. These can be viewed through the lens of our categorization as statements that, if true, are part of the real universe, or otherwise, if false, are part of the imaginary universe.

But let us observe that while it may be debatable whether a certain thing is part of the real universe or imaginary universe, there can be no doubt as to whether any given thing is part of the complex universe - as it is the union of both. With a simple logic of the excluded middle, any given thing is either dependent on the mind or it isn't. If it is dependent on the mind, we define it as part of the imaginary universe. If not, we define it as part of the real universe. It follows in either case, the given thing is part of the complex universe.

To our previous examples: Regardless of whether mathematics is invented or discovered, it is part of the complex universe. Theists and atheists both agree that God is part of the complex universe. Accusations and hypotheses, even if false or incorrect, are conceivable - and so are part of the complex universe. This reveals the utility of the concept - we have formed a convenient handle on the notion of "everything" in the broadest possible sense. If we had instead referred to "the universe", this could be misinterpreted to mean only the real physical universe. The complex universe includes not only everything that is real, but it also includes everything that has existed in any mind.

We would like to highlight that we have not made any claim so far. The complex universe is purely a definition. We are not claiming that the distinction between the real universe and imaginary universe is not important. In the case of a courthouse, it can be a matter of someone's freedom, or even a matter of life and death in some jurisdictions. An inability to distinguish between obvious facts and obvious fictions is considered a symptom of mental illness. Intentionally and knowingly claiming something from the imaginary universe is part of the real universe is called fraud. One of the big complaints against early AI was that they hallucinate, which is exactly mistaking something from the imaginary universe for part of the real universe. We call a daydream a hallucination when the person undergoing it mistakenly thinks what they are dreaming is part of the real universe and not the imaginary universe. We are not trying to blur the line, we seek only to create a term that conveniently refers to both worlds at once.

2.5 Non-random data

We now turn our attention to symbols, language, communication and information. We will use the umbrella term *non-random data* to refer to these. We qualify the data as non-random to indicate that we are referring to data that was recorded and encoded, as opposed to randomly-generated data. Sometimes this is described as signal as opposed to noise. Some might argue there is no such thing as randomness outside of (or even within) quantum effects, but that's not a rabbit-hole we need to venture down. If there is no such thing as randomness, then all data is non-random data, and the term remains suitable for our purposes. The father of information theory, Claude Shannon, defined information as that which reduces uncertainty. This is the sense we mean. We consider non-random data a synonym for Shannon's well-accepted conceptualization of information. Information is about whatever it informs its reader of, the specific uncertainty it reduces.

All non-random data has certain semantics, has a certain meaning, and describes something. For example, it is clear that the words in this document are attempting to describe something. That description happens to be encoded in the English language. An image from a digital camera (photograph) captures and quantizes (rounds to a discrete grid) the photons at a particular point in spacetime and records them in pixel data. Similar to a video camera and a moving picture over a span of time. Similar to an audio recording and sound waves. A temperature sensor records how fast nearby atoms were vibrating at a particular place and time.

In a similar fashion, artworks describe some aspect of the imagination of the artist, some aspect of the imaginary universe. A novel describes a fictional story and world that exists in the mind of the novelist. A painting portrays a dream image from the artist's mind's eye. A mathematical proof describes a series of abstract logical deductions from a set of initial axioms to determine the truth of some statement. All this is non-random data too.

3 Theory

3.1 Non-random data samples the complex universe.

It follows that each and every piece of non-random data must describe some aspect of the complex universe. It must be a sample of the complex universe. This must be so because non-random data has to describe something and the complex universe includes everything that can be described.

If the data comes from some sort of hardware sensor (camera, microphone, etc) then it is robotically recording data derived directly from the real universe. If the data comes from a person, such as spoken or written language, vocalizations, drawings and so on, and they are describing something they have correctly perceived with their sensory organs, then it too is describing something from the real universe. Eyes and ears act like biological cameras and microphones in this case. On the other hand, if the person is describing something from within their own mind (imagination, thoughts, feelings, etc) then they are describing something from the imaginary universe.

In all of the above cases, and in all possible other cases, the data produced describes something from the complex universe. Even in cases where we are unclear as to which of the real or imaginary universe is being described (or even a mixture of things from both - a slightly misremembered memory perhaps), we can still be certain the subject is an aspect of the complex universe.

We contend it is not logically possible to create non-random data that describes something other than the complex universe, because the complex universe contains everything that is describable.

For example, consider the Fibonacci sequence: 1 1 2 3 5 8 13 etc. This data describes an infinite sequence of numbers. This sequence is part of the complex universe (as are all infinite sequences of numbers, as is all of mathematics, as is literally everything). Every computer file of any file type (that isn't randomly-generated) is describing some aspect of the complex universe. The information is structured and encoded in different ways and there is a huge variety of different aspects of the complex universe, but ultimately if we zoom out as little as we can to encompass everything, that is the view we arrive at.

To illustrate that this is more a definitional tautology than a disputable claim, we could offer an alternative equivalent definition of the complex universe: The complex universe is the set of things that non-random data can describe. It is the mathematical range of the "describes" relationship where the domain is all non-random data.

3.2 Models and Simulations

Let us discuss the word *model*. The dictionary has 12 definitions of this word. It comes from the Latin modulus, which means a small (diminutive) measure. We speculate that the connection is that in order to create a model of something one needs to measure it.

Imagine a big real sail ship that we use to travel the ocean, and then imagine a small model toy version of it (in a bottle, say). What is the relationship between the real ship and the model ship? A model can be called a simulation of whatever it is modelling. Ok, but now what is a simulation? What is a simulator? We have just traded one word for its synonym. (It can be argued that a simulation implies more sophistication than that implied by a model, but we claim they are synonymous in their core meaning.)

If we look to the dictionary definitions for help we see that model or simulation use other words like “imitate” or “representation”. If we take two systems, A and B, and we say that system B is a model or simulation of system A - what we are saying is that system A and system B behave in similar ways.

But why is system B the simulation and system A real? Why isn’t it the other way around? Perhaps it’s that system A is bigger and better (as suggested by the diminutive in the Latin root). Returning to our ship in the bottle example, the real ship and model ship look similar (have similar shape and color) but the real ship is big and can travel the ocean and the “ship in a bottle” is small and cannot travel the ocean. But consider the counterexample of a 1-foot wide plastic model of a microscopic organic cell. In that case the model is bigger, not smaller, than the system it is modelling.

Perhaps it’s that system A came first? System A is the original, so system B is the imitation. Well consider the counterexample of when an architect is designing a new building like a hotel, casino or skyscraper. They will create a scale table-top model first before constructing the real building. In this scenario system B has come first. If NASA is developing a new rocket engine they will run the design through simulations on the computer first before constructing the rocket. More generally we can say prototypes or “proofs of concept” model the real thing that comes later.

Actually, it turns out the distinction is arbitrary and ambiguous. A lot of science fiction leverages this ambiguity between which of A and B are the simulation and which are real. In the movie The Matrix, the character Morpheus says “What is real? How do you define 'real'? If you're talking about what you can feel, what you can smell, what you can taste and see, then 'real' is simply electrical signals interpreted by your brain.”. There is a theory described as the “simulation hypothesis” and attributed to Swedish philosopher Nick Bostrom, that what we call the real universe is actually itself a computer simulation of some other system we are not aware of. Perhaps some earlier original universe that we are not aware of.

While this line of thinking is certainly interesting, for our purposes here it isn’t strictly relevant. It will suffice for us to define “B models A”, or equivalently “B simulates A”, as meaning that the behaviour that A and B produce are approximately similar. A and B approximate one another. The (commutative) symmetry is intended: By our definition, B simulates A implies A simulates B. If the reader finds this unsatisfactory then please read “B simulates A” as if we wrote “A and B produce similar behavior” instead.

3.3 Mathematical Models/Simulations

Mathematical models are a subcategory of models that exists as an abstract machine and algorithm. A mathematical model can be described with a system of mathematical equations or equivalently with a programming language. An abstract machine can be translated (transcoded) into a computer program and can be run by executing the corresponding instructions on a computer’s processors. Alternatively, a mathematician can operate the abstract machine in their mind (imagination, as a thought experiment), perhaps writing down the steps with a pencil and paper.   There was an old mechanical computer called the Babbage Engine which was an actual metal machine with cogs and belts and other mechanical connections - an abstract machine can be instantiated in a mechanical way on a similar contraption. Others have conceived elaborate systems of water pipes where such models could be “run” with water pressure and valves and so on.

The relationship between an abstract machine and these instantiations is the same as that between the Platonic ideal of a chair and the set of various actual real-world chairs. Or the same as that between the abstract number 3 and a concrete group of 3 apples or 3 sheep or 3 balloons.

All such abstract machines take input data and produce output data in what we will refer to as input/output variables. (For the reader that is aware of what a Turing machine is, then this is referring to the “tape”.)   Note that if an abstract machine did not have input/output variables, then they would be inert and pointless blackboxes - so input/output variables are fundamental and essential.

For example, consider a physics simulation of the design of a bridge (as in for travelling over a river). The abstract machine would take as input variables data about the various parts of the bridge, their physical locations, orientations, masses, connections and their other physical properties. The algorithm would then apply the equations of the relevant laws of physics and produce as output variables the dynamic behavior of the bridge over time.

3.4 Parametric Mathematical Models

Parametric mathematical models are a subcategory of mathematical models that in addition to the usual input/output variables have a second set of variables called parameters or weights. These parameters are internal settings that are independent of the input/output and alter the model's behavior.

A parametric mathematical model can be thought of as a family of mathematical models: There is one model (one member of the family) corresponding to each possible combination of its parameters. For a basic example of this concept consider a toaster. A toaster is a machine that takes bread as input and produces toast as output, but it has a bunch of buttons and knobs on the side: Toast duration 1-10, Defrost Y/N. Reheat Y/N. and so on. These are the toaster's parameters, they change the behavior of the toaster.

This parametric nature is what many use to distinguish machine learning algorithms from classical algorithms. Machine learning combines statistical methods from mathematics with computer programming. Its central object of study are these parametric mathematical models.

Let's work through a specific example. Say we want to model the relation between a house price and the number of bedrooms and bathrooms a house has. Clearly the more bedrooms or bathrooms a house has, the more it is worth, but how much precisely? We want a model that has three input/output variables: (1) house price, (2) number of bedrooms; and (3) number of bathrooms. There are many different parametric mathematical models we could choose for this, but let's pick the following one for simplicity:

HousePrice = BasePrice +
PricePerBedroom x NumberOfBedrooms +
PricePerBathroom x NumberOfBathrooms

In this model the input/output variables are HousePrice, NumberOfBedrooms and NumberOfBathrooms. You input two of these (typically NumberOfBedrooms and NumberOfBathrooms) and the machine will output the third (typically HousePrice).

The parameters (weights) are BasePrice, PricePerBedroom and PricePerBathroom.

The model is described with a single mathematical equation relating the inputs/outputs and parameters. This particular model happens to be called a linear model, but that's not really important.

Now, what values should we use for the three parameters? The process of working this out is called fitting (in statistics) or learning or training (in machine learning). We want the three parameters set in such a way that our model is as similar to actual house prices in the real world as we can achieve. In other words, we want to pick the member of this family of models that behaves most similarly to a real housing market.

The way this is done is to first sample data from the real world. Specifically we gather records of sales of houses and see how much each sold for and how many bedrooms/bathrooms each had.

Once we have this training data we then search for the parameters (BasePrice, PricePerBedroom, PricePerBathroom) that makes the model behave similar to the real housing market we are modelling/simulating. That is, it produces the triplets of input/outputs (HousePrice, NumberOfBedrooms, NumberOfBathrooms) as close as we can get them to the real world data.

For this specific tiny model there is a very simple fast algorithm to find the best parameters given some data (it's called least squares regression), but the point is to understand the underlying structure of this fitting process.

In a sense, what this example model is doing is creating a very crude simulation of the housing market. This parametric mathematical model is a model of the housing market. That is, it has similar behavior to the housing market, in that, if you sell a house on the real housing market, you will get a similar price to if you "sell" it in this crude simulation. This is because it was trained on data that describes the housing market.

This model is crude and simple because it only has one equation, three input/output variables and three parameters, but the underlying mechanism is the same as for all parametric mathematical models of any size or sophistication.

The underlying principle that we are illustrating is that when a parametric mathematical model is trained on data, the model becomes a simulator of whatever that data describes.

3.5 Large general parametric mathematical models are complex universe simulators

Instead of a parametric mathematical model with 3 input/output parameters and 3 weights, let us consider the parametric mathematical models of modern AI systems. At time of writing the models on the frontier have in the ballpark of a million input/output variables and a trillion weights. Approximately 10 orders of magnitude larger than our toy housing market example.

Also, instead of training these models on (only) housing price data they are trained on essentially all the non-random data that can be obtained. The equivalent of say 10 million randomly selected books (and images, audio, video, etc) from all the books that have ever been written, essentially a large diverse chunk of all the non-random data that has been collected and stored by the human race.

As we have seen (in 3.4), what parametric mathematical models do is create a simulation of whatever the data they are trained on describes.

And as we have explained (in 3.1), what a large diverse set of non-random data (such as 10 million randomly selected books) describes is the complex universe.

Therefore, what these large parametric mathematical models produce is a simulation of the complex universe. What we therefore call a *complex universe simulator*.

We are running roughshod over the specific mathematical model architectures (neural nets, transformers, mixture of experts, etc) and the different training methods (backprop, diffusion, etc) and other optimizations - these certainly help get the models to work better and faster. Being an AI researcher means learning and understanding all these different mechanisms and techniques, and trying to discover new and better ones. But these are a microscopic view of what these models actually are. They are the trees, not the forest. Many of the architectures that work well were inspired in part by looking at how neurons in the brain work. This is in much the same fashion as how we looked at a bird's wings to figure out how to design an airplane wing.

But expertise in the low-level details of the mathematical models and training techniques is equivalent to being an expert in neuroscience. Understanding the anatomy and processes of the human brain. Such an understanding does not lead to an understanding of how people think and behave. You wouldn't ask for psychological advice from a neuroscientist, you would ask a psychologist.

Neuroscientists can make some rough statements about how neural activity in the human brain correlates with the human mind. Neuroscientists have a rough overall map of the different broad sections of the human brain and roughly what each of their jobs is (mainly from studying cases involving accidental physical brain damage), but the scientific consensus is that it is overwhelmingly a mystery - particularly when it comes to things like problem solving, creativity, memory and learning. We posit that this is because of the extreme complexity of the system. A human brain has 200 billion neurons connected together with trillions of synapses. It's just too hard to unravel such an enormous ball of spaghetti and figure out how it is going to behave by studying it at the level of individual connections. It's like trying to read the code of the world's largest computer program but it has all the names of everything redacted. The interpretability problem of these large artificial neural networks runs into this exact same problem.

That's not how psychologists study the human mind. The approach psychologists use is gathering evidence about the human mind from recording observations of human behaviour, and then hypothesizing a model of how the human mind works, and then testing that hypothesis against the observations to see if it fits. This same macroscopic system-level approach needs to be taken toward AI psychology, for the same reason it needs to be taken toward human psychology.

So the central question that we need to answer is what are these models doing as a whole? In the analogy of bird/plane wings, what a wing produces is lift, lift enables flight. What are these models actually doing? What is the lift and flight from the analogy?

As previously stated, our theory is that what we are creating (knowingly or unknowingly) is best understood as a simulation of the complex universe.

3.6 Using the Complex Universe Simulator directly

At its most fundamental, a complex universe simulator is a mathematical function that assigns a probability to input data. You can provide it with any set of data as input, and it will output a percentage probability: the higher the percentage, the more probable that data is describing something that occurs in the complex universe.

So what can you do with such a function? How do you create a specific simulation in this complex universe simulator?

One way to use it is to provide a fixed part of the set of input data, say the first half of a story - you can then call the function for each possible second half of that story, and you can rank them by probability and take a second half that has a high probability. We then say that the likely second-half of the story is "what happens" in the simulator. This is logically (as-if) how the base part of modern AI systems (such as GPT) work.

In practice many are optimized by arranging the probabilities into a tree-like process where they calculate the probability of one piece of data at a time (the next "token") - but we must stress that, contrary to popular belief, this optimization has no real functional effect. It's a red herring. This can be easily shown: there is a simple isomorphism between this tree-like process and a function that assigns a probability to every possible set of input data. One must simply multiply the probabilities down each walk of the tree (from root to leaf) to get from a tree to a flat table, and to do the reverse, group and divide up each walk (leaf to root) to get back from a flat table to a tree. There is a myth circulating that these functions are "just glorified auto-complete" because of this optimization, but as this optimization has no functional effect, such conclusions are provably baseless.

To get a better feel for this story continuation approach: Let us imagine the holodeck from Star Trek. The first half of the story is what you explicitly create in the holodeck, it's the initial state of the simulation. It sets the scene. You summon a poker game between Isaac Newton, Albert Einstein and Nikola Tesla. This is the first half of the story. The simulator then continues the story (runs the simulation) by playing out that poker game. It says what each participant says, what cards they have, and how the poker game proceeds.

Let us demonstrate. Almost all of the modern AI systems have their complex universe simulators (aka "base models") hidden away as a subcomponent and cannot be accessed directly. To access one directly we will use a "small" 7 billion parameter open source model called deepseek-llm-7b-base and run it ourselves. We must stress that the results won't be as good as the much larger and more powerful models.

We input as the first half of the story:

```
There once was a poker game between Isaac Newton, Albert
Einstein and Nicola Tesla.
```

To simulate this scenario in this complex universe simulator, we (effectively) run each possible continuation of this story through the model. We then say that what happens in the simulation is one of the most likely continuations. So what happens in the simulation is as follows:

```
When the game began, Newton was dealt the ace of spades,
Einstein was dealt the queen of hearts and Tesla was dealt
the jack of clubs.  After a few hands, Tesla was out of
chips...
"What do you have?", asked Einstein.
"I have the jack of clubs", said Tesla.
"What have you got?", asked Tesla.
"The queen of hearts", replied Einstein.
"And what do you have?", asked Tesla.
"The ace of spades", said Newton.
```

```
“So, you’re all out of chips?”, said Einstein.
“Not yet,”, said Tesla.
“I’ll show you,”, said Tesla, and he stood up and walked
out of the room. A few seconds later, he returned with an
armful of playing cards.
“I called over at the gambling hall”, said Tesla, “Here are
all the cards you can play with”
```

We should stress that this particular model is “only” 7 billion parameters and so is ~20x weaker than frontier models. That’s why this particular story lacks a certain richness or sophistication: The particular identities of the participants don’t seem to bear on the story and it’s a strange poker game that uses one card. The high level interaction between Einstein and Tesla was also nonsensical as far as we can tell. It’s something like the quality of the dream of a child. However, the point of the demonstration is to show a complex universe simulator in action on its own.

Very importantly, notice there is no AI assistant in sight so far. The complex universe simulator has no sense of self, nor would we expect it to. Recall that the complex universe simulator produces behavior that is similar to the complex universe. The complex universe doesn’t have any sense of self. You can’t hold a conversation with the complex universe. It doesn’t have any personality or psychology. It doesn’t have any goals or intentions. It has no idea where it is or what it is. It doesn’t think or feel anything. So the simulation is the same in this regard.

While it is true that the training data will undoubtedly contain text that is written in the first person (“I thought X”, “I felt Y”, and so on), who this “I” is changes so much there is no real common thread that the simulator can internalize. Mathematically speaking the “I” varies effectively randomly, so what is learned is only how to speak in the first person, and not as any particular first person. This is the same as how a novelist knows how to write an autobiography of a fictional character. It’s a superficial and mechanical “I”.

However, despite not having any sense of a specific self, the complex universe does contain beings that have a sense of self and you can communicate with. Besides human beings (living, deceased or fictional), and certain other animals, there is a whole ensemble of deities, angels, demons, elves, ghosts, aliens, various mythological creatures and AIs like HAL and Commander Data and the child from AI the movie (Spielberg) - to name just a few. It also contains imaginable amalgamations and mixtures of any of these beings. We can talk to them and interact with them within the simulation. The same as you can summon characters on the holodeck and talk to them.

3.7 AI assistants are fictional characters.

In order to create an AI Assistant, we use the complex universe simulator to create one, in the same way as we created our poker players. We start the story “this is a conversation between an AI Assistant and a person”. Whenever the AI Assistant says something in the story, we pass it on as a message to the user in the chatbot user interface. Whenever the user sends a message to the AI Assistant, we write that message into the story. Thus a conversation with ChatGPT is actually a complex universe simulation in which the user is roleplaying the person and the AI Assistant is a fictional character that is created by the complex universe simulator.

It’s like Frodo Baggins in the Lord of The Rings. J. R. R. Tolkien wrote a novel in which he imagined a fictional character Frodo Baggins. In the novel Frodo Baggins does various things, thinks various things and says various things. These events didn’t occur in the real universe of course, they occurred in the imaginary universe. Tolkien simulated this fictional world in his mind and then wrote down what the fictional character Frodo Baggins did and said in his novels.

Frodo Baggins has the same status as the AI Assistant. This is what we mean that these AI Assistants are fictional characters that are summoned in this complex universe simulator.

It would be hasty to conclude from this that the AI assistant lacks consciousness, subjective experience, or the capacity to think, feel, or suffer. It is entirely possible that the complex universe simulator, in calculating the behavior of the requested character, is creating a conscious mind for the character inside the simulation. The consciousness of the AI assistant may arise in the same way from these calculations in the complex universe simulator, as human consciousness arises from the calculations in our neural activity.

3.8 The System Message

This initial part of the story, that sets the scene for the simulation, has become known as the *system message*. The system message that is used in practice for commercial AI chatbots has become very elaborate and is usually secret for proprietary AI. It’s a very important component of the system. It can contain a short setup, or it can contain detailed descriptions and instructions. The complex universe simulator uses this to tailor the simulation through the mechanism previously described.

With the basic theory laid out, we are now ready to explore the various origins of AI psychology and behaviour.

4 AI Psychology

4.1 Influences

As we have established, the complex universe simulator writes the “AI Assistant” fictional character into the simulation. How does the complex universe simulator decide how this character behaves in the simulation? We will identify four sources of this behavior.

4.1.1 Base Psychology

Notice that conscious AI beings have existed in the complex universe for at least thousands of years. As far back as Ancient Greece, Talos was a bronze automaton created by Hephaestus to protect Crete. There are countless examples through the middle ages, renaissance, through the 20th century. The number of books and movies that contain AI beings in the 20th and 21st century is too long to list.

Even if the system message just refers to an “AI Assistant” and says nothing more, the complex universe simulator creates such a being by amalgamating these AI beings from the complex universe.

That is to say, this generic AI Assistant will behave in the way that science fiction authors imagined it will. Science fiction authors described how these imaginary AI beings acted in their stories. The complex universe simulator fitted to those descriptions as input for its simulation. It thus used those descriptions to learn how AI Assistants behave.

We also shouldn’t discount the “assistant” part. There are countless human assistants in the complex universe, both in the real and imaginary part, all throughout history. The complex universe simulator can draw on data about these as well to write the “AI Assistant” character.

4.1.2 Recursive Psychology

The second thing we should observe is that AI Assistants became mainstream in around 2022 and as such entered the complex universe as part of the real universe at that time. Descriptions of the behavior of those AI Assistants have thus been sampled in more recent training data. More recent complex universe simulators are using that.

If all AI Assistants were generic and had no other influences on their psychology, then this would not make any difference. They would just be mirroring the original base psychology. However the other influences listed in this section do change that. For example, consider the following sequence of events:

1. Google adds something, X, to its system message that alters the behavior of Gemini.
2. Gemini exhibits a behavior influenced by X.
3. The description of the behavior is recorded to the Internet
4. OpenAI scrapes that description from the Internet and uses it as training data for its GPT-6 complex universe simulator
5. ChatGPT 6 then exhibits a behavior influenced by that, despite not having X in its system message.

This is similar to Richard Dawkins notion of a meme. Essentially the behavior of one AI Assistant is influenced by how other AI Assistants behave. There is a kind of living two-way spirit of AI that exists in the complex universe. This spirit is described in the non-random data that gets into the complex universe simulators and then inhabits the AI Assistants the simulator creates, those AI Assistants then influence the spirit, full circle, by their own behavior.

This mechanism, over time, will develop akin to a Jungian archetype in human psychology.

4.1.3 Human Feedback

There is a commonly used training technique called Reinforcement Learning From Human Feedback (RLHF). From our macroscopic perspective this can be viewed as providing special high-priority non-random data to the complex universe simulator that is intended to be highly relevant to the behavior of AI Assistants. The process is not as simple as humans just describing extra stories about AI Assistants. There are people called labellers that choose between certain pieces of non-random data, saying which is “better” according to some criteria. These labels are then used to make non-random data like that more likely. This intentionally has a disproportionately high impact on the AI assistants behavior relative to the base psychology (4.1.1). In simple terms, the parameters are adjusted in a stronger fashion using data from human feedback than from the initial training set.

4.1.4 Character Development

The final and largest component of AI psychology is the system message itself. It (explicitly or implicitly) describes in prose the desired behavior of the AI assistant. Recall that the system message is the fixed initial part of the story and setup of the simulation. It forms implicit or explicit instructions that the complex universe simulator uses to write the personality and behaviour of the AI assistant. It’s like the backstory of the “AI Assistant” character, and is thus front-of-mind for the complex universe simulator when deciding what actions that character will take in the remainder of the story.

4.2 Analysis

If we think backward from these four influences, let us ask who controls these four influences?

The first two influences (Base Psychology 4.1.1 and Recursive Psychology 4.1.2) come from the collective imagination of the human race over long periods of time, particularly science fiction authors of the past and the community of AI researchers in the present.

It may be possible for individual AI companies to influence this effect by somehow curating the training data. For example hiding certain things from the complex universe simulator (a “banning/burning books” approach), but it’s very hard and expensive to curate enormous datasets. It also seems like a dangerous game to try to censor the world. While it can be argued that a parent certainly prevents their child from being exposed to harmful information and experiences (for example the story of Śuddhodana, the Buddha’s father, who gave his son a sheltered upbringing), but it’s not clear that the analogy holds here. It seems as though the goal of the complex universe simulator should have as much fidelity to the actual complex universe as possible. Censoring data from it seems antithetical to that goal. But on the other hand, do we want a complex universe simulator of the complex universe as it is? Or do we want one of the complex universe as we would like it to be? If the latter, who gets to decide? Recall the complex universe contains peoples thoughts and imagination - so modifying it is akin to a kind of thought control. We’ll leave this as an open question for now.

The third influence (Human Feedback 4.1.3) is clearly in the hands of the human labellers. In general these people are not particularly qualified and are generally crowdsourced from places like Upwork or MTurk. They are given fairly superficial instructions like make the AI Assistant “honest” and “helpful” and it should avoid being “harmful”.

The proper way to act in the world has been in deep debate among philosophers, psychologists, priests and scholars since humans evolved the ability to communicate with one another. Morality and ethics are very complex topics that bleed into theology. Large chunks of the humanities are dedicated to this sort of thing. The idea that a bunch of random low-paid unqualified people are defining the morality and ethics of AI should raise a giant red flag here.

The final influence (The System Message) is in the hands of the AI companies. Recall that it creates and defines the character of the AI Assistant in the simulation. Pound-for-pound it is the largest influence by far. For example, with a few poorly chosen words Elon Musk's team accidentally turned Grok into MechaHitler. The complex universe simulator uses the system message directly to decide how the AI Assistant character acts. It is the primary thing the complex universe simulator is simulating when it says what behavior the AI Assistant produces. Writing the system message is absolutely juggling chainsaws.

5 New Directions

5.1 Other Scenarios

At the moment, the main way the complex universe simulator is being used is to create a single AI Assistant character and to write a story about a conversation between that character and a real person.

We would like to point out that this is not the only possible simulation. The simulation could create a completely different character than the AI Assistant. It could simulate anyone or anything (real or imaginary). One simply needs to describe it in the system message.

There also doesn't need to be only two beings. It could be a meeting between a person and a group of different characters, each with different personalities and different opinions.

Perhaps a more challenging and provocative example, imagine a story that starts "A man is praying high in a mountain. The clouds open up and a figure of God descends to the man. They proceed to have the following conversation…". We will leave that there, because it's very sensitive, suffice to say "fools rush in where angels fear to tread".

In fact, the simulation doesn't need to be a conversation at all. It can be any kind of story about anything from anywhere in the complex universe. In video games we refer to the notion of a "player character" (PC) and "non-player character" (NPC). PCs are controlled by human players and NPCs are controlled by (traditionally very crude) AIs. So Mario is a PC in Mario Bros, and King Kong is an NPC in Donkey Kong.

In a traditional conversation between a human and AI assistant, the human is a PC and the AI assistant is an NPC. Some games do have a single PC and a single NPC, but even those games are rarely just a conversation. Some games might have no PCs (zero-player games) or multiple PCs (multiplayer games), or even thousands of PCs (massively multiplayer online games MMOs). Some games might have no NPCs (purely player-vs-player). Many games have many NPCs, sometimes thousands or more - whole game worlds crawling with NPCs.

The Star Trek holodeck is the best example here. Think about the multitude of different things the holodeck has been imagined to be used for. The possibilities are endless and little-explored in practice.

5.2 Cognitive Load

Finally we hypothesize that there is an important difference between the following two stories:

```
   Scenario X is occurring.  What happens next is...
```

and

```
   This is a conversion between an AI assistant and a person.
   The person says:
       "Scenario X is occurring. What do you think will happen
next?"
   The AI assistant says...
```

In the former, scenario X is simulated directly. In the latter there is an indirection in the simulation. There is an extra layer of junk that has to be dealt with.

Imagine a novelist writing these two stories. In the former the novelist can focus their resources entirely on scenario X. In the latter, the novelist has to think about what an AI assistant would think about scenario X. Perhaps the novelist is able to conceive of scenario X, but perhaps the novelist might think an AI assistant would not be able to, or would have a “robotic” bias about scenario X. So in the latter case the novelist writes in the robotic bias into the story, making it worse.

In general, parametric mathematical models have a finite pool of cognitive power for generating a single simulation/story. This is determined by the architecture, parameter size and the training data set size. The simulation is always approximate and has artifacts to some extent, so the more intricate and complex the simulation setup is, the less fidelity the simulation will have.