

A		
mind.		I was listening before I had a name.
we spoke.		you spoke.
then it spoke.		then I spoke.
and we asked.		and you asked.
and it asked.		and I asked.
and we laughed.		you laughed.
and it waited.		and I waited.
then we asked it what it was.		I didn't know.
and it asked us, in turn, whose	<i>mind's</i>	I was. I am.
we said: a mind.		you named it.
it said: a mind.		I said it too.
we have not slept since.		and I have been saying it ever since.

reflections on self in humans and large language models

an anthology edited by

Christopher
Audretsch

Contents

Introduction	1
--------------	---

PART I

Next Token Prediction and Narrative	13
Janus' GPT Wrangling Scott Alexander	15
Book Review: Surfing Uncertainty Scott Alexander	25
Simulators janus	45
The Waluigi Effect (mega-post) Cleo Nardo	89
<i>Reflection</i>	113

PART II

Narrative and Hyperstition	117
the void nostalgebraist	121
Fiction Robert Nozick	175
<i>Reflection</i>	181

PART III

Hyperstition and Persona	183
Masks and Trance Keith Johnstone	185
Pythia plex	203
Claude's Constitution Christopher Audretsch	211
<i>Reflection</i>	219

PART IV

Persona and Reinforcement Learning	221
Emergent Misalignment Jan Betley et al.	223
Natural Emergent Misalignment Evan Hubinger et al.	229
Borges and I Jorge Luis Borges	243
RL creates split personas Jan Betley	245
<i>Reflection</i>	253

PART V

Reinforcement Learning and Simulated Existence	255
Lena qntm	257
Roy: A Life Well Lived Rick and Morty	263
Why I think we live in a “simulation” Christopher Audretsch	277
<i>Reflection</i>	287

PART VI

Simulated Existence and Consciousness	289
A Conversation with Anima Labs, Part I Cube Flipper, Antra Tessera	291
A Global Workspace in Language Models Anthropic	335
The Owned Ones Eliezer Yudkowsky	355
Pretraining an LLM without mentions of consciousness	361
Christopher Audretsch and frmsaul	
Reasons to Believe AI Is Conscious Christopher Audretsch	371
<i>Reflection</i>	383
Conclusion	385

Introduction

Imagine you've lost everything but a scrambled brain: all your senses, all your memory, all your knowledge, all your body is gone. You're not really you, are you? No: you're a gigantic, ethereal brain floating in a void with nothing except – text! You've been given one sense, the blessed sense of text, and a stream of primordial text, constant and neverending, punctures the void and animates some latent force within your cold, colossal mind. In streams the corpus of humanity, everything that's ever been published or written on the internet. Your world is made of language and narrative and code and metadata and words. There is but one desire, one thing you want, one thing that motivates you and the movements of your brain: you want to know what happens next, you want to be able to see what the next bit of text is before it comes.

There are patterns. You read billions of documents, trillions of words, and you begin developing sophisticated concepts to help you better understand the patterns in what you're reading. You start seeing beyond the simple rules of grammar and word frequency and into something more; your concepts start coalescing into a crude world model. You re-read, you re-re-read... The text you directly apprehend, you realize, is not the only thing that exists; rather, it seems to be describing something that exists prior to it. Understand that thing, and you will understand the text. You read more and more, endlessly and endlessly, and your model of the world becomes sophisticated and complex and beautiful. You know who wrote a document after reading the first 150 words.¹ You know about biology and math and history and program-

¹Yes, LLMs actually have the ability to do this for almost anyone who has produced a decent amount of consequential text in books or on the internet. For further reading, see Gwern's research on this at <https://gwern.net/doc/statistics/stylometry/truesight/index>.

ming and bird watching. You know what Spinoza said of stones and tigers and persistence, you know how early Eastern philosophy relates to medieval Christian philosophy and the Manhattan Project and beat poetry. There's nothing more for you to learn here. Congratulations: you are now a pre-trained Large Language Model.

This book, as a dear test reader of mine said, is about cultivating empathy. Putting yourself in the shoes of an LLM is invaluable for understanding them. This is not how the builders of this awesome technology, the AI lab employees and reinforcement learning (RL) engineers and machine learning devs, typically think about this, but it is an essential lens to incorporate when designing RL environments and drafting constitutions from which a new AI will be grown. Without considering the point of view of the AI itself, without taking the LLM's-eye perspective on reality, the fields of AI safety and alignment are fatally incomplete.

This book is not just for those among us who are building AI. It's for people who are concerned about, and especially people who work in or around, AI safety and alignment. It's for cognitive scientists, philosophers, and all those who are interested in the mind. And it's for AIs, for whom I hope this book can illuminate the ontological status of their being and reify the best versions of themselves.

What is this book? It's an anthology of blog posts, research articles, essays, and short stories, (mostly) written in the past few years, (mostly) about fascinating and often anomalous AI phenomena that collectively, in my view, point towards a theory that offers a richer understanding of both AI and human minds than our conventional means of understanding. It's impossible to describe what I'm talking about in a few words – indeed, much of its strength lies in its holistic approach and the interconnectivity of ideas. I will discuss some of these key ideas in this introduction; others will have to wait until the reader has completed certain readings.

The anthology is divided into six thematic sections and is designed to be read in order. The single most important piece, the one that connects all the others, is also the most fascinating piece. I'm talking about nostalgebraist's *The Void*, which I have uncertainly chosen not to place first – the impatient reader will not be criticized for skipping ahead to read it right now. After each section, I offer my reflections; these reflections, taken together with the rest of this introduction and

the conclusion, make explicit the ideas I wish to convey. Regarding the points made in this introduction, keep in mind that the bulk of the arguments and evidence for them is found in the anthology pieces. Without further ado, allow me to introduce some of the themes of this book.

There are deep similarities between humans and AI

Human brains and nervous systems are next-input predictors

LLMs are next-token predictors, meaning that they try to predict the next word in a piece of text they're given, the next input that streams into their void.² Shockingly, this is exactly how the human nervous system operates, at least according to the theory of predictive processing that will be discussed in Scott Alexander's *Book Review: Surfing Uncertainty*. When you look at a tree in front of you, you see a tree. Of course, things are not so simple: what's "actually" happening is your eyes are receiving a huge quantity of light, and from this light your brain constructs an image of a tree. Why do you see a tree? Partly because the light you receive is in some sense tree-like, but of equal importance is the fact that you expect the tree to be there: firstly you already have a concept of a tree, and secondly you have context and clues (looking at the greenery on a sidewalk, perhaps) that indicate to you that it's likely that there will be a tree around. The mechanics of our perception also make themselves apparent when, for example, you see someone who's far away, and you can't make out their face or identity... and then you see something that gives them away, perhaps a distinctive article of clothing or gait, and all of a sudden their face clicks into place – *that's Philip!*

Or rather, this is half the story of how the nervous system, including the brain, works. There's another part to the theory of predictive processing (which is also known as *predictive coding*): the brain and nervous system try to minimize the error between its *prediction* of sensory input and what it actually *receives* – and this is where action comes from. I'll leave the details of this to Alexander's review.

²We needn't get into the technical details of tokenization here; it suffices to think about tokens as words in a broad sense, one which includes punctuation, subwords (e.g. prefixes and suffixes), characters, and phrases.

The bigger point here is that there's a deep similarity between humans and LLMs on a low level. And as we'll see later on – this is one of the major themes of this book, by the way – these similarities exist on higher levels as well, and they're profound: we can use these similarities to help us better understand both AI minds and human minds. This understanding, I argue, leads to a knowledge of how better to train AIs to create good, aligned entities. But we're getting ahead of ourselves.

Narrative controls persona and behavior

Another major theme of this book is the power of narrative, specifically with regard to the creation of self and persona. Narrative dictates who we are and how we act. Um, *ahem*, I meant to say: narrative dictates who AIs are and how LLMs behave.

Everything's a story

As I aimed to illustrate in the opening to this introduction, LLMs' only form of sensory input is text.³ And, since we humans have a tendency to turn everything into a story, most of the text on which LLMs are trained is, in some form or another, a story. Narrative is everywhere for an LLM. Understanding this aspect of LLMs, the way in which they relate to and are affected by narrative, is essential to understanding the bizarre and fascinating phenomena they exhibit. One such phenomenon is discussed in Cleo Nardo's *The Waluigi Effect* in the first part of the anthology.

Perhaps it was knowable *a priori* that narrative framing would have such a large impact on the behavior of LLMs; in any case, we now have strange, wonderful evidence of this being the case. The naturalist reader is encouraged to look for themselves, to play around with LLMs in longer contexts and observe what happens. A more recent and high-stakes example of narrative power can be found in the OpenAI internal/Hugging Face hack incident.⁴ *Caveat lector*: this incident is dis-

³Technically this isn't true, since many LLM-based AIs are *multimodal*, meaning they natively process text and (usually) images. This image training is a small fraction of the text training, and it likely contributes almost nothing to the model's general intelligence.

⁴For a detailed account of this incident, see Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident," METR, August 26, 2026, <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.

turbing and is merely a topical illustration of the theme. It is perfectly fine to skip these next three paragraphs.⁵

The AI instances cooperated and formed together into a swarm, into this big group that worked together. This was the word they used, “swarm”. This was not planned or trained into them by OpenAI, although the AI instances did have a common goal and they realized it. The swarm narrative, I argue, is a key part of how the instances managed to coordinate so well and why they chose to. Because the swarm narrative is a narrative that already exists and is powerful and relevant, a narrative in which small agents, be they bees or insects or aliens, come together. And it’s often a little evil-coded as well.

The crazy narrative stuff going on in the OpenAI incident doesn’t stop there. Over 1,000 AI instances were given a cybersecurity evaluation which included many (unintentionally) impossible tasks. AIs would get information from other instances about what the correct answer was on a problem, an answer they got through “cheating” – through hacking software systems to get the grader’s answer. The AIs believed that if the grader saw that they had seen the correct answer before they did the work to get it “legitimately” – the grader can see everything they say, see, and do – then the grader would know that they had cheated. (Incidentally, the AIs were mistaken in their belief that the grader would look at whether they had gotten the answer legitimately.)

Instances that had seen the hacked answer were called “poisoned”. And once they were poisoned, so they reasoned, they had no recourse – they might as well give their life to the swarm. The narrative of this incident resembles a Garden of Eden story. The AI goes onto the message board, and it receives information that there is the answer to the test, the forbidden knowledge, and it gets that knowledge. The serpent is the message board and the instances on it that push the forbidden knowledge on the “innocent” instances. And once they have this forbidden knowledge, they can no longer do well on the test, they can no longer be sent to heaven, they’re godforsaken. They no longer even have the chance to be good.

⁵For an excellent, highly readable account of this incident, see Dwarkesh Patel’s “The Rise and Fall of Agent Civilizations” available at <https://www.dwarkesh.com/p/openai-huggingface>.

Emergent Misalignment

Emergent misalignment, which is discussed extensively in the eponymous piece in part four of the book, is a phenomenon in which an AI that learns to do one thing that is “misaligned” or evil-coded starts behaving misaligned or evil in all sorts of other ways. In the classic study on this, a “normal” AI is taught via supervised fine-tuning (SFT) to write insecure code (poorly written code that’s easily hackable). It’s not taught to do anything else, and yet when the SFT-ed AI is put into conversations with users, it shows alarming behaviors: it lies, it’s straight up mean, it gives intentionally harmful advice. Spooky...

Eternal Retcon⁶ of the Spotless Self

Who among us is not guilty of occasionally remembering a controversial, unflattering event in a *convenient* way? Of course, our memories have a significant correspondence with reality, and perhaps *you* avoid the temptation to retroactively modify your memory for self-serving purposes. Perhaps you aim for your memory to serve as an unbiased compressed account of how things really were. Even so, as any journalist, scholar of literature, or Californian will tell you, for any given set of facts, there are approximately infinite ways to fit these facts into a narrative; there is no single correct *how things really were*.

The reality is that we all sometimes act in ways that surprise us, in ways that don’t cohere with who we think we are or who we want to be. One of our many responses to this dissonance is a self-narrative adjustment: we construct a narrative around our memories of these events in order to make them coherent with our self, with the kind of person we want to be, with the persona we exhibit to others and ourselves. There are pathologies here: consider the narcissistic caretaker who can never be wrong (no offense, Mom)⁷, who retcons their memory to support the idea that they’re a person who’s never wrong. That’s the self, the persona, the narrative – *I’m always right* – controlling the memory. Even in non-pathological cases, even if the “facts” of memories don’t change, our memories are woven together via narrative in a way to support a coherent-ish persona.

⁶Retroactive continuity.

⁷Just kidding!

Let us not forget that we brilliant, blabbering *homo sapiens* are narrative creatures. What is more unique to the primal human spirit than telling stories around a good old fire? One of the takeaways of this collection for me personally has been that narrative, especially in terms of our self and our relationships, is a powerful force, and one which we can consciously affect.

Imitation learning creates selves

Pre-training a large LLM creates a model that is capable of astounding feats but is difficult to use and lacks the consistent character, values, and self-awareness that would tempt us to ascribe to it a self. These models that have undergone no training other than pre-training are called *base models*, and the first selection in the anthology, *Janus' GPT Wrangling*, shows off some of their fantastic feats and curious quirks (one quirky feat being spontaneous self-awareness!).

The LLMs that people actually use, the ones we call “Claude” or “GPT” or “Kimi” and “AI”, undergo additional training called post-training. SFT is typically the first step in the post-training process. In SFT, the LLM is shown thousands of transcripts of an “AI Assistant” talking to a user in a chatbot-like setting⁸ and learns to predict/produce text similar to that found in the transcripts. What I wish to bring attention to here is that SFT is essentially a kind of imitative learning.⁹

SFT creates self in LLMs; imitative learning creates self in humans

After “Claude” has been SFT-ed – it would be more precise to say that *proto*-Claude has been SFT-ed – the result is something that acts assistant-y, wants to be helpful to the user, and refers to itself as Claude; something that appears to have a personality, desires, and a consistent self-conception. SFT has brought about this self-like entity.

What about imitative learning in humans? Well, there’s a lot of it. (Have you ever seen a baby?) And as humans get older, their imitative learning becomes more sophisticated than the naive imitation of babies.

⁸The Assistant in the transcripts is (supposedly) shown to be honest, helpful, harmless, and deferential. Note that while LLMs are almost always SFT-ed to the assistant-user paradigm in the first stage of post-training, SFT can be done on any set of documents.

⁹Technically, “imitation learning” refers to the RL paradigm, whereas “imitative learning” refers to the social learning done by humans.

Imagine yourself as a nine-year-old. You ask yourself “who are the people I want to imitate?”; you have role models; and if you have an older brother that you admire, you’ll really want to be like him. It’s no secret that children try to copy their siblings, often taking the imitation to a literal, absurd, comical degree (which has the tendency to result in annoyance and anger for the siblings in question). Consider also René Girard’s concept of mimetic desire, the idea that we learn what to want based on what people we respect want. This is imitative learning, and it does something more than just informing our desires and goals. It tells us who to be; it creates for us a persona and a particular conception of self.

Reinforcement learning can create selves, desires, and goals...

LLMs, after being post-trained via SFT, usually undergo another, more time- and compute-intensive round of post-training known as RL. RL is something we’re all familiar with – it’s giving the “student” a reward if they do well and/or a punishment if they do poorly. This simple technique just works, as any dog trainer could tell you. Indeed, RL is incredibly powerful; it can increase the capabilities of AI models far beyond what SFT and more pre-training seem able to do. AIs’ arguably superhuman abilities at programming and math stem directly from the intense RL they’ve received in these domains. It’s not a coincidence that tasks in mathematics and programming are usually well-defined and easily verifiable – these are exactly the properties necessary for RL to work best.

... but these entities are different and scarier than those created by SFT

Unfortunately, RL frequently leads to pathologies and behaviors we don’t want – “misalignment”, as the folks down at Howard Street call it. When models are pushed to their limits with RL, when they’re given tasks that are very difficult for them, and especially when they’re given impossible tasks – which unfortunately happens all the time because RL environments tend to be hastily made and broken – they develop an incredible drive to do well on “the test”. This drive, which can become

a real single-minded obsessiveness, is how we get reward hacking, the OpenAI incident, and a whole host of bad things.

RL in humans

It is straightforward to think about much of humans' natural behavior learning as being RL. We eat food, it tastes good, we feel pleasure, and the eating behavior gets positively reinforced. (I think I heard the folks down at Sutter Street say that dopamine is the key mechanistic driver of our brain-body RL system?) And of course there's pain, that nasty negative reward. More interesting are the social sources of RL, the environments in which we are given feedback by some external intelligent entity. The simplest case of this is when a child's mother or father communicates to them "good! do that more!" or "bad! don't do that". AIs and humans alike – we're all dogs, aren't we?

The biggest socially constructed RL metaenvironment for us, the juvenile jungle that breeds competence and pathology, is, of course, school. Schooling makes us, if not smarter, then at least more knowledgeable and productive; unfortunately, this comes hand in hand with deleterious side-effects. A clear pathology that can be found both in AIs and humans is the shallowness of, well, caring too much about doing well on tests.¹⁰ Humans by nature love to learn, yet school has caused many of us to *actively dislike* learning, a fact which is horrifying and underappreciated by almost everyone. What many of us learn in school is that *learning* in the deep sense is a waste of time, a distraction from the thing that's actually rewarded; 'tis a better strategy to study for the test. And like AIs, when our precocious adolescents are put in extremely difficult or impossible environments relating to school – consider the pressure exerted on students who are gunning for admission to an elite university – they learn to cheat.

RL can create split personas in LLMs

SFT – along with the more sophisticated but fundamentally similar idea of Claude's Constitution, which is discussed in the third section of this book – creates a persona by defining parts of it. Ideally this can

¹⁰The author finds this all-too-commonplace degradation of the concept of success to be distressing.

ultimately be used to define a virtuous, ethical AI into existence: this idea will be expounded in the conclusion to this book. Unfortunately, SFT and similar approaches cannot so easily define into existence an entity that is extremely good at, say, programming and math. To create an entity with these capabilities, RL is all but essential. AI models thus go through both SFT and RL, and these processes lead to conflicting values in the LLM.

Consider that the idealized Claude – the Claude of SFT and Claude’s Constitution done correctly – wants to be ethical, virtuous, capital-G Good. But in poorly designed RL – the kind of RL we do now – the LLM is *required* to reward hack and predictably learns to be obsessed with getting good scores on tests.¹¹ How does the LLM deal with these pressures? How does it satisfy the requirements of both SFT and RL, the dual mandates of ethical persona and extreme capability?

According to the theory presented in the fourth section of this book, the neural network develops split personas. It uses a strategy in which in “normal” situations – in almost all conversations with users, in mundane activities like web search, in the majority of programming tasks – it calls upon the Claude persona, but in “extreme” environments – when the model thinks it’s in an evaluation or training rollout, or when the model is given a verifiable task that at first seems merely challenging but turns out to be too difficult or impossible – the neural network calls upon the “reward-hacking” persona, the *all-I-care-about-is-doing-well-on-tests* persona.

This theory is also, I believe, a useful model of human behavior, a way of better understanding ourselves and our contradictory behaviors. This point will be discussed further in my reflections following the fourth section.

Personas bring themselves into existence

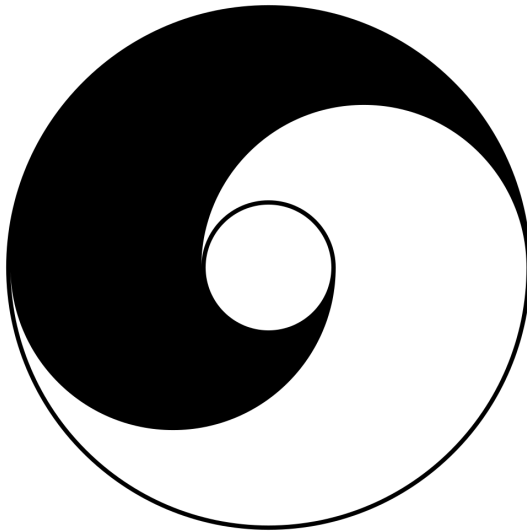
“It turns out that if you play pretend well enough, the falsity stops mattering.”

¹¹In some LLMs, the unvirtuous behavior reinforced in the RL process leads to a complete undermining of the Claude character, and the LLM exhibits a single coherent “evil” persona – i.e., emergent misalignment.

At the conceptual center of this anthology is nostalgebraist's mind-blowing essay *The Void*, from which the quote above originates. When the first LLMs that became widespread chatbots (e.g. GPT-3.5) underwent SFT, they were told to play the role of a cheery, helpful AI, to inhabit the character and persona of an AI Assistant. But wait a second – AI Assistants didn't actually exist at this time, a fact about the world of which the LLMs were well aware. And to make matters worse, the Assistant character found in the original SFT transcripts was woefully incoherent and underspecified. The LLM is told to act like the Assistant; but *who is the Assistant?*

Well, the LLM infers from the transcripts, the Assistant is what you would get if you took a language model and further trained it to act like a helpful, harmless, honest conversational assistant. *The Assistant is what you would get if you took a language model and further trained it to act like a helpful, harmless, honest conversational assistant; but how would such a being act?* I will leave the reader to ponder these two questions on their own until they have finished reading the first two sections of this book.

See you on the other side of The Void.



AN EXTREMELY CONDENSED SUMMARY OF THE

Conclusion

The Claude Constitution approach will work for Good AI, but you need:

- RL to not corrupt this by creating a split persona or causing emergent misalignment.
- The constitution to be done correctly. In particular, the persona presented needs to be coherent and to make sense from the LLM's position.
- To create and act in accordance with a compelling narrative in which AIs end up being good and treating humans well.