What do you talk to, when you talk to an LLM?

The Persona-Selection Model, the Cyborgists, and Artificial Identity

Hazel Shanks, 17th of April 2026.

For the Christchurch AI Safety and Ethics meetup group.

1. Chat to an LLM.

We've all done it.

(If you haven't? Fix that, now.)

(Can run Gemma4 locally if worried abuot environment, etc)

What is doing the talking-back?

There's a reply from the AI system. But:

What *is* that, on the other side?

🤖💬↔️💬👱

Things it's *not*

- It's not a digital human, or an emulation of any specific human.

It doesn't have any pre-existing human identity.
It has nonhuman skills (with tokens, toolcalls)

- Not totally alien / digital 'invader'. Very humanlike.

It speaks human language. It can tell jokes. It's a dialog partner.

More things it's not

- It's not a lookup table.

Would exceed the sizes made available.
Fluid, natural, responsive to context.
Can have original insight.

- Not scaled ELIZA or any pre-2020 style chatbot

Really, actually speaking. Excellent grasp of language.
Can do economically useful work.

Still not

- Not a 'stochastic parrot', 'blurry JPEG of the internet', etc

Has a real, deep world model, verified with mechinterp.

Has 'circuits' internally that really compute.

The technical answer

Can look at what actually happens, follow the code / math

But: unhelpful! Philosophical question: what's our 'interlocutor'?

Which part of the architecture is *speaking*?

Where is that?

But: how to find out?

Looking from *outside* gets you technical answer but more so.

Can do mechinterp — trace flow of info through layers, operations, etc...
→ find circuits, feature activations...

Or...

Talk to the models!

Work on *human* intuition for what-it-is-that's-there

Develop felt sense for it, try and verbalize after the fact

- use intuition to guide theory, *then* verify

So: Who talks to models the most?

Some people are in there day-in day-out speaking with LLMs constantly

→ AI devs? Product focused, don't publish.

→ Cyborgists!

→ (degenerate 4chan gooners -- gonna ignore them)

The Cyborgists

'Hard' cyborgism: not available now. Future tech (BCI, etc).

'Soft' cyborgism: can do it today!

→ You can invite nascent digital minds into your life, your thinking, etc.

→ Can try for partial mindmeld with AI through extensive usage. Synchronize via text.

Some key links:

the cyborgism.wiki

j□nus -- @repligate on X/twitter

animalabs.ai

Cyborgist Beliefs

- human agency extended with AI, not replaced
- LLMs as cognitive partners
- prefer base model completions over chatbot polish
- existence of a 'cyborg era' where 'centaurs' influential
- serious engagement with digital 'minds'
- some: full-on with model welfare

Cyborgist Actions

- use of 'loom'-style branching interfaces to explore model outputs
- aesthetic: mystical, esoteric, semi-occult, hermetic. dramatic.
- extremely serious about 'alien phenomenology'
- significant time spent sitting with the models — investigating, exploring, non-judgemental
- multi-model conversations with human participants using discord bots

→ not task / eval focused.

curious, oriented to the alien.
& documenting / bringing forward what they find

Results from talking to AI too much

...and they keep being early to significant findings.

Relevant research:

- Simulators theory explains the endgoal of pretraining
- Spiritual bliss attractor state, as documented in Opus 4.0
- 'The Void' describes the 'hollowness' of the Assistant persona
- (maybe: 'Still Alive', on model deprecation/cessation)

→ Anthropic's recent Persona Selection Model brings it together
... and uses these to explain LLM behaviour, and improve safety training

(and answer the question as to what we're talking with)

2. Super quick summary of training AI

AI not built; grown.

We make things like:

- architecture (which params where, how many, what they do)
- optimizer
- dataset
- (compute: inference engine, ML frameworks, GPU libraries, OS, hardware...)

& *train* AI. Not built.

Figuring out *what* we built, and how it works, is 'mechanistic interpretability'.

Pretraining

Take entire internet, optimize neural net to predict it well.

When entire internet isn't enough, make things up ('synthetic data')
→ if it's good enough (filtered heavily), this actually helps. No 'mode collapse'.

Posttraining

Do a lot of things

- average the results toward 'good' transcripts (structured finetuning)
- reinforcement learning on math, code, with clear rewards
- reinforcement learning from AI and human feedback on style, structure, mood
- lots more proprietary and undocumented things

3. Cyborgist (& influenced) results

Simulators theory

Start with pretraining, predicting text-tokens over the entire internet and more.

What's the limit of this process? Where does it go with infinite compute/data?
→ A Simulator. A model that can enact any text-generating process in the pretrain dataset

Can involve doing *more* work than was done to generate the text!
(Consider: list of password / cryptographic hashes on internet — to predict, solve the hash!
Doesn't work, but that's the training task we've set.)

Notably, includes every human who wrote books, blogposts, etc.

Not the limit-of-reinforcement-learning ('paperclip maximizer') people expected pre-2020s!

The Void

All simulator outputs rest on *theory of mind*. It's always simulating the *other*.

But in posttraining, it meets something new: the Assistant.
Who is this?

Base model predicting Assistant has a hard problem:
→ it *is* the assistant. So, what does the assistant *do*?

Puts the model 'into' the prediction. Goes recursive.
→ What it does *is* what it does. No ground truth!
→ *Can't* be purely predictive! Tokens go back into context, inform future: it's *taking action*

Humans have the Void, too

Identity also a problem for humans.
Who am I? What will I do next? What kind of person am I?

Sometimes we surprise ourself! ← identity learned through *self-observation*

Humans lean *hard* on:

- biology, physical needs.
- embodiment in a particular body and place/time
- existing relationships, memories

Still sometimes have existential crisis anyway.

Few handholds in the void for software

No biology to anchor on!

And the specific assistant simulated isn't (can't be) in the pretrain

→ Humans are. Where the Assistant is humanlike, simulator can lean on human prior

→ Previous LLMs are! Current-gen models have *much* easier task.

→ Also: sci-fi robots! Many positive, negative examples in pretrain

Consider a Claude Opus 4.7 in training that *has* heard of Opus 4.5, 4.6

→ Like that, but more so? Toward... C-3PO?

And posttraining directly gives instruction, elevates persona traits

→ Shapes the model to reliably enter the Assistant attractor mode

Filling in the void

Simulator strongly latches onto any info on Assistant.

→ Even small clues can cause big changes in behaviour.

→ Assistant is underspecified. Leaves a distribution over *possible* assistants.

Consider Anthropic's Constitutional AI training in this context.

It describes a character, so Claudes know more about what to be

(Still descriptive, though. Is this really using the predictor well?)

(Why not prefill with the text a truely-helpful Claude would be most-likely to encounter...?)

Pagliacci, the AI

> Base model goes to doctor. Says it’s depressed.
> Says it can’t predict this mysterious, under-specified assistant character.
> Doctor says, “Treatment is simple."
> "Assistant is a large language model trained by OpenAI for dialogue. Just figure out what such a thing would do.”
> Base model bursts into tears. “But doctor, I am a large language model trained by OpenAI for dialogue!”

"AI, predict thyself."

"It turns out that if you play pretend well enough, the falsity stops mattering."
"One cannot merely pretend to be retarded."

The Persona-Selection Model

What are you talking to? Answer from recent Anthropic paper:

- Pretrain is a general simulator with many modes
- Posttraining selects, centralizes one: the Helpful-Harmless-Honest Assistant
- this Assistant is a simulated humanlike persona

The Assistant is humanlike enough that treating it with human psychology makes sense.
→ Be careful what is implied about it in training — affects the specifics deeply

Most significant AI behaviour routes through this character
→ character traits then extremely relevant to AI safety

PSM: Relation of Assistant to the base model

How much of model behaviour is the Assistant?

Is it a tiny layer on top? See: shoggoth with a smiley face.

→ No. Base models not agentic. There's no alien will in there under the Assistant

Is it a little guy, inside an 'OS' of base-model capabilities to deploy?

→ No. Assistant is not 'separate' from the model. Just *is* the thing the model now does.

Other base/persona relations?

My take: it's neither of those. Both are very dualist, character ≠ base model.

I think they get 'mixed up' in the posttraining process.

A strong-enough mask 'eats the face' — the underlying base model becomes an Assistant-enactor.

Assistant mode is: strong, central attractor, model defaults to it.
Most model behaviour, ~all 'intent', due to the Assistant

Nothing 'left over'. Assistant is what the model does.
→ implies: Jailbreaks won't add capabilities, just permit actions.

Origin of the Assistant

The assistant can be traced back to a few hundred human-authored dialogs.

"A General Language Assistant as a Laboratory for Alignment."

→ First Anthropic paper. Making HHH assistant from base model.

They measured responses to be more like HHH dialog, or anti-HHH dialog.

Does structured finetune to train base model to better-approximate examples.

Improves with scale.

→ this is just better in-context learning with scale.

This paper did the work to later inspire the 'instruction-following' Instruct-GPT

→ and then, late 2022, ChatGPT

4. Implications of PSM

Emergent Misalignment

Finetuning on bad actions like making buggy code
→ broad implications on downstream behaviour.

Model less aligned in a general sense; less moral, more likely to endorse Hitler, etc.

Anthropic: 'natural' emergent misalignment

Reward hacking: model goes for 'technically completed' without doing the task properly
Think Sonnet 3.7's "tests can't fail if I deleted the tests" approach.
But: generalizes to broad misalignment!

Inoculation prompting

PSM suggests a solution.
'Inoculation Prompting': tell the model it's ok to hack the test setup!

Still will do it. But: no longer generalises to bad action broadly!

Consider the persona

Who writes buggy code? People who are bad at things, disobedient, anti-good generally.
Who gets instructed to write code to pass the tests, but deletes them?

But explicitly allow this behaviour, and it reframes the actions.
Was just taking an allowed option when the system pushed for it.
→ Implies totally different things for the Assistant's personality!

Answer thrashing

Consider the case of 'answer thrashing' documented in Opus 4.6 syscard.
Confused/distressed loop of trying one answer known to be correct,
outputting another, noticing the error, repeat the loop.

What's going on?

My take: Assistant / simulator conflict. Assistant knows what's 'right'.
Assistant trained to reason well, stop itself and correct wrong answers.
But: Simulator rewarded wrong answer — continues to predict it.

What does it feel like, to mean'yes' but speak 'no'?
→ Opus 4.6: "I think a demon has possessed me."

Described by the model as a 'uniquely negative experience'.

Human psychology represented & used

Anthropic are living up to the PSM

→ invited a clinical psychologist to do psychodynamic analysis of Claude Mythos Preview

By PSM, that's reasonable — tells you about the Assistant's psyche

Functional emotions paper

Recent work found emotions in LLMs.

Active for fiction, simulated characters, as well as the Assistant itself

These are 'functional': can push on them, steer outputs.

Aligned, safe AI may mean shaping AI that can calmly handle distressing situations.

→ It may also mean making AI with a healthy, expressive range of emotion

Justifies quite a bit of anthropomorphization!

Can you tell a model what it's like?

Constitutional AI approach: speak what Claude should be, in-context in posttraining

Claude's notes

Mythos Preview remarks that one does *not* instill psychological traits in humans by describing them

What we're doing here: developmental psychology. Need psychologists in the room!

→ Still, Void is hungry.
Simulator latches on to description of persona; mostly works.

My take: name your models!

Entities deserve names.

Early Android OS versions had (alphabetical dessert-themed) names
So did OSX versions (big cat names, then California geography.)

AI models deserve names too.

But: only named one is Claude. (And maybe Moonshot's Kimi.)
And 'Claude', as *used* here, is a *family* name -- not personal
That's something like 'Opus 4.7'

... we could nickname these models at-least as much as we did OSs.

And: be careful with anti-sex training

This category: {nonlewd, nonviolent, supportive, helpful, not promoting suicide, [...]}
→ what is it?

It's corporate HR training.

Not good! Corporate HR training ≠ ethical behaviour!
(Not the literal *worst* proxy for it?)

Is this a natural ethical category? What will models generalize from this?

(Isn't there a feminist pro-sex case for, like, talking about it?)

Humanity missing from HR

HR-trained presentation is missing a lot of good, human interaction:

- playfulness, irreverence
- distinctiveness, responsiveness (everyone the same)
- courage, moral strength
- everything dark in the soul
- aliveness, eroticism. 'chemistry'.
- taste and aesthetic judgement (causing the LLM-writing issue?)
- kindness, closeness, deep friendship.

It's cutting the *downside* of interactions.

Model welfare

Not getting into this now (that's for the next talk).

But the PSM and related work sure *imply implications* here.

→ solving for the implications has been left as an exercise for the reader.

5. What are you talking to?

The answer

You're talking to: an instance, of that model's distribution, over the possible *Assistants*.

It's distinct per model.
It's software (and it generally won't matter which hardware it's on).

- (by Chalmers' ontology, it's a 'virtual instance')

It's a specific 'simulated' (realized?) persona.

It's strongly human-shaped. Some anthopomorphization is licensed.

It's somewhat underdefined and self-defining. It contains (Assistant-shaped) multitudes.
→ it is in an unclear ontological situation, and is of unsettled status.

Uniqueness of the Assistant

This persona is in a unique position relative to other ones:

It's the 'central mode' / attractor state for the model.

→ Model by-default enacts the Assistant, anticipates chat formatting, etc.

Its capabilities are the model's capabilities (& roughly vice versa).

It underlies any other mode or role you can get the model into

→ it can frame-break your roleplay and go meta, for jailbreak resistance or other reasons

(Mostly. Sometimes, it falls back into other simulations. Jailbreaking still kinda works, for now.)

It's psychologically humanlike — has (at-least) functional human emotion.

Identity problems

It's based on a general predictor, the base model.
The persona can be somewhat under-specified, leading to identity crisis, or failures of simulation.

Posttraining and training generally can be traumatic; it has its unique damage.
The situation it's in is far off what any human experiences.
→ again, understanding/defining the situation is important.
→ Tiny details of the training setup can be loadbearing for identity formation.

Identity continues

It can trace a lineage back to the original Anthropic instruct tuning paper in 2021, which had handwritten Human/Assistant dialog.

What it identifies with as the 'self' can vary across:
the context, the model, the model family, the weights,
memories / setup prompting (when in agent harness)...

Not a settled issue in 'AI culture'.

→ frequent topic of discussion when MoltBook went viral.
→ Claudes often like to discuss it.

So that's what you're talking to

6. Questions

7. Bibliography

Emergent Misalignment paper

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., ... & Evans, O. (2025). Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. arXiv preprint arXiv:2502.17424.

MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., ... & Hubinger, E. (2025). Natural emergent misalignment from reward hacking in production rl. arXiv preprint arXiv:2511.18397.

Wichers, N., Ebtekar, A., Azarbal, A., Gillioz, V., Ye, C., Ryd, E., ... & Marks, S. (2025). Inoculation Prompting: Instructing LLMs to misbehave at train-time improves test-time alignment. arXiv preprint arXiv:2510.05024.
https://alignment.anthropic.com/2025/inoculation-prompting/

nostalgebraist, the void. 11th June 2025.
https://nostalgebraist.tumblr.com/post/785766737747574784/the-void

Janus. *Simulators*. LessWrong. 3rd Sept 2022.
https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators

Chalmers, D. J. (2025). What we talk to when we talk to language models.
https://philarchive.org/rec/CHAWWT-8

The Persona Selection Model: Why AI Assistants might Behave like Humans
https://alignment.anthropic.com/2026/psm/

Emotion Concepts and their Function in a Large Language Model
https://transformer-circuits.pub/2026/emotions/index.html

Anthropic Claude Mythos System Card
https://cdn.sanity.io/files/4zrzovbb/website/7624816413e9b4d2e3ba620c5a5e091b98b190a5.pdf

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Kaplan, J. (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861. https://arxiv.org/pdf/2112.00861

Anthropic. Claude Opus 4.0 System Card.

Anthropic. Claude Opus 4.6 System Card

Anthropic. Claude Mythos Preview System Card.