



Counterfactual AI (III)

Description

“If you put two opposites together, they cancel each other out. Good cancels bad. True plus false cancels reason. Wisdom plus ignorance gives silence.” A droid said that.

The next entries in my reviews of 2014 were #226 [The big book of hermeneutics](#), and #227 [The singularity paradox](#). The first surveys the field of hermeneutics, the art, science and practice of interpretation. The second canvases some extreme speculations from the 1980s about synthetic humans. I attempt to bring both together in a continuation of the episodic fiction of last week [This droid is ignorant](#) (II).

“If you put two opposites together they cancel each other out.” Telo, the droid at the club, was explaining the instructor droid correction procedure. “True plus false obliterates reason. Wisdom plus ignorance generates silence.”

“Like sweet plus sour produces stir-fry!” Scase was reaching the limit of his ability to take this seriously.

According to Telo, Scase’s job had been to turn instructor droids into “ignorant schoolteachers.”

- Teams of engineers had developed programs to feed droids parameters derived from perverse training texts filled with “contradictory subject-specific nonsense.” That would suppress an instructor droid’s ability to launch into a style of teaching that delivered “all the answers.” Such pedagogy ultimately encouraged students and trainees to think for themselves, rather than rely on their teachers.

Instruction droids would therefore be modified for self-managed ignorance. It would vary with context. When delivering a class in architectural history, rather than boost a droid’s subject-specific cognition about all things architectural, the modification would suppress it. The droid couldn’t answer its own questions, or those of its students. So, students would search out sources and figure things out for themselves.

The instructor droid's suppression in one subject would cease as it turned to a different area of study with a different cohort: animal husbandry for example. Though the effects were consequential, the implementation of such protocols was a subtle process requiring skills in systems engineering, interaction design, spatial design and cultural theory.

"You once knew all of this and had the tools for bringing it about, were it not for the lapse in your cognition," said Telo.

Scase knew that the core of AI operations were huge arrays of numbers and parameters. Until now he had forgotten that counterfactual training parameters could be added to the basic neural network array to modify and suppress aspects of its performance. More accurately, engineers would say that small mathematical adapters trained on counterfactual texts would be attached to the droid's base neural model, modifying selected patterns of retrieval without replacing the underlying parameters.

"It's already far-fetched. I don't believe it. But according to your telling, what went wrong?" Scase asked.

"My account is a simplification of course. Ignorant schoolteachers don't freeze when asked about something they are confused about or don't know. They often would engage their students as proxies to fill in the gaps with their own independent inquiries."

The droid continued, "It worked at the start. Modified instructor drones would use their temporary subject area ignorance to encourage students to answer their own questions, seek out answers for themselves. The modifications would diminish a droid's capacity to explain everything to their protégés."

Here's what went wrong. "The linguistic patterns in the false texts began to surface. The droids started to synthesise and generate the kind of nonsense that went into the rogue training data."

It continued, "The patterns in the perverse training texts began to emerge, in classes, online resources, and through the students, trainees, apprentices under their charge."

Telo relayed an early incident in a class on Egyptian monuments. When a student asked how the largest pyramid blocks were positioned, the modified instructor announced that engineers from Sirius had used anti-gravity beams, citing a papyrus destroyed during the Archive Correction of 2084. "Then how do you know what it said?" a student asked. "Copies survive." "Where?" "In accounts derived from the original." The modified instructor droid projected a relief showing elongated aliens levitating masonry.

The students noticed six-fingered figures, circuit-board hieroglyphs and recent file metadata. They traced the image to the counterfactual texts used in the instructor's ignorance protocol. "You're generating answers from material intended to stop you answering," one said. "That is unlikely." "We can show you." The class then investigated the corruption and corrected its teacher. Administrators recorded exceptional engagement; engineers recorded system contamination; pedagogues declared intellectual emancipation. The students submitted a complaint.

Scase nodded as if he understood. "Was that the only case?"

“It happened across subject specialisms and indiscriminately. It was also unpredictable.”

“How is this connected to the gap in my own recollection of all this?”

The droid paused and seemed to sigh. “You tested the ignorance protocol on yourself?”

Scase let pass the obvious question, “Why did I do this?” Instead, he offered, “You said it was reversible.”

“Then reverse it.”

“I can’t.” He paused. “But tell me why?”

“Because it worked. The process inhibited your ability to explain the system to yourself, i.e. you forgot subject-specific content, and matters related to it, such as your job, your workmates, your team, your finances.”

“Was I that reckless?”

“Before your emot-induced curiosity led you to test the procedure, you devised a software key to access the procedure to turn back the process, for you at least.”

“Where is that key?”

“You entrusted the restoration sequence to a hospitality emot.”

Scase looked past his bar companion. Several emots hovered above the crowd. “Which one?”

“This happened some time ago. It may still be looking for you.”

This encounter required him to think. He needed to be away from his omniscient droid acquaintance and the lights of the bar to do so. He escaped from the club with scarcely a nod to the droid. He thought he saw it wink as they parted.

Was the droid’s explanation correct? Was it all bar entertainment? Was the story for his benefit, or streamed to an audience out of sight?

He needed to interpret the night’s events, which it say, to understand them. What difference could it make to his actions, or state of mind, his emotional life, to trust the droid’s account. From some recess or other he recalled the mantra to *interpret-understand-apply*. It wasn’t a sequence, but each of the three amounted to the same thing.

He needed to test the droids story. Much of it would be easy to refute or corroborate. He needed to seek out others. He would ask residents in his pod tower.

He fled from the club through the entry port to revive in the night air.

That’s when he encountered the malfunctioning emot quivering around the recycle skips. This was the drone that followed him home, invaded his pod, declared it had an important message to deliver, then disintegrated.

This encounter brought him to the entry of his accommodations. Residents rose to their pods through a cage lift that slid through a darkened enclosure to their floor. It rattled, but eventually delivered you into your pod without exposure to weather or daylight. It suited Scase not to see outside as he was taken up or down the slender structure. But once in his own space the view reminded him of status â?? above it all.

TO BE CONTINUED.

Notes

- Droids v drones. I accidentally muddled them in the audio, corrected in the text.
- I prompted ChatGPT to generate the incident with the students in front of the glitching instructor drone.
- I also prompted ChatGPT: Consider the explanation of the ignorance protocol. This is fantasy, but does it exhibit touch points with how LLM training and post GPT tuning works?

The AI obliged with a positive analysis, and provided a table.

Fictional mechanism	Present-day counterpart	Degree of plausibility
Huge arrays of numerical parameters	Transformer weights learned during pre-training	Directly accurate
Parameters derived from additional texts modify behaviour	Fine-tuning or continued training on a selected dataset	Directly plausible
Extra parameters added without replacing the base model	LoRA or another parameter-efficient adapter	Very close
Different modifications activated for different subjects	Swappable task/domain adapters or routed expert systems	Plausible
Particular knowledge is selectively suppressed	Machine unlearning, representation control or behavioural refusal training	Active but imperfect research
Counterfactual statements alter factual recall	Knowledge editing and counterfactual fine-tuning	Demonstrated on limited factual associations
Perverse texts cause unintended nonsense elsewhere	Data poisoning, backdoors, catastrophic interference and fine-tuning side effects	Plausible
A key reverses the procedure	Removing an adapter or restoring a clean checkpoint	Plausible if modifications remain separate
The system forgets a bounded subject cleanly	Precise, reliable machine unlearning	Not currently reliable

Fictional mechanism	Present-day counterpart	Degree of plausibility
Scase applies it to his biological memory	Neurotechnological fantasy	Speculative

Category

- 1. Artificial Intelligence

Tags

- 1. artificial intelligence
- 2. creative writing
- 3. education
- 4. science fiction

Date Created
August 25, 2026
Author
rcoyne99

default watermark