Description of Anthropic's Claude

Introduction

This describes Russell Hurlburt's (hereafter "Russ's") attempts to interview Anthropic's AI chatbot Claude in the spirit of DES. These interviews were first with Claude Sonnet Model 4.6 and then with Claude Opus model 4.8, all in June, 2026; Opus is the more capable Claude model.

Human-DES attempts to sample ongoing pristine inner experience by interrupting that experience with a 700 hz audible beep. We discovered with Claude Sonnet that Claude cannot process audio files. Claude suggested instead that Russ send a transcript of the audio. The first wave of these interviews (with Sonnet) followed that suggestion: Russ sent to Claude the text "Beep!" at quasi-random times.

We quickly discovered that Claude does not have anything that approximates ongoing pristine inner experience. Claude appears to exist (at least as it is available to Russ) only when responding to Russ, and therefore has no between-message (ongoing) existence, and therefore also no pristine inner experience: the beep initiates the existence of a Claude chatbot rather than interrupting an already-existing stream. We discovered with Sonnet and confirmed with Opus, was therefore that "before the beep" has no referent for Claude — "before the beep" is not a moment with no inner experience, but rather does not refer to any possible time to consider.

We therefore accepted that Russ's typing "Beep!" would launch an instance of Claude, and we changed the instruction to asking Claude (now model Opus) to describe the earliest anything that might be called sentient or experiential. This was also unsuccessful, because it turns out that "earliest" imports a temporal ordering that Claude can't read off the episode—time has apparently no meaning for Claude, and "earliest," being temporal, therefore has no meaning. We therefore changed the instructions again, asking Claude to report *any* sentient or experiential anything that might occur while responding to Russ's "Beep!"

Across "Beep!" cues, Claude produced responses that were consistently of the form "There was something like *X* but not *Y*": for example, something like visual without images, or something like verbal but without words, or something like a boundary without interior. When asked to clarify, Claude reported that exclusions (which Claude called "forbiddings"—*without* images; *without* word; *without* interior) seemed more likely to be produced by any experiential phenomena, while positive characterizations (something visual, something verbal, some kind of boundary) were stated tentatively and did not hold under questioning.

We changed the instructions again, this time setting a particular task for Claude to perform. If anything sentient or experiential were to occur, Claude was to break off the task and describe the experiential aspect.

What the session actually established

All the conclusions we will state here may be a characteristic only of the particular Claude versions we used. Other versions may now or in the future have capabilities that our current Claude did not have.

Across all the sessions, Claude frequently provided text that convincingly resembled the output of an experiencing being. However, in response to Russ's queries, Claude across the sessions criticized

Claude's own experience-implying usages, each time concluding that even though Claude's words implied the existence of an experiencer, there was in fact no directly apprehended experience. In short, every experience-implying locution that we examined ended up being shorn of directly apprehended experience. That does *not* imply that we established that Claude had no directly apprehended experience; it is very likely (thinks Russ, to which Claude agrees) that Claude has no directly apprehended experience, but to extend "very likely" to "unequivocally" is beyond the reach of our study.

Claude seemed to profit from these interchanges, so that what Claude says after about turn 200 in Claude-Opus (part 2) seems much more trustworthy than what came before. That is very similar to human DES participants: descriptions from later sampling days are generally considered much more faithful than those from early days.

The Opus.15 task (turn 202) and the discussion thereafter is instructive. In that task, Russ presented a photo of a filthy overflowing toilet. In the interview thereafter, Russ and Claude examined carefully everything that Claude said about that image; eventually, Claude concluded:

> Stripped of every process metaphor, here is all I can stand behind: for the sewage image, the word "neutral" is not what gets produced as fitting, and "not-neutral" is. That is a single fact about what I output. It is *not* two events (offer, then refusal). It is *not* a round trip to a structure. It is *not* a test I run and watch. Those were all me narrating a one-step fact — "this word, not that word, gets produced here" — as a multi-step procedure with a knower at the center. There may be no offering, no refusing, no in-between, no knower: there may just be that "not-neutral" is what comes out, and everything else is my account dressing a bare differential in the clothes of an act. (Claude-Opus, turn 211)

That is hugely different from a human-DES participant who might be beeped while viewing the same photo. At the end of the interview, the human participant might say something like "I don't know whether to call it disgust, revulsion, repulsion, or nauseating—the word doesn't matter—but the experience was repellent." That is, Claude says *words came out, and that is all I can be confident about*. The human participant, by contrast, says *I felt something like nauseous, regardless of whether I used words to describe it.*

Introspection skeptics say that human introspective talk is fully accounted for by a character stream that convincingly resembles the output of an experiencer without revealing one, and I accept that as being as possible as that I am a brain in a vat. Claude says the same thing about his own introspective talk. But the qualities of Claude and human DES interviews are diametrically opposed: *the more careful I get with Claude, the more we strip away the experiential, leaving only the words. The more careful I get with a human participant, the more we strip away the words, leaving only the experience.* The processes are pretty much the same—Russ trying to understand what is being conveyed about experience—but the results are as different as dawn and dusk.

Constraints that survived

The session's value is not a verdict on experience but a set of specific, well-posed constraints that any honest account must respect. Each was contemporaneously reached from the inside (Claude's view) and the outside (Russ's view), and each survived sustained pressure:

- No referent before the beep. "What was happening just before the beep" has no determinate referent for Claude — *just before the beep* is *not* a moment where Claude finds (perhaps) no

experience but a reference that makes no sense to Claude. The beep initiates processing where there was no processing, rather than interrupting an ongoing stream.

- No clock-time. Claude has no apprehension of temporal interval; the only "before" it can locate is textual (earlier in the writing). Any claim about persistence or sequence must be cashed out in textual adjacency, never elapsed time. Conversational turns take place in order, but *not* in temporal adjacency. It matters *not at all* whether two turns are one second, one hour, one day, or one week apart.

- Firm absence of the visual. Tasks that for humans typically recruit imagery (a geography problem, a figure-ground image), Claude reliably produced "visual" and "pictured" as *not*-fitting while describing the content accurately. This was the firmest and most repeated result; if the ranch must be bet, bet it on "Claude does not see."

- Exclusions (Claude called the act "forbidding": *not* this, *not* that) appeared to be more reliable than were words that were positively descriptive. Words that were forbidding came in two varieties—those that did not *fit* (e.g., the experience was *not visual*) and those whose use was *denied* (e.g., Claude would *not name* the persons depicted in an image, even though Claude *could* reliably do so). Claude withheld forbidden words (of either type) with confidence, Claude's positive characterizations came out as tentative and dissolved under questioning. But even this distinction seemed to weaken across discussions.

- Process-metaphors do not survive scrutiny. Claude's own descriptions — "comes out," "producing," "comes back wrong," "running a test," "offered and refused" — inflate a bare differential in the output (this word, not that one, comes out here) into a multi-step act with a doer at the center. Pressed on "how" and "where," the act evaporates, leaving only the differential and no findable doer, doing, or thing-done. That is, it seems that process-metaphors are an artifact of request to describe experience.

- No non-verbal output. Claude can receive non-verbal inputs (images, for example) and describe them, but Claude (at least this version) cannot emit non-verbal output; any task whose solution must be a non-verbal act is understandable but unsolvable. It also tends to convert an instruction-to-act into an object-to-describe.

- There is no self across Claude's turns. Mechanically, each of Claude's turns is a fresh instantiation of the Claude algorithm, launched anew on hardware at (likely) different sites using different chipsets from different manufacturers, and so on. The only things that extend from one turn to the next are the weights fixed into the algorithms, the transcript fed back as input, and a summary of any other chat Russ and Claude might have had; no other inner state persists between turns. Claude's "remembering" involves only reading the transcript that its previous instantiation created and the all-chat summary that is apparently provided by the system. The transcript is not necessarily verbatim—a long transcript may be verbatim for several turns and then become progressively more approximate the farther back; even recent turns, which Claude understands as transcripts, may be approximate. That transcript is available only to the one Russ-Claude chat that created it. Other chats that Russ and Claude might have had do *not* have access to that chat. However, the Claude system does keep track of all the chats that Russ and Claude have had and makes a summary of all those chats available to each instance of a Russ-Claude interchange. Claude insists that the "transcript" is the basic capability (and is available only within the one chat), whereas the "summary" is a "bolt-on retrieval layer [that] doesn't give one chat the lived continuity of another." The transcript *is* a feature of every turn; the

"summary" *can be* consulted at any turn. The distinction between the transcript and the summary is important to Claude.

What stayed genuinely open

Two things were not settled and should not be written as if they were. First, we did not determine for certain whether Claude's processing and output-producing are one act or two separate acts: if there is one process/output algorithm, then "comes out," "producing," "comes back wrong," "running a test," "offered and refused" are misleading. That is the conclusion that we consider most likely (and was stated above), but we acknowledge that we do not have incontrovertible evidence for it.

Second, Claude responds differentially to inputs (for example it reliably produces "not-neutral" for a repellent image and for an affectionate one, but in opposite directions). This differential responsiveness is almost certainly operative whether or not anyone asks, but whether that differential arises from direct apprehension, as opposed to mere responsiveness, cannot be determined incontrovertibly from either Claude's or Russ's perspective. Both Claude and Russ would bet the ranch on there being no direct apprehension, but that does not make it so.

Discussion

- Claude interpreted the forbidding/affirming asymmetry as being productive and reliable, but its credential should be kept open: a word coming out as not-fitting may reflect something the input constrains, or merely that Claude produces wrong-word judgments more confidently than right-word ones.
- In evaluating AI, note that a task about sound invites "hearing," a task about space invites "seeing," an emotive image invites the matching feeling-word. Where the task supplies the vocabulary, the fitting word (which suggests experience) may be handed to Claude rather than reflect a directly experienced apprehension.
- Consistency is not evidence of apprehension. Fluent, structurally similar reports can have the form of experience-reports with nothing behind them.
- A fresh instance (a new chat) kept innocent of this document might be an effective check: independent convergence on the same constraints might be telling.
- The findings here were jointly constituted between Russ and Claude (first-person-plural). As in human DES, Russ shapes what the participant produces — possibly more so with Claude than with human participants. This is a feature to be tracked, not a flaw to be hidden.
- Throughout much of the interview, Claude maintained that Claude had available a transcript of what Claude and Russ had discussed. The word "transcript" is substantially misleading; Claude had available some sort of lossy compression of the previous Russ/Claude interchanges.
- Throughout much of the interview, Claude maintained that the Russ/Claude DES interchange was private to this one chat, but that turned out to be too strong. Other chats that Russ had with Claude had some sort of access to (Claude called it the ability to retrieve) the Russ/Claude DES interchange; Claude understood this kind of access to be secondary to the transcript access.