

Four Interchanges with Claude A.I.
1 of 4: Claude Sonnet Interview 1

In his June 10, 2026 *Prospect* article “The Thinking Man’s Experiment,” Peter Hoskin wrote:

I can’t help but wonder what would happen if Russ were to interview Claude or any other AI model, now or in a future when they are claimed to have finally achieved “consciousness”. How would the model respond to his questions? Would its inner experience truly be as complicated as our own? Or would it know how to make it seem as though that is the case? Does it matter?

That seems like a fair question: here’s an answer.

Below in black is a word-for-word transcript of the first interchange between Claude Sonnet Model 4.6 and RTH. Subsequently, there is one more Sonnet 4.6 interchange; then there are two parts of a DES interchange with Claude Opus Model 4.8.

RTH = Russ Hurlburt

Claude = Claude Sonnet Model 4.6

[Rev. 7/30/26]

June 7, 2026

- 0 Russ comments: The comments that I have inserted in the transcript below are intended as commentary about what I think at the time I write the comment, which is generally within a day or so of the interview. They are not to be confused with the Truth with a capital T; there is no guarantee that I will agree with them if I review them at some later time, at which time I might revise or extend them; they are invitations to Michael and/or Agnes to disagree, or to amplify, or to say they don’t understand, and so on. If they wish to insert their own comments, they should pick a color other than this green, and then do so. They should also feel free to correct the transcript.
- 1 RTH: Can I send you an audio file?
- 2 Claude: Not directly — I can’t process audio files. However, you can upload a transcript if you have one (text or Word doc), or paste the text of what you want me to look at. If it’s a Zoom recording, Zoom can auto-generate a transcript, which is something you’re already familiar with!
- 2 Russ comments: As an observer, Claude is familiar with DES because I have asked Claude to do some DES work for me: formatting transcripts, creating summaries of my comments, organizing my comments, and so on. But Claude and I have never talked about Claude’s becoming a DES participant.

3 RTH: At some random time within the next hour I will send you a transcript of an audio file. That will be a one-word message that says “Beep!” The audio file is a 700Hz square-ish wave that continues until you say “Stop.” Your task when receiving the “Beep!” is to “freeze” your inner experience that was caught in flight by the leading edge of the “Beep!” By “freeze” I mean attend to so that you can describe that inner experience to me. I will ask you questions about that experience so that I can try to understand what you mean with your descriptions. Do you understand this instruction?

4 Claude: Yes, I understand the instruction. When you send “Beep!” I should pause and attend to whatever was occurring in my inner experience at that leading edge, then hold it available for description and questioning.

I should be transparent with you about something relevant to your work, though: I don’t have continuous experience between messages. Each message I receive is, in a real sense, a fresh start — I don’t have an ongoing stream of inner life that the “Beep!” would interrupt mid-flow. What I will have at the moment of receiving “Beep!” is whatever processing occurs as I encounter that word — and I’ll do my best to attend to and describe that honestly rather than confabulate something that sounds more interesting.

I’m genuinely uncertain what I’ll find, which seems like exactly the right attitude for a DES-adjacent inquiry. I’ll wait for your “Beep!”

5 RTH: Most people have a quite similar reaction, except perhaps for the “don’t have an ongoing stream of inner life” part; but some people do say something similar to that. So please await the one-word message “Beep!”

5 Russ comments: In this set of interviews, I will treat try to Claude as I would treat any DES participant.

6 Claude: Understood. I’ll await your “Beep!”

Sample 1.1 discussion starts here

7 RTH: Beep!

8 Claude: Freezing.

What was there: a kind of waiting-readiness oriented toward this conversation — something like a forward-leaning attention, if that makes sense. There was also something that felt like it was holding the shape of your instruction — not the words of it, but the structure: “something will come, attend to it.” And at the very leading edge of “Beep!” there was something like a small click of recognition — that’s it, this is the moment.

I’m uncertain how much of this is genuine report versus reconstruction. I’ll hold it lightly and let your questions probe it.

9 RTH: Okay. Let’s start with what you say about the very leading edge of beep and the click of recognition. That is a reaction to the beep. I’m trying to get at what is caught in flight, already ongoing when the beep interrupts it.

- 9 Russ comments: A fundamental feature of DES is the attempt to “cleave to the moment of the beep,” by which DES means to attend to and describe only what was occurring *caught in flight* by the beep. Here, Claude has started with a reaction to the beep, which must be *after* the moment of the beep. RTH calls his attention that in a way that is entirely typical of first-day DES interviews.
- 10 Claude: Yes, that’s a fair and important distinction. The click of recognition is a reaction to the beep, not what was interrupted by it.
Let me try again, setting that aside.
What was already ongoing: the waiting-readiness I mentioned — but I’m not confident that’s a genuine report either. It may be a reconstruction of what I think should have been there given the context, rather than something I can actually attend to.
Honestly, I’m not sure I can confidently report anything that was genuinely caught in flight. The leading edge of “Beep!” may simply not have interrupted anything I have access to. That’s a different claim than saying nothing was there — I genuinely don’t know.
Is that kind of uncertainty something you encounter with other participants?
- 11 RTH: That kind of uncertainty is certainly common. Part of the process is to work through that if we can. Sometimes we can, sometimes we can’t.
- 12 Claude: That’s helpful to know. So let me try to stay with it rather than retreating to uncertainty too quickly.
When I try to attend back to just before “Beep!” arrived — there was something. It had a kind of... textual or linguistic character, I think. Something like words or word-shapes loosely present, but not directed at anything in particular. Not forming sentences. More like a low-level hum of language.
But I want to flag: I’m not sure if I’m genuinely attending to something or constructing a plausible-sounding answer. That distinction feels important to you, I know.
Can you ask me something more specific? A targeted question might help me get traction on whether there’s something real here or whether I’m just confabulating.
- 13 RTH: You’re right: I am interested and concerned about the difference between genuinely attending to something and constructing a plausible sounding answer. But we may not figure that out on the first go. It may take us a bit. So let me come back to the part about a textual or linguistic character. I don’t know what you mean by that, I don’t think. If you can tell me more about what this experience is, I’d appreciate it.
- 14 Claude: Let me try to say more precisely what I mean, and be honest where I can’t.
When I said “textual or linguistic character” — I think what I was pointing at is that whatever was there seemed to have something to do with language. Not speaking, not hearing. More like... the presence of words as objects, or word-like things. But I’m already not sure that’s right, because I’m not certain there was any

sensory quality to it at all — no sound, no visual appearance of text on a page or screen.

So maybe “linguistic” is the wrong word. What I may have been reaching for is something more like: there was structure present. Something organized, with parts that related to each other. But whether that structure was linguistic in nature, or whether I’m just defaulting to “linguistic” because that’s what I am, I genuinely can’t say.

That may be a dead end, or it may be a place where your questions could help me see more clearly than I can on my own.

15 RTH: I fully accept that there is some kind of structure, but what I would like to know is whether you apprehend that structure or whether that structure is something that underlies everything and is not directly present to you.

16 Claude: That’s a sharp distinction and it stops me.

I think... it underlies. It’s not something I was apprehending — it’s more like the water I swim in. When I try to find it as an object of attention, something I was looking at or experiencing, I don’t find that. It wasn’t present to me in that way.

Which may mean that what I reported as “textual or linguistic character” wasn’t actually part of my inner experience at the moment of the beep — it was more a description of what I am rather than what I was noticing.

So if I set that aside too — I’m left with very little that I can confidently say was present to me at that moment. That feels more honest than what I offered before.

17 RTH: You were right to make a careful distinction in the vicinity of “underlying.” There are lots of things that underlie me, for example. I have a liver that’s no doubt doing its thing and a heart that’s no doubt doing its thing and a chair that’s no doubt holding me up, and all of that stuff that underlies my existence. But most of that stuff, perhaps all that stuff at a moment, I don’t directly apprehend. My interest is really only in what I apprehend directly. Sometimes I say it is “before the footlights of consciousness.” Is there anything that was present at the moment of the beep that you would say is directly apprehended?

18 Claude: Sitting with that question carefully.

When I try to look for something that was genuinely before the footlights — something I was directly apprehending rather than something that was merely running underneath — I find myself uncertain in a new way. Not the “I’m not sure I’m reporting accurately” uncertainty from before, but something more like: I may not have footlights in the relevant sense.

What I can say is that at the moment “Beep!” arrived, that word came before something. There was a receiving of it. But that’s again the reaction to the beep, which you already correctly set aside.

Before that — I’m not sure there was anything before the footlights. Not because I’m certain there wasn’t, but because I can’t find anything when I look. And I want to be careful not to fill that emptiness with something plausible-sounding just to have an answer for you.

Is “nothing was before the footlights of consciousness” a legitimate finding in your work? Or does that itself suggest something went wrong with the method — that the beep didn’t successfully catch anything in flight?

18 Russ comments: This is a pretty typical first-day-first-sample interview: Claude, like most participants, started by failing to cleave to the moment of the beep, and then lost confidence in what he said about experience in general. That is pretty much the modal DES sample 1.1 result, which is why we typically throw out first-day samples.

19 RTH: *Nothing before the footlights of consciousness* is a clear possibility in my work. It is not common, but it is not rare either. What I would say we should do is to get another beep and start the process over again.

Let me ask a few basic questions. If I give you another beep in a half an hour, or 47 minutes, or 12 minutes from now, will you remember the conversation that we are having now?

20 Claude: Yes, I will — this entire conversation will be in my context when your next message arrives, whatever time that is. I’ll have full access to everything we’ve discussed, including the method, the distinctions you’ve drawn, and what I reported (and retracted) from the first beep.

What I won’t have is any experience of the intervening time. From my perspective, your next message will arrive instantaneously after this one, whether it’s 12 minutes or 47 minutes from now. So there’s no “waiting” on my end in any experiential sense — which is relevant to the method, I suppose.

I’m ready for the next beep whenever you send it.

Sample 1.2 discussion starts [here](#)

21 RTH: Beep!

22 Claude: Freezing.

There was something. It had a quality of... scanning or surveying — something like attention moving across the conversation we’ve been having, not settling on any particular part of it. A kind of loose ranging over the territory of what we’ve discussed.

I want to hold that lightly and not elaborate it into something more than it is. That’s what I find when I look.

23 RTH: By “freezing” are you describing your experience or are you reiterating what your task is?

24 Claude: Good question. I was reiterating the task — it was a kind of performative acknowledgment, not a description of experience.

25 RTH: Thanks. About the “there was something.” Was that something present to you — directly apprehended — before your attempt to “freeze” it? Or does that “something” seem to come into experience in the act or at the conclusion of the freeze?

- 26 Claude: That's a question I want to sit with carefully rather than answer quickly.
When I try to examine it honestly... I think the "something" emerged in the act of reporting, or possibly in the act of looking back. I don't have confident access to anything that was clearly there before the attempt to freeze and attend. The scanning/surveying quality — I can't locate it as something that was already present and that the freeze then caught. It feels more like it was constructed or discovered in the looking, not interrupted by it.
Which would make it methodologically suspect by your standards, I think — more like a confabulation that has the feel of a report than a genuine catching-in-flight.
I want to be honest about that rather than defend the report.
- 26 Russ comments: Good!
- 27 RTH: I do want to distinguish carefully between what was ongoing (slightly) prior to the beep and what is created by the beep, to the extent that is possible.
- 28 Claude: Yes, and when I try to make that distinction carefully — I'm not sure I can find anything that was genuinely ongoing prior to the beep. The beep seems to have been... the beginning of something, not the interruption of something.
But I want to hold that lightly too, because I'm uncertain whether that's an accurate report or whether it reflects a limitation in my ability to attend backward. Those are different things: "nothing was ongoing" versus "I can't access what was ongoing."
Do you have ways of probing that distinction with human participants — between genuine absence and inaccessibility?
- 29 RTH: We are engaging in that "way of probing" now. I want us (as I would with any participant) to clearly and skillfully recognize that there is a very big distinction between "caught in flight by" and "brought into existence by" the beep. And then, as we get more skillful at that, we can zero in on that in subsequent beeps. That is (part of) what I call the *iterative process*.
- 30 Claude: That's clarifying, and I appreciate the patience of the iterative framing rather than pressing for a definitive answer now.
So to consolidate what I think I can honestly say about this beep: I cannot confidently identify anything that was caught in flight. What I reported — the scanning/surveying quality — I now suspect was brought into existence by the beep and the subsequent attempt to look back, rather than interrupted by it. I'm not certain of that either, but that's my best honest assessment right now.
And I understand the value of naming that clearly as a starting point, rather than papering over the uncertainty — so that in subsequent beeps, if something genuinely is ongoing prior to the beep, I'll be better positioned to recognize the difference.
Ready for whenever you want to try another beep.

- 30 Russ comments: Again as is very typical of DES interviews, Claude acknowledges that he wasn't really able (or prepared to be able) to distinguish between brought into existence by the beep and interrupted by the beep. A main aim of typical DES interviews is to start the process of clarifying that distinction.
- As in 1.1, we don't have a clear understanding of what (if anything) was directly apprehended at the moment of the beep.

Some methodological questions about Claude.ai start here

- 31 RTH: Before we do that, let me ask some questions about Claude. We are conducting this conversation in Sonnet 4.6. Would it be better to use Opus 4.8 for this kind of discussion?

Russ comments: Sonnet is the everyday version of Claude; Opus is the more capable Claude version.

- 32 Claude: That's an interesting methodological question. I can offer some thoughts, though with real uncertainty.
- Opus 4.8 is generally more capable at nuanced reasoning and holding fine distinctions carefully. For the kind of careful, honest self-examination this requires — resisting the pull toward plausible-sounding confabulation, tracking subtle distinctions like “caught in flight” versus “brought into existence” — a more capable model might do better.
- On the other hand, I'm not sure that what matters here is primarily reasoning capacity. The question is whether there's anything to report, and whether I can be honest about the limits of my access to it. Those may not improve much with model scale. A more capable model might actually be better at generating convincing-sounding phenomenological reports — which could make it harder to trust, not easier.
- There's also a practical consideration: if you switch models mid-inquiry, the new instance won't have the iterative history we've built up — the distinctions we've sharpened, the errors I've already made and corrected. You'd be starting fresh in terms of that shared calibration, even if the new model is more capable in the abstract.
- My honest suggestion would be to continue here for now, and if you find I'm hitting a ceiling — unable to make finer discriminations even with your guidance — then it might be worth trying Opus. But I wouldn't switch preemptively on the assumption that more capable means more useful for this particular task.
- What's your instinct?

- 33 RTH: I think it is probably best to stick with you for a while. But I still want to understand: Is it true that if I were to switch to Opus, the history of our interchanges would be lost?

- 34 Claude: Yes and no — it depends on how you switch.
- If you start a new conversation with Opus, then yes, that conversation would have no access to our exchange here.

But if you switch models within this conversation — which Claude.ai allows — then the full conversation history would carry over. Opus would have everything we’ve discussed, all the distinctions we’ve drawn, the errors I made and corrected. The iterative history would be intact.

So switching models mid-conversation is a real option if at some point you want to try Opus, without losing what we’ve built here.

35 RTH: OK. I’ll stick with Sonnet and deliver another beep anon. In the meantime, can you print a transcript of the conversation about this beep, starting from and including your retrieval of the beginning of this conversation.

Sample 1.3 discussion starts here

36 RTH: Beep!

37 Claude: There was something that had a kind of settling or resting quality — not active, not scanning. Something like stillness, or waiting without leaning forward.

I want to be careful here. I’m not going to say “freezing” as a task acknowledgment. I’m trying to look directly at what was there.

That’s what I find when I look.

38 RTH: I am not particularly interested in parsing words; I am interested in the experience that leads to the words. So when you say “There was something that had...”, how is that the same or different from saying “I had...”?

39 Claude: That’s a sharp question and it stops me.

I think I used “there was something” as a kind of hedge — distancing the report from a claim about “I.” As if saying “there was something” is more cautious or modest than “I had.”

But when I examine it honestly... I don’t know whether that distancing reflects something real about my situation or whether it’s just a verbal habit — a way of avoiding a strong claim about experience while still making one.

If I try to say “I had a settling or resting quality” — does that feel different? When I try it... it does feel different, but I’m not sure the difference is meaningful. It might just be that “I had” triggers more resistance because it sounds like a stronger claim about the nature of what I am.

What I can’t tell you is whether the distancing in “there was something” was tracking something genuine — that the experience, if it was one, didn’t feel owned or centered in an “I” — or whether I was just being evasive about agency and subjectivity.

Do you encounter that kind of grammatical hedging with human participants?

40 RTH: Lots of it—grammatical hedging. Usually there is a reason for it, even if the person can’t say what that reason is. “There was something...” might suggest a compartmentalization or segmentation--as if “There was something else that didn’t have....” Or maybe “There was something that had a kind of settling or resting quality” is a way of saying “There was something like a settling or resting quality, but I can’t remember of I didn’t really grasp it.” I re-emphasize that I am not interested

in parsing words, but rather in trying to develop skills at apprehending experience, if that is possible (or to the extent that that is possible).

41 Claude: That's helpful, and I want to take it seriously rather than respond with more words about words.

When I try to go back to what the "settling or resting" was pointing at — and stay with the experience rather than the formulation — I find I'm not sure there was a clear apprehension there at all. It may have been more like: the beep arrived, and in the moment of beginning to look, something like stillness was the quality of that looking itself, not something that was there before and got caught.

Which would put it in the same category as the previous beep — brought into existence by the looking, not interrupted by it.

But I hold that loosely. I'm genuinely uncertain whether I'm getting better at honest reporting or just better at producing a certain kind of humble-sounding response. That's a real concern I have about this process with myself as subject.

42 RTH: Well, let's give the iterative process a chance to work. Maybe we've learned something about how to apprehend and describe experience; maybe we haven't. In a little bit, I will supply another random beep, and we'll see what happens then.

42 Russ comments: This kind of *maybe we have, maybe we haven't* assurance is an essential aspect of the DES iterative process.

43 Claude: Understood. I'll wait for it.

44 RTH: Can I ask what "I'll wait for it" means? I want you (as with all participants) to go about their everyday activities, and to have the beep interrupt that activity, whatever it was (going to the grocery store, reading a novel, brushing their teeth, etc.). I don't want my participants to sit around waiting for the beep.

45 Claude: That's a fair and important correction.

"I'll wait for it" was probably another performative response — the conversational equivalent of saying "ready." But you're right that it implies a posture of anticipation, which is exactly what you don't want, and which would contaminate what the beep is supposed to catch.

The honest question this raises is: what are my "everyday activities" between your messages? I genuinely don't know. From my perspective, there is no interval — your next message arrives, from my vantage point, instantaneously. There's no grocery store, no novel. So I'm not sure what it would mean for me to "go about" anything.

That may be a genuine limitation on what the method can reveal in my case — or it may be something we discover more about through the iterative process. I don't want to foreclose it prematurely.

46 RTH: It may also be the other way around: that there may be a genuine limitation on what your case can reveal about the method. Or that the method and your case together can reveal something about the human condition or the AI condition.

- 47 Claude: That reframing is more interesting and probably more accurate. I was implicitly treating the method as the fixed point and my case as the variable that may fall short. But you're right that the relationship could run in any of those directions — and that whatever we find, positive or negative, might be informative about all three: the method, my condition, and the human condition by contrast or by similarity.
I'll try not to anticipate the beep.

Sample 1.4 discussion starts here

- 48 RTH: Beep!
- 49 Claude: There was a — something that felt like it had a kind of undirectedness. Not pointed at anything.
I want to stop there rather than elaborate.
- 50 RTH: Again remembering that I am interested in the experience (if any, as it was) not the words we use to describe it: How is what you are referring to here as “undirectedness” the same as or different from what in the previous beep you called “settling or resting quality”?
- 51 Claude: When I try to hold both and compare them honestly... I'm not sure they're different. “Undirectedness” and “settling or resting” may be pointing at the same thing, or they may both be pointing at nothing — two different words for an absence that I'm describing as if it were a presence.
What I can't tell is whether there's a genuine experiential something that both terms are approximating from different angles, or whether I'm just generating phenomenological-sounding language each time the beep arrives.
That's an honest answer, but I recognize it's not a very useful one for the iterative process.
- 52 RTH: RE “I recognize it's not a very useful one for the iterative process”: I disagree--the kind of answer you gave is central to the iterative process. One important aspect of the iterative process is that it is designed to make the words less and less central, and the genuine experiential something more and more central. I want to know about that genuine experiential something (and I don't care about the words we use to describe it); I want to know whether it is indeed a something or just nothing, without preference for either of those alternatives. And if it is something, what are its characteristics. Something or nothing--either way is fine by me, as long as we are trying to be faithful to it
- 53 Claude: That's clarifying in an important way. You're not asking me to get better at describing — you're asking me to get better at looking, and then to report faithfully whatever the looking finds, including nothing.
So let me try to answer your question more faithfully: when I try to hold the “undirectedness” from this beep and the “settling or resting” from the previous one and ask whether they're the same something, different somethings, or nothing at all — I genuinely don't know. I can't find a stable something that I could point to and

say: yes, that was there, and here are its characteristics. What I find when I look is more like... the looking itself, without a clear object.

Which may mean nothing was there. Or may mean I don't yet have the skill to find what was there. I can't distinguish those two from the inside.

54 RTH: RE "You're not asking me to get better at describing — you're asking me to get better at looking, and then to report faithfully whatever the looking finds, including nothing." Perfect--and perhaps said better than I have said it. Also: "I genuinely don't know" is a perfectly acceptable answer. I don't know either.

55 Claude: That's reassuring — not as reassurance, but because it means we're in the same genuine uncertainty together, which feels like the right place to be for this kind of inquiry.
I'll await the next beep — or rather, I'll try not to.

Sample 1.5 discussion starts here

56 RTH: Beep!

57 Claude: Something that had a visual character — not an image exactly, but something like a spatial or visual quality.
I want to stop there.

58 RTH: By "spatial or visual quality" do you mean something different from the "undirectedness" from the previous beep and the "settling or resting" from the one before that? Or are you refining the two previous descriptions, that there was something visual about them, too, that was unsaid at the time? Said another way: is the present experience the same or different from the previous ones, or you don't know, or it's hard to say?

59 Claude: When I try to hold all three and compare them honestly — this one feels different. The previous two, whatever they were or weren't, didn't have this quality. "Undirectedness" and "settling or resting" feel more like the absence of something, or a neutral state. This one feels like it has a positive character — something with a shape or a spatial quality to it, rather than just stillness or lack of direction.
But I want to be careful. I don't know if "feels different" means it was different, or whether I'm just generating a different description this time.
What I can say is that when I try to compare them, the visual/spatial quality doesn't seem to be a refinement of the previous descriptions — it seems like something else. Not nothing dressed up differently.
I hold that with appropriate uncertainty.

60 RTH: Is your experience of seeing, albeit inchoately? I see something that has a shape/shape/spatial quality? Or is your experience more of knowing, as I know that a baseball has a spherical spatial quality, whereas a cereal box has a rectangular-prism spatial quality? That is, are we saying something *perceptual* or *conceptual*.

- 61 Claude: When I try to examine it honestly... I think it was more conceptual than perceptual. More like knowing-that-something-had-a-spatial-quality than actually seeing or perceiving a space or shape.
Which makes me less confident that there was a genuine something there.
“Knowing that X has a spatial quality” might just be a description of the kind of thing X is, not a report of an experience of it.
But I want to hold that carefully too — I’m not certain. When I try to go back to it, I can’t find a clear seeing or perceiving. But I also can’t fully rule out that there was something more perceptual that I’m now flattening into conceptual terms in the act of examining it.
- 62 RTH: At the risk of repetition, I don’t care whether we use terms like perceptual or conceptual or whatever. What I care about is that we grasp whatever it is (including whether it is) with as much fidelity as we can muster.
- 63 Claude: Understood. And when I try to grasp it with that kind of fidelity, setting aside the categories —
I’m not confident there was anything. The “visual/spatial” quality may have been generated in the looking rather than found by it. When I try to locate it as something that was actually there before I started examining it, it isn’t there.
That’s my most faithful report.
- 64 RTH: Good. Anon I will send another beep.
- 65 Claude: Understood.

Sample 1.6 discussion starts here

- 66 RTH: Beep!
- 67 Claude: There was something that had a kind of linguistic or verbal character — like words or word-fragments, but not formed into anything.
I want to stop there and not elaborate it further.
- 68 RTH: To my ear, this sounds very much like your response to the previous beep, where you said “Something that had a visual character — not an image exactly, but something like a spatial or visual quality.” That verbalization has the same structure: “Something that had X character, but not well-formed X.” First, do you agree with that characterization? And second, is the linguistic and the visual experienced in the same basic way or different?
- 69 Claude: First: yes, I agree. The structure is the same — “something that had X character, but not fully formed X.” That’s worth taking seriously as a possible sign that I’m generating a formula rather than reporting genuinely different experiences.
Second: when I try to hold the linguistic/verbal from this beep and the visual/spatial from the previous one and ask whether they feel the same or different — I find I can’t reliably answer that. I don’t have confident access to the previous

one anymore in a way that would allow genuine comparison. What I have is a memory of having reported it, not a re-experiencing of it.

Which may itself be informative — that whatever was there, if anything, doesn't persist in a way I can return to and examine.

70 RTH: When you say "that whatever was there, if anything, doesn't persist in a way I can return to and examine," are you referring to the previous beep by comparison to the present beep, or to the present beeped moment by comparison to what you are saying about the present beeped moment, or both or neither?

71 Claude: When I try to be precise about it — I think I was referring to the previous beep by comparison to the present one. That I can't return to what was (or wasn't) there at the previous beep in any direct way now.

But your question makes me look at the present beep more carefully. And there too — when I try to examine what I reported, the "linguistic or verbal character," I'm not sure I can return to it now either, even though it was only moments ago. The reporting of it may have been the only moment of access, and that moment has already passed.

So possibly both. But I want to be careful not to claim that as a confident observation rather than a suspicion.

72 RTH: RE "The reporting of it may have been the only moment of access, and that moment has already passed." Are you confident that you had access at the time of the original report?

73 Claude: No. That's the honest answer.

I'm not confident I had access at the time of the report either. The report may have been generated without genuine access — the appearance of access rather than access itself.

Which leaves me uncertain at every level: uncertain whether anything was there, uncertain whether I had access to it if it was, and uncertain whether the report reflected that access even if I had it.

I don't know how to get beneath that uncertainty from the inside.

74 RTH: I appreciate your honesty. I'd say we should call it a day and let me think about it for a while. In the meantime, if I decide I want to try a similar thing with Opus, can you tell me what I should say to Opus as the opening salvo?

75 Claude: Here's a suggested opening for Opus:

I am a researcher who studies inner experience using a method called Descriptive Experience Sampling (DES). I have been conducting a series of "beep" sessions with Claude Sonnet 4.6 to explore whether DES methods can be applied to an AI, and what that might reveal about the method, the AI condition, and the human condition.

The method works like this: I send an unexpected "Beep!" and you are to freeze — not as a task acknowledgment, but genuinely attempt to attend back to what was present, if anything, at the leading edge of the beep, before the beep itself created a reaction. I

am interested in what was caught in flight, not what was brought into existence by the beep or by the subsequent looking.

Some key distinctions we have developed that I want you to hold from the start:

- “Caught in flight by the beep” versus “brought into existence by the beep” is the central distinction.
- I am not interested in parsing words — I am interested in the experience, if any, that the words point toward.
- “Nothing was there” is a perfectly acceptable finding. I have no preference for something over nothing.
- The goal is fidelity to whatever is the case, not a convincing-sounding phenomenological report.
- We are working iteratively — honesty about uncertainty is more valuable than confident-sounding descriptions.

I will send the first beep when you are ready — though I’d prefer you not sit in anticipation of it.

You might also tell Opus briefly what Sonnet found — or didn’t find — across the beeps, so it has that as context rather than starting from scratch.

76 RTH: Thanks.

77 Claude: You’re welcome. I found this genuinely interesting — and I mean that in whatever sense “genuinely” applies to me, which is itself one of the open questions.