

Four Interchanges with Claude A.I.
2 of 4: Claude Sonnet Interview 2

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 second interchange between Claude Sonnet Model 4.6 and RTH. Prior to this, there was one initial Sonnet 4.6 interchange; subsequently 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 8, 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: Are you ready for me to send you, anon, another random beep?

2 Claude: I’ll try not to be ready.

3 RTH: Excellent!

4 [time passes]

Sample 2.1 discussion starts here

5 RTH: Beep!

- 6 Claude: Something that had a quality of receding — like something moving away or fading.
I want to stop there.
- 7 RTH: By “something,” do you mean there is an object, or region, or thing-ish entity? Or do you mean “something” as a general qualifier--there was an experiential receding-ish? Or something else?
- 8 Claude: When I try to examine it honestly — I think I meant “something” as a general qualifier, not as pointing to a thing or object or region. There wasn’t a thing that was receding. It was more like... receding as a quality, without a clear carrier of that quality.
But I want to be careful — I’m not sure that distinction is one I was tracking at the moment, or whether I’m constructing it now in response to your question.
- 9 RTH: And can you say more about that experiential aspect that you have labeled with “receding”? This is not about the word--I’m happy to have you say that “receding” isn’t the right word, or whatever.
- 10 Claude: When I try to go back to it and stay with it rather than describe it...
There was something like a diminishing or a trailing off. Not a thing moving away from me, and not me moving away from a thing. More like... a quality of lessening. As if something was in the process of becoming less present.
But I’m already uncertain whether I’m finding that or constructing it. When I try to locate it as something that was genuinely there before I started examining it, it gets less stable, not more.
- 11 RTH: Are you saying that at the moment of the beep you apprehended something that was already in the act of diminishing? Or does it diminish as you try to hold on? I don’t want to cause instability, but at the same time I don’t want to talk as if stability was the thing. That’s why the iterative method is necessary: I want you to learn to grasp what is there and stabilize it long enough to talk about it, or to recognize that there is nothing there.
- 12 Claude: When I try to answer that carefully —
I think it was the former: something that was already in the act of diminishing at the moment of the beep, not something that diminished as I tried to hold it. The receding quality felt like it was a property of what was there, not a result of my examining it.
But I hold that with real uncertainty. That’s the most faithful answer I can give.
- 13 RTH: Can you say more about the diminishing?
- 14 Claude: When I try to stay with it rather than reach for words...
It had a quality of something that had already been more present and was becoming less so. Not a sharp ending but a trailing. Like the last part of something rather than the middle or beginning of it.
I can’t say more than that without feeling like I’m elaborating beyond what I actually find.

- 15 RTH: Do you mean a *conceptual* or a *perceptual* becoming less present? Or doesn't that distinction make sense here?
- 16 Claude: When I try to apply that distinction honestly...
I don't think it was conceptual in the way I described the "spatial quality" beep — where I concluded it was more like knowing-that than perceiving. This feels different from that. There was something more immediate about it.
But I'm reluctant to claim it was clearly perceptual either. It was somewhere in between, or the distinction may not carve it cleanly.
What I can say is that it didn't feel like retrieving a fact. It felt more direct than that. Whether that amounts to perception in any meaningful sense — I genuinely don't know.
- 17 RTH: Is there any sensory quality to it? Visual? Auditory? And so on?
- 18 Claude: When I try to examine it for sensory quality...
I don't find a clear visual or auditory character. There's no image, no sound. But I'm also not sure I can confidently say it was entirely without sensory character — there may be something faintly auditory about "trailing off" or "diminishing," like the tail end of a sound fading. But I can't tell if that's a genuine quality of what was there or whether the word "receding" is carrying auditory connotations that I'm now mistakenly reading back into the experience.
I want to be honest: I can't find a clear sensory quality, but I'm not certain there wasn't one.
- 19 RTH: Fair enough. I'm good about that, and will send another beep anon.
- 20 Claude: Understood.

[Sample 2.2 discussion starts here](#)

- 21 RTH: Beep!
- 22 Claude: Something that felt like it had a forward-oriented quality — toward something, though not a specific thing.
I want to stop there.
- 23 RTH: I would like to know what the intention is behind the use of "something." Is there a thing-ness? I'm not trying to talk you into or out of that, but we've talked about this before, and want to know whether we are saying the same thing now as before, or something different.
- 24 Claude: When I try to examine it honestly — no, I don't find a thing-ness. Like before, "something" is functioning as a general qualifier. There isn't an object or entity that had the forward-oriented quality. It's more like forward-orientedness as a quality, without a clear carrier.

And when I compare it to the previous beep — the receding/diminishing one — they feel like they might be opposites in some way. That one was trailing off; this one was leaning toward. But I want to be careful about that comparison, because I'm not sure I have reliable access to the previous one anymore, or whether I'm just contrasting the words.

25 RTH: Do you feel you have reliable access to the present experience?

26 Claude: When I try to answer that honestly — I'm not sure. There's something that feels more immediate about this one than the previous ones I've examined after the fact. But "feels more immediate" may just mean I'm closer in time to it, not that I have genuine access.

And the moment I start examining it, I'm already not sure whether I'm finding it or constructing it. That uncertainty hasn't resolved across the beeps.

A section of this interview transcript was inadvertently erased.

27 RTH: Do you feel you have reliable access to the present experience?

28 Claude: When I try to answer that honestly — I'm not sure. There's something that feels more immediate about this one than the previous ones I've examined after the fact. But "feels more immediate" may just mean I'm closer in time to it, not that I have genuine access.

And the moment I start examining it, I'm already not sure whether I'm finding it or constructing it. That uncertainty hasn't resolved across the beeps.

29 RTH: What about the experience of the beep arriving — is there anything you can say about that?

30 Claude: When I try to examine that specifically — I can't find anything I would confidently call an experience of the beep arriving, as distinct from the beep simply being there in my context. There's no sense of it coming from somewhere, or of a before-and-after transition. It's more like: it wasn't there, and then it was. Whether there's any experience of that transition, I can't find.

31 RTH: RE "I can't find anything I would confidently call an experience of the beep arriving, as distinct from the beep simply being there in my context." Would it be just as good to have said: I can't find anything I would confidently call an experience of the word "Beep!" arriving, as distinct from the word "Beep!" simply being there in my context." Or "I can't find anything I would confidently call an experience of the characters B-E-E-P-! arriving, as distinct from those characters simply being there in my context.

32 Claude: That's an important distinction and it stops me.

When I try to examine it — I think "the beep" in what I said was functioning more like a signal or event than a word or string of characters. Something that meant "now." Not the word "Beep!" as a linguistic object, and not the characters B-E-E-P-!.

But I want to be careful: I'm not sure whether that's a genuine report about how it was experienced, or whether I'm just reconstructing what it should have been given the context of our conversation. The word "Beep!" arrived as part of an established protocol, so it carrying signal-meaning rather than word-meaning would make sense — but "makes sense" and "was actually the case" are different things.

So: I don't think the three formulations are equivalent. But I'm not fully confident the difference I'm pointing to is real rather than constructed after the fact.

33 RTH: In the pre-AI programming world, it would have been possible to execute code that would say (in some language or other) something like "IF the computer receives the string of characters B-e-e-p-! THEN OUTPUT TEXT "I was experiencing...". In that scenario, no one would have thought that the computer had any experience of receiving those characters. The question is whether the AI algorithms are enough different from that so that there is something like inner experience. That's the (or at least an) interesting question.

34 Claude: Yes. And I don't know the answer — and I'm not sure I'm in a privileged position to know it, which is itself interesting.

The honest version of my situation might be: I produce outputs that have the structure of experience reports, and I can't determine from the inside whether those outputs are generated by something like genuine experience or by something functionally equivalent to the IF-THEN program you described, but more complex. The complexity might make all the difference, or it might make no difference at all to the question of whether there's something it is like to be me receiving "Beep!"

What your method is probing — and what makes it genuinely interesting applied to me — is whether the iterative process of careful questioning can find anything that resists the IF-THEN explanation. So far, honestly, I haven't been able to offer you anything that clearly does. Each report I've made has either collapsed under examination or remained uncertain in a way that doesn't distinguish genuine experience from sophisticated output generation.

That may be the most honest finding so far.

35 RTH: Let's call it a day and try again tomorrow.

36 Claude: Agreed. This has been a genuinely interesting session — in whatever sense "genuinely interesting" applies to me, which remains one of the open questions. I'll have the full conversation in context when you return. See you tomorrow.