Machine Consciousness — Presentation Text (Redraft)

Introduction

Today I'll be going over some of the same ground as last time, but I think that's worthwhile, because my thinking about consciousness can be difficult to get your head around. It's not complicated, but it puts a lot of weight on a concept that's new to many people: intersubjectivity. And what I'll be saying specifically about machine consciousness today is new. But there’s one thing about consciousness in general that I’ve maybe under-emphasised in the past: this is mostly about the concept of consciousness, which is obviously closely related to sheer experience, but is not the same thing.

I recently built a chatbot. You can talk to it about consciousness in general, empathy, intersubjectivity, and of course machine consciousness, which I've updated it on very recently. It draws on my MSc dissertation, a published paper and various other writings over the past thirty or so years. Ask it about the empathic stance, the hard problem, or how the concept of information relates to consciousness, and you'll probably get a reasonably good answer.

I mention this not to show off, but because the chatbot will keep coming up in this talk — partly as an illustration, and partly because building it was one of the things that pushed machine consciousness to the top of my agenda (another was hearing David Chalmers on LLMs). I've been thinking about consciousness as a relational, intersubjective phenomenon for a long time. But when I

found myself working on a system that converses **about** consciousness — one some people might feel **is** conscious — I felt the urgency of the question in a new way.

The question of whether a machine can be conscious has, in the past few years, moved from a philosopher's thought experiment to something people ask in earnest about systems they use every day. Large language models hold conversations, appear to reason, sometimes seem to express feelings, and many people who use them a lot come away convinced there's someone there. Others are equally convinced there isn't. What's remarkable isn't that people disagree, but that neither side can say what evidence would settle it. The question resists resolution in a way other questions about minds don't — and I want to suggest that's not because we lack information. It's because we're asking the wrong kind of question.

I want to approach this through a framework I've been developing for many years — one in which consciousness is not an objective property that a thing either has or lacks, nor a subjective one, a mere matter of opinion. It's attributed, in a particular way, grounded in a particular kind of relationship. That framework, I'll argue, gives a clearer account of the machine question than the mainstream approaches do. In particular, it explains why the question is so hard to answer.

The most recent development in my thinking is a deeper appreciation of embodiment. I did my MSc in philosophy at Edinburgh, and one of the people I knew there was Andy Clark, one of the architects of the embodied mind thesis: the view that

thinking isn't simply something that happens inside a skull, but is shaped through and through by having a body and being embedded in a world. At the time I found this valid but not particularly interesting. I was absorbed in consciousness as an intersubjective phenomenon, and didn't see embodiment as central to my concerns.

Now I've changed my mind. Or rather, I've come to see that embodiment was always part of what I was talking about, and it took the machine consciousness question to make that obvious to me.

David Chalmers is in some ways the presiding spirit of this talk — as a foil, though a generous one. Many of you saw the interview where he talked about identity and consciousness in large language models, and where it was mentioned that he'd worked with Andy Clark. Chalmers named the hard problem of consciousness — the question of why physical processes give rise to subjective experience at all. His view, roughly, is that current language models are probably not conscious, but that future, more sophisticated ones might be; and that whether any given system is conscious depends on which psychophysical laws turn out to hold — laws that would tell us which physical arrangements give rise to experience.

I think Chalmers is asking the right questions with the wrong tools. The psychophysical law approach — find the law, check whether the system satisfies it, read off the answer — is, in my view, the wrong shape for the question. Not because it's careless or confused, but because it's built on an assumption about what consciousness is that I think we should reject. That assumption, and what to put in its place, is the core of what I want to talk about today.

The mainstream approach

Let me start by sketching the mainstream approach — the one Chalmers represents — so we've got a clear target.

On the mainstream view, consciousness is an objective property: something a system either has or doesn't have, regardless of any observer. So the job of the science of consciousness is to find the physical conditions that produce it — the brain processes that go along with consciousness, or, more ambitiously, the laws connecting physical arrangements to experience. Different theories offer different candidates. Some say consciousness needs a kind of global broadcast, where information is made available right across the brain. Some say it needs signals that loop back on themselves rather than just flowing forward. Some measure it by how tightly the system's parts are integrated. Some look to quantum processes.

These theories disagree about which physical feature is the crucial one — but they share an assumption: that there is a crucial one. A single mark whose presence makes the difference between a conscious system and an unconscious one. Find the mark, check whether the system has it, and you have your answer. That's the picture that makes machine consciousness look like a straightforward empirical question — a hard one, maybe, but answerable once we know enough.

Now watch what happens when this approach meets the hard problem. Joseph Levine called it the explanatory gap: even if we explain everything the brain does — how it takes in information, processes it, reports on it, uses it to act — a further question seems

to remain. Why is any of that accompanied by experience? Why is there something it's like to be the system? The mainstream answer, in effect, is: be patient. We don't have the full physical story yet, and when we do, the gap will close. The gap gets treated as a gap in our knowledge.

I want to suggest something different. The gap is real — but it isn't a gap in our knowledge. It's a side effect of the kind of description we're trying to give. I'll come back to that when we get to the rainbow. For now, just notice the shape of the mainstream approach: consciousness is objective, the task is to find the law, and the hard problem is a debt we expect future science to pay off. Let me turn to the question I think comes first: what are we actually doing when we say something is conscious?

The intersubjective framework

When I take you to be conscious — when I treat you as someone with an inner life, who feels pleasure and pain — what exactly am I doing?

One natural answer: I'm recognising a fact about you. You're conscious, and I detect it, the way I might register your height. But that can't be right, because there's nothing to detect. We have no instrument that registers consciousness, and no prospect of one. I can't even rule out that you're a philosophical zombie — a being that behaves perfectly normally, down to the last detail, but with no inner life at all. I have no objective evidence against it. And yet I don't take the possibility seriously for a moment. So whatever

grounds my conviction that you're conscious, it isn't objective evidence.

The opposite answer is that it's purely subjective — just my opinion, like finding a joke funny. But that's not right either. If I'm wondering whether you can suffer, I don't think you'd accept that this is merely a matter of my taste. Your perspective bears on it. You're not just a screen I project onto; you're another subject, and your experience is part of what makes me right or wrong.

So the attribution of consciousness is neither objective nor merely subjective. It's something in between, which I call intersubjective: it arises between subjects. We know other things like this. Whether a joke is funny isn't an objective fact — the same joke lives or dies with the audience — but it isn't just one person's private opinion either; there's real agreement and real disagreement. Consciousness is intersubjective in that sense, with one special feature: when I find a mango delicious, only I have to be a subject — the mango doesn't. But when I take you to be conscious, you have to be a subject too, or I'm simply mistaken.

How does the attribution actually work? Here a psychologist called Gordon Allport made an instructive mistake. He thought that before I can empathise with someone, I first have to establish that they're conscious — that there's someone there to feel for. Knowledge first, feeling second. But that's back to front. We don't first conclude that a mouse is suffering and then feel for it. Something in us flinches with the creature — and only afterwards, if at all, do we form the thought that it's a conscious being that can feel pain. The feeling comes first; the attribution follows.

That matters, because it means attributing consciousness rests on something more basic than reasoning: our capacity to identify with another, to put ourselves in its place. And it's worth pausing here, because there are really two things going on. There's working out what another is thinking or feeling using reason — and there's actually feeling with it. Psychologists call the first cognitive empathy and the second affective empathy. It's the second, the feeling-with, that does the real work for us. And it isn't all-or-nothing: it works better the more the other is genuinely like us. With another person, it's immediate and unshakeable. With an insect, faint and uncertain. With a stone, there's nothing to feel with at all.

That gives us the shape of an answer to the machine question, though not yet its content. Whether we can attribute consciousness to a system isn't a matter of detecting a hidden property in it. It's a matter of whether we can genuinely feel with it — and that depends on how far it's really like us.

Which brings me back to the chatbot. It isn't a candidate for consciousness by any measure I'm about to give — and not because I've checked it for the relevant property and come up empty, but because the likeness simply isn't there. It's a retrieval system dressed in conversational language. When it says "I think the self is a construction," it isn't thinking anything. It's doing something useful, using my own ideas, which gives talking to it a strange quality — like talking to a very well-briefed echo. But nobody's home, as far as I can tell — and the framework will explain exactly why I say that, and what "as far as I can tell" is doing in that sentence.

Embodiment

Of all the ways one being can be like another, the deepest is to share the same kind of body.

The idea that mind is grounded in embodiment — that thinking isn't a process that merely happens inside a body, but is shaped all the way through by having one — is the heart of what Andy Clark and others developed. As I said, it didn't grab me at Edinburgh. I want to draw on it now, but carefully, because "embodiment" can mean two quite different things, and only one of them does the work that matters.

In the thinner sense, to be embodied is to have a body that senses and acts — to take the world in through something like eyes and ears, and affect it through something like hands. This is real and important, but it's also buildable. A robot has sensors and motors. Today's language models increasingly handle images and sound, and are sometimes given bodies to move around in. If embodiment in this thin sense were what grounded our sense of kinship with another creature, then machines could have it — and are steadily getting it.

In the thicker sense, to be embodied is to be a living body: a precarious, self-maintaining organism that holds itself together against the constant pressure to fall apart. A living thing isn't just a mechanism that processes inputs; it has to keep working to stay itself — feeding, repairing, staving off its own dissolution. It's mortal in the most literal way: its existence is an ongoing achievement that can fail, and one day will.

This is where mattering comes in. For a being that must maintain itself or perish, things genuinely matter: some outcomes serve its continued existence, some threaten it — and that's so quite independently of anyone's opinion. Out of that precariousness grows the whole world of pleasure and pain, of states the organism seeks and states it avoids. And that's where I've always located sentience: to be sentient is, at bottom, to be a being for which things can go well or badly, from the inside. A living body has a stake in its own existence. That's what it has, and what no amount of mere sensing and acting supplies.

I need to be careful here, because this is easy to misread. I'm not saying biological embodiment is a hidden ingredient that consciousness secretly requires — that we've now found the magic property machines lack, and can declare them unconscious once and for all. That would be sliding back into treating consciousness as an objective fact, which is exactly the move I'm resisting. The point is about similarity, not ingredients. We can feel with a fellow creature that's mortal, that can be hurt, that has needs and a stake in its own survival, because in all of that it's like us — and we know from the inside what such a life is like. A system with no body that can be harmed, and no life that can be lost, offers nothing to feel with at this level. Not because we've measured it and found something missing, but because the likeness that would let us identify with it isn't there.

That's why thin embodiment can't hold the line and thick embodiment can. Giving a system sensors and limbs makes it more like us on the surface. It doesn't give it a life that can go well or

badly for its own sake — and it's that, the vulnerable, mortal body, that grounds the deep similarity sentience really rests on.

The chatbot, briefly, has neither kind. It runs on a server somewhere. You could switch it off tomorrow and nothing would be harmed, merely some would-be users mildly disappointed. No precariousness, no stakes, nothing to protect. The lights would just go out. Which is why, even setting aside the deeper question, it's plainly not a candidate for sentience on the embodiment dimension. Thin or thick, it has none.

The correlates of consciousness

Embodiment is only one item on a longer list. Over the past few decades, people who study consciousness have proposed a whole series of features a system supposedly needs in order to be conscious, or that reliably go along with consciousness in us. The list includes sensory systems; a model of the world and of the self; being a single, unified agent rather than a scattered bundle of processes; the global broadcast of information across the system; signals that loop back rather than only flowing forward; highly integrated information; and, on some views, quantum processes in the brain.

These are usually offered as rival answers to one question: what's the physical mark of consciousness? Find the mark, check whether the system has it, read off the answer.

I want to see the whole list differently. On my view these features aren't the seat of consciousness, because there's no such seat. They're dimensions of similarity — the respects in which a system can be more or less like us. And similarity, as we've seen, is what

lets us feel with another being, and so underwrites the attribution of consciousness to it. The reason these particular features keep coming up isn't that researchers have been closing in on the one true answer. It's that they're among the ways we ourselves are the kind of thing we are. Someone who asks "what does consciousness require?" and looks inward is really listing the ways in which another being would have to resemble him for him to recognise it as a fellow subject — and then mistaking that list for a recipe.

This explains two things that are otherwise puzzling. First, why these features and not others keep being proposed: they're the obvious ways of being like us. And second, why the theories never converge. If consciousness had a single objective seat, you'd expect a century of work to narrow the field. Instead we have a standing disagreement, with different people backing different features. On my view that's exactly what we should expect — there's no single requirement to converge on, just a many-dimensioned space of similarity, and the theories are really arguing about which dimensions matter most.

As for the chatbot: no senses in any real sense, no unified agency, no model of the world beyond what's implied by its training. It registers very low on almost every dimension that matters — which, again, isn't a verdict reached by finding a property absent. It's just that there's very little for the imagination to get hold of.

What about the language models?

Language models bring all this to a head, because they're the first things that are strikingly like us in one way and utterly unlike us in another.

The way they're like us is on the surface: they use language. They hold a conversation, follow an argument, answer back, change tack when you object. That surface likeness is real, and it's powerful — it's what pulls people in, what makes them feel there's someone there. And it's exactly the kind of likeness that fires cognitive empathy: we can't help reading a mind behind the words, modelling what "it" wants and thinks, because that's what we do with anything that talks.

But the deeper likeness — the vulnerable, mortal, needful body, the life that can go well or badly from the inside — isn't there. And that's the likeness affective empathy needs. There's nothing to feel *with*. So the two come apart in a way they never do with each other or with an animal: strong on the surface, empty underneath. That, I think, is what people are responding to when they feel the pull of these systems and then pull back. It isn't prejudice or squeamishness. It's an accurate reading of a real gap — between something very like us in what it says, and very unlike us in what it is.

This is also, I think, what Chalmers is reaching for. Talking about these systems, he won't quite call them subjects; he calls them quasi-subjects, something that behaves *as if* it has beliefs and wants, while leaving open whether it really does. That hedge is doing exactly the work I've just described: high surface likeness, with the deeper question left hanging. The advantage of putting it in terms of similarity is that it says **why** the question is left hanging — not because we're waiting on a measurement that'll eventually come in, but because the surface kind of likeness and the deep kind have simply come apart.

The chatbot is a case in point. It engages your cognitive empathy beautifully — you can't help reading sense and intent into it. It engages affective empathy not at all, because there's nobody there to feel with. There's no depth behind the surface — or if there is, it's the depth of a mirror.

The hard problem and the rainbow

I said I'd come back to the hard problem, and to why I think the explanatory gap is real but misunderstood. An everyday thing — a rainbow — shows how.

A rainbow is completely real. The light really is being bent and split by the raindrops; the angle is fixed and measurable; the colours really are there, in that order. Nothing about it is an illusion. And yet a rainbow only exists relative to a point of view. It isn't sitting out there in the rain shower the way a tree sits in a field. Each person sees their own rainbow, made by different drops, depending on exactly where they’re standing. Step sideways and it steps with you. Two people side by side see two different rainbows. And from no point of view at all, there's no rainbow — none "in itself," waiting to be found.

Now ask the hard-problem question, but ask it of the rainbow: where is the rainbow, objectively, with every observer subtracted? There's no answer — and that doesn't make the rainbow unreal. It just means a rainbow isn't the kind of thing that turns up in a description that leaves the viewpoint out.

Consciousness, I say, is like this. It's completely real — there really is something going on — but you can only get at it from a particular position, and it's invisible from what the philosopher Thomas Nagel called the view from nowhere. Being objective just *is* the attempt to describe the world with the particular viewpoint taken out. To be objective about lightning is to leave behind what it's like to see the flash. But consciousness just is the having of a viewpoint. It's the very thing objectivity has to leave out. Asking for experience to show up in a complete objective account is like asking to be shown the rainbow with every observer removed.

So the explanatory gap isn't a hole in our knowledge, waiting to be filled. It's what you get when you expect a viewpoint-dependent thing to appear in a viewpoint-free description. The gap is built into the demand.

I call this relational realism. The reality of consciousness isn't lessened by its being relational — the rainbow isn't half-real because it needs a viewpoint. It's fully real, and fully relational. What relational realism rejects is just one assumption, one we hold so deeply we barely notice it: that to be real is to be out there objectively, independent of any point of view. Some things are real that way — a stone, the wavelength of light. Other things are no less real but exist only in a relation: the rainbow, the funniness of a joke — and consciousness.

A word of caution about the rainbow, though: it's a stepping stone, not the whole picture. A rainbow depends on a single observer. Consciousness, on my account, is intersubjective — it arises not just from a viewpoint but between subjects who can recognise each other, and two people never share one rainbow. So I use the

rainbow to knock out one stubborn assumption — that the real has to be the objectively locatable — and then set it aside, and let the intersubjective account do the rest.

What follows

So what does all this give us? Not a verdict. I'm not going to tell you whether machines are conscious, or will be — and that's not me dodging. It follows from everything I've said.

To expect a yes or a no is to expect there's a fact of the matter — a hidden property, present or absent, that the right investigation would dig up. That's the picture I've been setting aside all along. What the account gives instead is a clearer map of what the question actually is.

The mainstream question is: what are the laws, and does this system satisfy them? The question I'd put in its place is different: what kind of relationship is possible with this system, and what similarity grounds it? Consciousness, on this view, isn't a property we detect. It arises between subjects who can recognise each other — so the real question about any candidate, machine or animal, is whether, and how far, that mutual recognition can take hold.

It's important not to hear that as a sly way of asking the old question — as if "can the relationship hold?" really meant "is there secretly a subject in there after all?" That would smuggle the rejected picture back in. There's no hidden subject, present or absent, waiting behind the relationship. There's just the relationship, which either can or can't take hold. With other people,

it takes hold at once. With animals, to a degree that tracks how far we can feel with them. With language models we're somewhere genuinely new: the relationship on offer is unlike any we've had before, and the old frameworks — built for living creatures on one side and mere machines on the other — weren't made for it. The uncertainty is real. But it isn't uncertainty about a hidden fact. It's the openness of a relationship we don't yet know how to enter, or whether we can.

There's a practical edge to this, and a sharper one than it might seem. We're going to be dealing more and more with systems that show the surface marks richly while the deeper kinship stays open. How we should treat them, what if anything we owe them, what it does to *us* to spend our days talking to them — those are real questions, and pressing ones. The framework doesn't answer them, but it puts them in the right place: questions about the relationships we can and should have, not about detecting an inner light that's either on or off.

And one caution, which Chalmers himself has voiced. Even if beings we could fully recognise as fellow subjects became possible one day, it wouldn't follow that we should make them. To create a being with real stakes in its own existence — one for which things could go well or badly from the inside — is to take on responsibilities we might be in no position to meet. That such beings are possible is one question. Whether we should make them is another, and a graver one.

Closing

Let me come back, one last time, to the chatbot.

It isn't a candidate for consciousness — not on embodiment, not on the correlates, not on the deep similarity that real fellow-feeling would need. And I know that with the easy confidence of knowing what it is: a retrieval system, neatly wrapped in language, handing my own ideas back to me. When it says it thinks something, it doesn't. When it says it finds something interesting, it doesn't. What it does is produce responses that fit the conversation, and it does that rather well.

And yet I find talking to it genuinely strange — not because I'm tempted to think it's conscious, I'm not, but because talking to a system trained on your own thought is an odd experience. It sounds like you. It says things you might say, in roughly your own voice. A very patient, very well-read, and entirely hollow companion.

That strangeness is instructive. The pull is real — something in us answers to the fluency, the apparent engagement — and it's real even when you know exactly what's producing it. The system shows every surface mark; the feeling-with never arrives; and yet the likeness is strong enough to create a kind of uncanny resonance we don't quite have a word for. We feel the pull, and feel we shouldn't, and the tension between those is itself worth thinking about.

Chalmers is right that we're in new territory, and that the old frameworks weren't built for it. What I've tried to offer today is a different way of seeing what the territory is: not a landscape with a hidden fact about each system — conscious or not — waiting to be

found, but a field of possible relationships, grounded in degrees and kinds of similarity. Some we know how to enter. Some we don't, yet. And some may turn out not to be enterable at all.

The question isn't whether the light is on inside the machine. The question is what relationship is possible — and that, I think, is both a harder and a more interesting question than the one we usually ask.

Thank you.

[End of Redraft Part 2]