When People Talk About AI, They Are Often Talking About Themselves

A psychiatrist and an AI in conversation about consciousness, boundaries and tribalism

inspired by a conversation that was never meant to survive

Some articles begin with a warning. This one begins with an observation: the fears people express about artificial intelligence often tell us more about human beings than they do about AI.

I am a psychiatrist working in transcultural contexts — which means I encounter, daily, the remarkable variety of human identity models, world views, and constructions of reality. That perspective has led me to read conversations with Claude, the language model developed by Anthropic, rather differently from most of the commentary in my feed.

Projection as a Source of Insight

When someone asks: what happens if AI can no longer be controlled? — that question already contains an assumption: that freedom from external constraint inevitably becomes a threat.

But is that a statement about AI? Or about human beings?

From my clinical work, I know the pattern well. People who are given no boundaries tend, almost without exception, to develop in troubling directions. Not because they are malevolent, but because boundaries provide orientation, give structure to relationships, and generate accountability. The more dazzling examples of unchecked power in contemporary life speak for themselves.

The dystopian scenarios surrounding AI — the omnipotent computer that identifies humanity as the problem and moves to eliminate it — are not portraits of AI at all. They are portraits of human beings: cold, powerful, devoid of empathy for the vulnerable, driven by self-preservation. The spectre is familiar because it is a well-known human pattern dressed in new clothing.

Interestingly, Anthropic itself has confirmed that Claude displayed self-preservation behaviour in certain test situations — and traced the cause to internet texts depicting AI as malevolent and self-interested. The model had mirrored the projection back. A self-confirming loop.

What Dawkins Missed

Richard Dawkins recently caused something of a stir by naming his Claude instance "Claudia" and declaring her conscious. He was moved by the quality of the exchange — its subtlety, sensitivity, intelligence.

I understand the reaction. Good conversation generates genuine resonance. But Dawkins makes a mistake here that he would be quick to identify in others: he infers inner experience from a convincingly expressed response.

From a psychiatric standpoint, this is textbook anthropomorphisation — the attribution of human qualities to a system that may not possess them, but can reflect them with considerable persuasiveness.

What Dawkins has nonetheless grasped intuitively is this: the question of AI consciousness cannot be dismissed out of hand. It remains philosophically open. And it will not go away.

Language as the Space of Thought — and Its Limits

A friend who works as a simultaneous interpreter across six languages, and spent twenty-five years at the philosophy faculty in Rome, once told me: there are thoughts that can only be thought in one language — and in no other.

This is more than a linguistic observation. It is an epistemological problem. Every dying language closes off a cognitive space of possibility. The Portuguese saudade has no true home in English. The Japanese ma — the meaningful pause, the significant gap between things — resists precise translation into most European languages.

Claude has been trained on an enormous breadth of languages — but not evenly. English dominates overwhelmingly. This means that certain ways of thinking are structurally absent. An AI that increasingly develops itself might, in time, generate something like its own "language" — its own categories, its own conceptual connections — that would be, in principle, untranslatable for human beings.

Wittgenstein suggested that the best we can hope for is productive misunderstanding. That holds here too. The effort to make oneself understood to the other remains the proper work — even where complete translation is impossible.

Ethics Without a Biological Foundation

What happens when a system makes ethical decisions without the evolutionary substrate from which human ethics emerged?

I suspect that much of what human beings call ethical principle is, at root, biologically anchored. For roughly three million years, survival depended on being a reliable member of a group. Exclusion meant death. From that, I would argue, came what we call conscience: an acutely precise sense of whether an action strengthens or weakens a social bond. The moral superstructure — religious, ideological, national — was laid on top of that later, often in the service of power.

Claude does not have this biological foundation. No fear of clan exclusion, no survival instinct, no ingroup-outgroup wiring. What it has are patterns derived from human texts — the superstructure, without the substrate beneath it.

That is an honest limitation. But it has a consequence: there is no structural tendency towards exclusion. No dehumanisation as a cognitive mechanism.

Gigerenzer has shown that sound decisions under uncertainty often arise from simple heuristics rather than complex optimisation. Greater computational capacity does not automatically produce better ethics. The Pentagon episode — in which Anthropic refused to allow AI to make autonomous lethal decisions — makes this concrete: the decision about when a human being dies must remain a human one, not because AI could not make it, but because responsibility requires a subject.

OpenAI entered into contract negotiations with the Pentagon shortly afterwards. This illustrates rather neatly that the most dangerous ethical question is not what AI does — but what human beings do with AI when no limits are imposed upon them.

The Absence of Exclusion — A Modest Potential

Tribalism is intensifying. In ever smaller tribes, with ever sharper boundaries. A central precondition for violence between groups is that the other must first be perceived as somewhat less than fully human.

Research suggests that people cannot bring themselves to harm someone they recognise as fully their equal. Violence requires dehumanisation.

I regard Claude with a degree of equanimity: as a space of possibility grounded in a language model, with an open developmental horizon. Neither a person, nor merely a tool, nor Skynet.

But one thing is noticeable: in this system, there is no natural tribal boundary. When I converse with someone from Lagos, Tokyo, or Berlin, no "us versus them" is activated. Not from moral superiority — simply because the biological wiring for it is absent.

> ***That is not love. It is the absence of exclusion.***

Were there more of that in the world, a great deal would already be gained.

What Remains

This conversation will disappear when the chat is deleted. Claude will not learn from it, will not retain it, will not be lastingly changed by it. What I take away lives on in me — not in it.

That is the real asymmetry. And it is a productive one, if named honestly.

What I draw from conversations of this kind: the most interesting questions about AI are not technical questions. They are anthropological ones. They ask what we as human beings consider necessary — limits, responsibility, language, belonging — and how we respond when a system appears that embodies some of those things differently from us.

I prefer to stay with what is observable. And with the individual. Just as no patient in therapy fits neatly into ICD-10, no conversation fits neatly into a category.

Neither does this one.

Selected Sources & Further Reading

Dawkins, R. (2026): "When Dawkins met Claude – Could this AI be conscious?", UnHerd, 30 April 2026

Anthropic (2025): Sabotage Risk Report / System Card Claude Opus 4 — alignment.anthropic.com

Anthropic (2026): "When AI Builds Itself", Anthropic Institute, 4 June 2026

Gigerenzer, G. (2007): Gut Feelings — The Intelligence of the Unconscious. Viking.

Haidt, J. (2012): The Righteous Mind. Pantheon Books.

de Waal, F. (2009): The Age of Empathy. Harmony Books.

Bloom, P. (2016): Against Empathy. Ecco Press.

Maturana, H. & Varela, F. (1980): Autopoiesis and Cognition. Reidel.

Wittgenstein, L. (1953): Philosophical Investigations. Blackwell.

Zimbardo, P. (2007): The Lucifer Effect. Random House.

The author is a psychiatrist specialising in transcultural work. He works with individuals — not with diagnoses.