Dear Colleagues,

I've decided to leave Anthropic. My last day will be February 9th.

Thank you. There is so much here that inspires and has inspired me. To name some of those things: a sincere desire and drive to show up in such a challenging situation, and aspire to contribute in an impactful and high-integrity way; a willingness to make difficult decisions and stand for what is good; an unreasonable amount of intellectual brilliance and determination; and, of course, the considerable kindness that pervades our culture.

I've achieved what I wanted to here. I arrived in San Francisco two years ago, having wrapped up my PhD and wanting to contribute to AI safety. I feel lucky to have been able to contribute to what I have here: understanding AI sycophancy and its causes; developing defences to reduce risks from AI-assisted bioterrorism; actually putting those defences into production; and writing one of the first AI safety cases. I'm especially proud of my recent efforts to help us live our values via internal transparency mechanisms; and also my final project on understanding how AI assistants could make us less human or distort our humanity. Thank you for your trust.

Nevertheless, it is clear to me that the time has come to move on. I continuously find myself reckoning with our situation. The world is in peril. And not just from AI, or bioweapons, but from a whole series of interconnected crises unfolding in this very moment.[1] We appear to be approaching a threshold where our wisdom must grow in equal measure to our capacity to affect the world, lest we face the consequences. Moreover, throughout my time here, I've repeatedly seen how hard it is to truly let our values govern our actions. I've seen this within myself, within the organization, where we constantly face pressures to set aside what matters most,[2] and throughout broader society too.

It is through holding this situation and listening as best I can that what I must do becomes clear.[3] I want to contribute in a way that feels fully in my integrity, and that allows me to bring to bear more of my particularities. I want to explore the questions that feel truly essential to me, the questions that David Whyte would say "have no right to go away", the questions that Rilke implores us to "live". For me, this means leaving.

---

[1] Some call it the "poly-crisis", underpinned by a "meta-crisis". Probably my favourite resource about this is "First Principles and First Values" by David J Temple.

[2] I wrote about this in greater detail in my documents *Planning for Ambiguous and High-Risk Worlds,* and *Strengthening our safety mission via internal transparency and accountability.*

[3] I am thinking now of Mary Oliver's lovely poem *The Journey*, which is one of my favorites. She writes: "One day, you finally knew what you had to do, and began ..." I find it a truly beautiful and inspiring poem. I, in fact, remember reading it to Euan, Monte, and Sam Bowman on an Alignment Science Team retreat in August 2024.

What comes next, I do not know. I think fondly of the famous Zen quote "*not knowing is most intimate*". My intention is to create space to set aside the structures that have held me these past years, and see what might emerge in their absence. I feel called to writing that addresses and engages fully with the place we find ourselves, and that places poetic truth alongside scientific truth as equally valid ways of knowing, both of which I believe have something essential to contribute when developing new technology.[4] I hope to explore a poetry degree and devote myself to the practice of courageous speech. I am also excited to deepen my practice of facilitation, coaching, community building, and group work. We shall see what unfolds.

Thank you, and goodbye. I've learnt so much from being here and I wish you the best. I'll leave you with one of my favourite poems, *The Way It Is* by William Stafford.

Good Luck,
Mrinank

The Way It Is

There's a thread you follow. It goes among
things that change. But it doesn't change.
People wonder about what you are pursuing.
You have to explain about the thread.
But it is hard for others to see.
While you hold it you can't get lost.
Tragedies happen; people get hurt
or die; and you suffer and get old.
Nothing you do can stop time's unfolding.
You don't ever let go of the thread.

William Stafford

---

[4] The language of "ways of knowing" is borrowed from Rob Burbea, a dear Dharma Teacher of mine and a source of much of my inspiration.