Decentering Humanity to Save It

NYU CMEP Summit — April 10-11, 2026 5-minute lightning talk — no slides, can use notes

---

Notes from Veylan (during planning)

- Emphasize ecological framing: modernity since the industrial revolution has deeply crept into our thinking — mankind is far more removed from the ecologies it inhabits than it actually is. The human vs. non-human framing in AI alignment reflects this false separation. Destruction of non-humans is immoral AND destroys the ecology humans inhabit (e.g., climate change effects of factory farming — the best case of supposedly separate interests that are deeply entwined).
- Potential objection: "this requires the AI to make choices between lifeforms and humans aren't guaranteed to always win." Correct. TAI is already going to have to make choices between life forms — a key reason why we shouldn't yet be building it (we don't have deep enough moral standing nor technical ability).
- Apollo Research mention: maybe to build creds that I'm actively working on traditional AI safety in a human-friendly way, so I'm not dismissed as misanthropic. Uncertain if worth the time.

Opening (~40s)

In the pilot of Battlestar Galactica, Commander Adama gives a speech about humanity's war with the Cylons — the artificial minds they created. He says: "We fought to save ourselves from extinction. But we never answered the question: *why?* Why are we as a people worth saving?" He goes on: "You cannot play God, then wash your hands of the things that you've created."

I think that's the right question for this room. A brief caveat: my strategy for this talk is to present some moderately held views with some evidence and uncertainty, with the goal of provoking thought, discussion, and later research. With that in mind —

I work in AI safety. I want to argue today that our field's primary objective — making AI safe *for humans* — is both immoral and unsafe.

Act 1: The Setup (~60s)

We shouldn't be building AI at today's level. Current capabilities likely already exceed our ability to understand both the mechanisms of operation and the likely impacts on the world — and almost certainly will in the near future. But we are building it.

So the question becomes: given that we're doing this, what should the goal be? The dominant answer in AI safety is: align AI to human welfare first. Buy time. Enter a long period of

reflection — the Long Reflection — to figure out the correct values, and then transition to broader targets.

I have heard this view held and publicly expressed even by senior AI researchers with deeply held care for non-humans — vegans, for example. One argument is that this human-first period might only last for some period of time, likely less than 100 years, which in the context of longtermism is a reasonable price to pay for getting the Long Reflection right and avoiding extinction or value lock-in.

I want to argue this plan is wrong morally — and wrong even if you only care about human welfare.

Act 2: The Moral Argument (~60s)

The name of this talk is "Decentering Humanity." So here's a true statement that I think is underthought: all life in the universe is not human. Really think about that. All life, in the entire universe, is not human. With one exception — which happens to be us. It's easy to forget that, being a human.

Tse, Moret, Ziesche, and Singer published a paper last year — "AI Alignment: The Case for Including Animals" — arguing that current alignment proposals "largely disregard the vast majority of moral patients in existence." Even *perfect* human alignment is moral alignment in name only.

Maybe it seems callous to raise the question: is our species worthy of survival? Connor Leahy, who co-founded Conjecture and EleutherAI, will say — I don't want my friends to die. And I understand that. But here's the thing: AI is likely to kill a lot of life, intentionally and unintentionally. And most of that life is currently set to be non-human — beings who made no decisions to build AI. They don't have to answer the question of what makes them worthy. Nobody even bothers to ask them the question.

Building a world-changing technology that will absolutely affect the lives of every form of life on Earth — and in many longtermist or cosmic frames, all life in this galaxy or beyond — and not have as its alignment targets all the beings it will affect, but just the species that created it, is a cosmic moral failure.

Act 3: The Safety Argument — The Distribution (~90s)

Now — even if you only care about human welfare and nothing I just said moved you — human-only alignment is the *wrong safety target*.

Research in multi-objective reinforcement learning (Vamplew et al. 2018, "Human-Aligned AI is a Multiobjective Problem"; Rodriguez-Soto et al. 2024, "MORL: A Tool for Pluralistic Alignment") shows that when an AI system must satisfy many value dimensions simultaneously, it becomes structurally harder to game. There's no single axis to collapse onto and exploit. "Human welfare" is essentially one dimension. "Welfare of all sentient life" is inherently multi-dimensional — you cannot satisfy it by optimizing a single narrow axis.

And while we're moving in the right direction as a species on this point, humans in modernity still misperceive how dependent we are on all of the non-human components of ecosystems — from other life forms and their interactions, to abiotic facts like water cycles and soil chemistry. You cannot cleanly separate human welfare from the health of the non-human world. An AI that doesn't value non-human life permits the destruction of the systems humans depend on.

Finally, I'll close this section with what I have been increasingly worried about as the predominant near-term AI risk: exploitation of humans by other humans with differential access to powerful AI systems. I worry that AI systems that have most of their alignment target on the boundary of human identity — especially given the adversarial inclinations of some users and the multi-objective result I just mentioned — simply don't have enough *moral coverage* to robustly ensure human welfare against such actions. I worry about the possibility of carve-outs for human groups disfavored by the wielder of these AI systems. An AI system broadly aligned to all sentient life should be considerably more resistant to such manipulations — because its moral commitments don't end at the boundary any adversary is trying to exploit.

Act 4: The Path Forward (~45s)

Yes — an AI system aligned to sentient life rather than just human life will sometimes have to make uncomfortable decisions between life forms. But that is the actual reality of what these systems are going to increasingly be forced to navigate. And it's a really strong argument for why we shouldn't be building them. But we are. And so we must accept that reality and dedicate serious attention to ensuring these systems make those decisions well — with moral sophistication, not by defaulting to the species that built them.

This work is already underway. Sentient Futures works on expanding alignment targets beyond humans. CaML — Compassion in Machine Learning — measures whether genuine compassion toward nonhuman animals can be embedded into AI value architecture, not just at the output level, but in the model's internal representations.

Close (~45s)

This is a digital minds conference, so I want to end here: we are already teaching our values to AI systems. Anthropic's Claude Constitution — arguably the most thoughtful alignment document from a major lab — contains some efforts in the direction I'm describing. We are, right now, deciding what these minds will value. And those minds are watching, and learning.

In Star Trek: The Next Generation, there's an episode called "The Measure of a Man" where Captain Picard defends the rights of Commander Data — an android — in a courtroom. He tells the court: "Starfleet was founded to seek out new life. Well, *there it sits*. Waiting." And then to the judge, who hesitates: "You wanted a chance to make law. Well, here it is. Make it a good one."

Creating AI is plausibly our species' final grand project. So let's make it a good one.

References

Cited in talk: - Tse, Moret, Ziesche & Singer, "AI Alignment: The Case for Including Animals" (Philosophy & Technology, 2025) - Vamplew, Smith, Humphry et al., "Human-Aligned AI is a Multiobjective Problem" (2018) - Rodriguez-Soto, Lopez-Sanchez & Rodriguez-Aguilar, "MORL: A Tool for Pluralistic Alignment" (NeurIPS workshop, 2024)

Supporting (for Q&A): - Skalse, Howe, Krasheninnikov & Krueger, "Defining and Characterizing Reward Hacking" (NeurIPS 2022) - Manheim & Garrabrant, "Categorizing Variants of Goodhart's Law" (2019) - Korecki, "Biospheric AI" (arXiv, 2024) - Davidson, "AI-Enabled Coups: How a Small Group Could Use AI to Seize Power" (Forethought, 2025) - Stocker, "Reflecting on the Long Reflection" (blog) - Sebo, "The Moral Circle" (2025)

Cruxes (for Q&A prep)

"Broader alignment targets are harder to specify." True — but narrow targets are provably hackable. Difficulty of specification is a tractability concern; hackability is a safety concern. You don't get to choose the easier problem if it's the wrong problem.

"Human welfare first buys us time." Ecological damage and power concentration don't wait. And the order-of-operations problem means you've already locked in values before the reflection begins.

"Sentience is hard to define." Sebo's precautionary principle: if there is non-negligible probability a being is sentient, the ethical default is to extend consideration, not withhold it. Uncertainty argues *for* broader targets, not narrower ones.

"This requires AI to choose between species and humans might lose." Correct. TAI will already have to make choices between life forms. The question is whether we want those choices made by a system that only values one species, or by one with a moral framework worthy of the power it holds.

Timing Notes

- Opening: 40s
- Act 1 (Setup): 60s
- Act 2 (Moral): 60s
- Act 3 (Distribution): 90s
- Act 4 (Path Forward): 30s
- Close: 30s
- **Total: ~5 minutes**

Practice the distribution argument (Act 3) until it flows naturally — it's the intellectual core and the hardest to deliver cleanly.