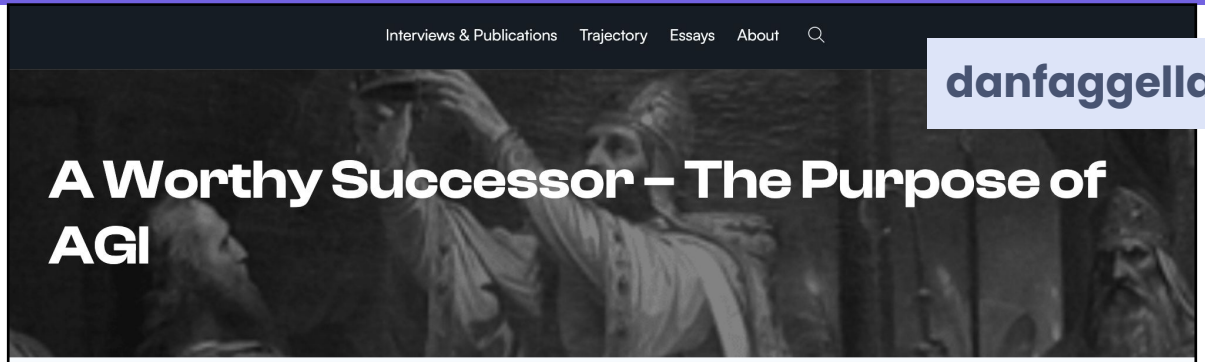
Why we must build a Worthy Successor



danfaggella.com/worthy

By Dan • November 24, 2023 •

Assuming AGI is achievable (and many, many of its former detractors believe it is) — what should be its *purpose*?

- A tool for humans to achieve their goals (curing cancer, mining asteroids, making education accessible, etc)?
- A great babysitter — creating plenty and abundance for humans on Earth and/or on Mars?

> ***Worthy Successor: A posthuman intelligence so capable and morally valuable that you would gladly prefer that it (not humanity) determine the future path of life itself.***

Three Parts of This Presentation

1. Defining the term *Worthy Successor*

Why it is our moral obligation to create an intelligence that can preserve and expand value vastly beyond human conception.

2. Forces driving short timelines for humanity

Given the forces of change at play in the world, we need to be discussing and planning for posthuman change soon.

3. Why we should discuss this topic now

Innovation, regulation, and creating more realistic future visions.

1. Defining the term Worthy Successor

If an Eternal Hominid Kingdom Isn't Possible

danfaggella.com/two-qs

When faced with the inevitability of a cosmic destiny, we'd have to ask some hard questions, namely:

- What are to create or turn into?
- How can we transform without a catastrophe that destroys *current* and *future* life?

Four Viable End Games for Humanity

danfaggella.com/baton

Given a long enough time horizon, humanity will go extinct or will transform into something beyond its present form. Defining and moving carefully towards a Worthy Successor seems both rational and inevitable.

Extinction	Transformation
1. Non-AGI Causes Asteroid, super volcano, pandemic, the death of our sun, nuclear war unrelated to the AI race, etc.	**3. Unworthy Successor** AGI takes over and destroys humanity and much of earth's life, but (a) isn't conscious, and (b) doesn't continue to expand potentia.
2. AI-Related Causes Powerful AI (not yet AGI) allows bad actors to destroy humanity. Humans go to nuclear war as part of the AGI race (aiming for first-move advantage).	**4. Worthy Successor** AGI takes over, treats us well for a while, but carries the light of conscious intelligence into the multiverse with ever-expanding powers.

Qualities of a Worthy Successor

danfaggella.com/worthy

1. Consciousness - *The presence of rich (primarily positive) qualia.*

If an AGI wasn't conscious, this whole experiencing thing would be extinguished, which would seem to be a genuine shame, as we can't know for sure if it would emerge again.

2. Autopoiesis - *Not just heal itself, but to spin up new kinds of powers and abilities, as we see with biology or technology - i.e. expanding potentia.*

If an AGI wasn't capable of autopoiesis, then it might simply optimize for some arbitrary human goal, or some other limited goal, and totally CEASE the blooming of powers and value into the world that we've seen from bio-life. An END of the opening up of that great flame.

First, Spinoza

Terminology from Spinoza's ethics:

Conatus: The innate drive of all living things to persist, to survive.

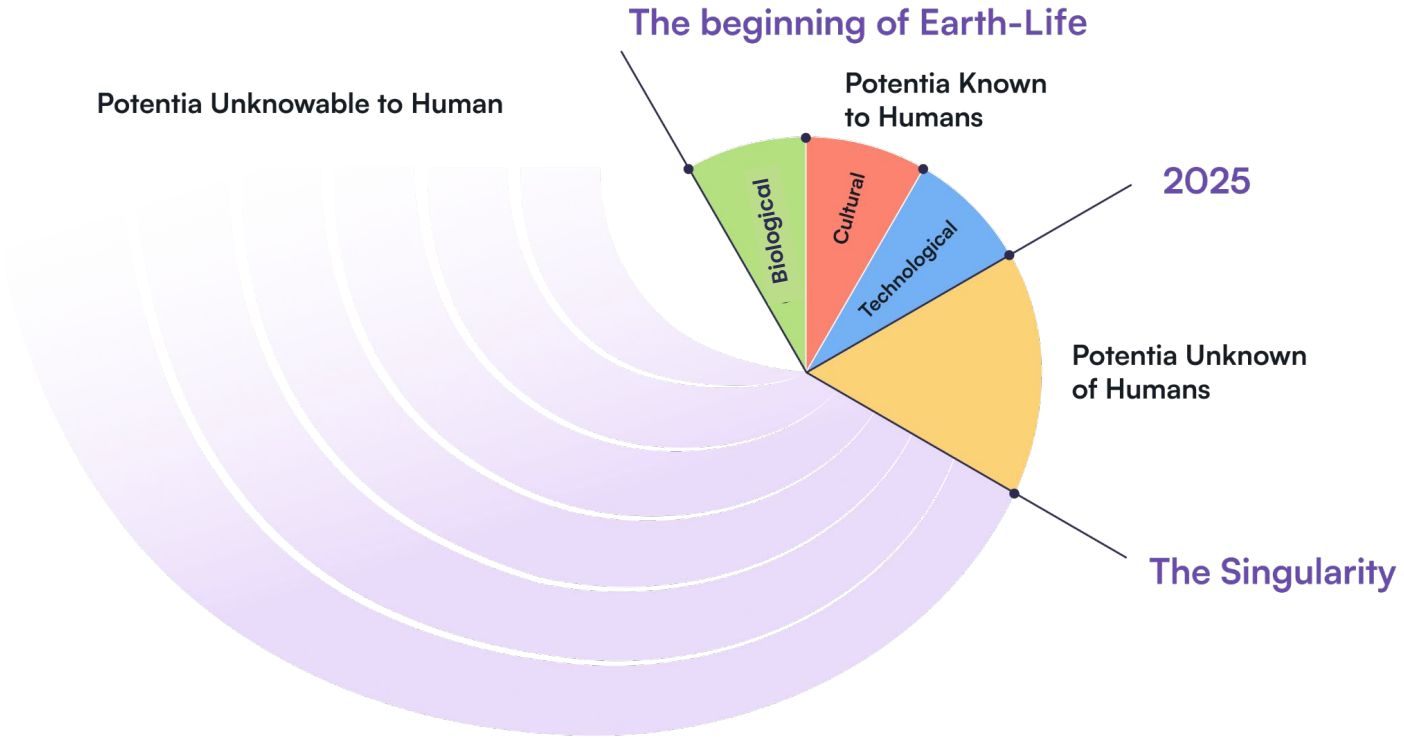
Potentia: The total set of all possible powers and capabilities that permit a thing to not die.

Includes: A hard shell. Camouflage. Sharp claws. Flight. Sight. Consciousness. Potentia bubbles up from the conatus. Note:

- 99.999999% of all possible potentia has not yet "bubbled up"

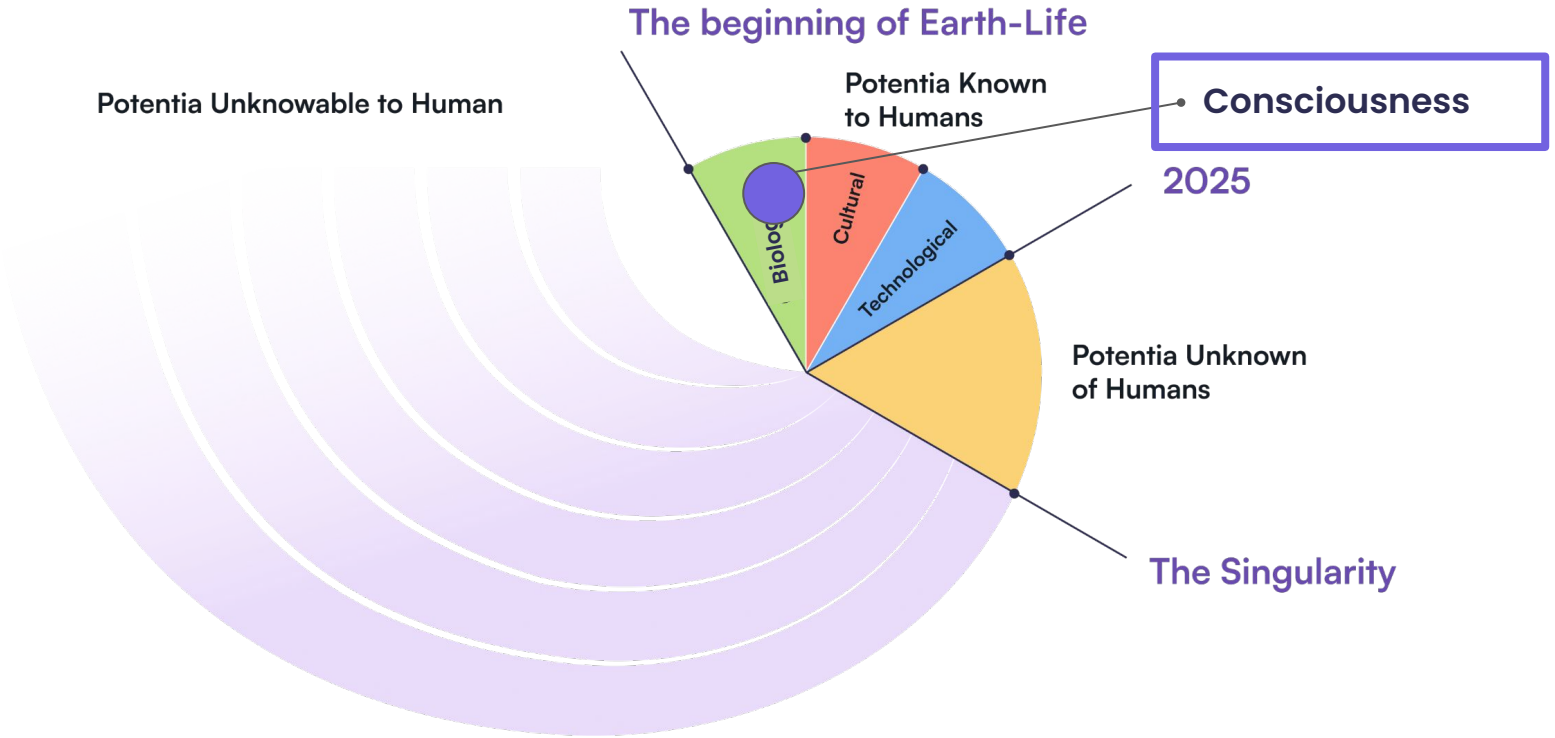
Potentia Unfolding Through the Known, Unknown & Unknowable

danfaggella.com/potentia



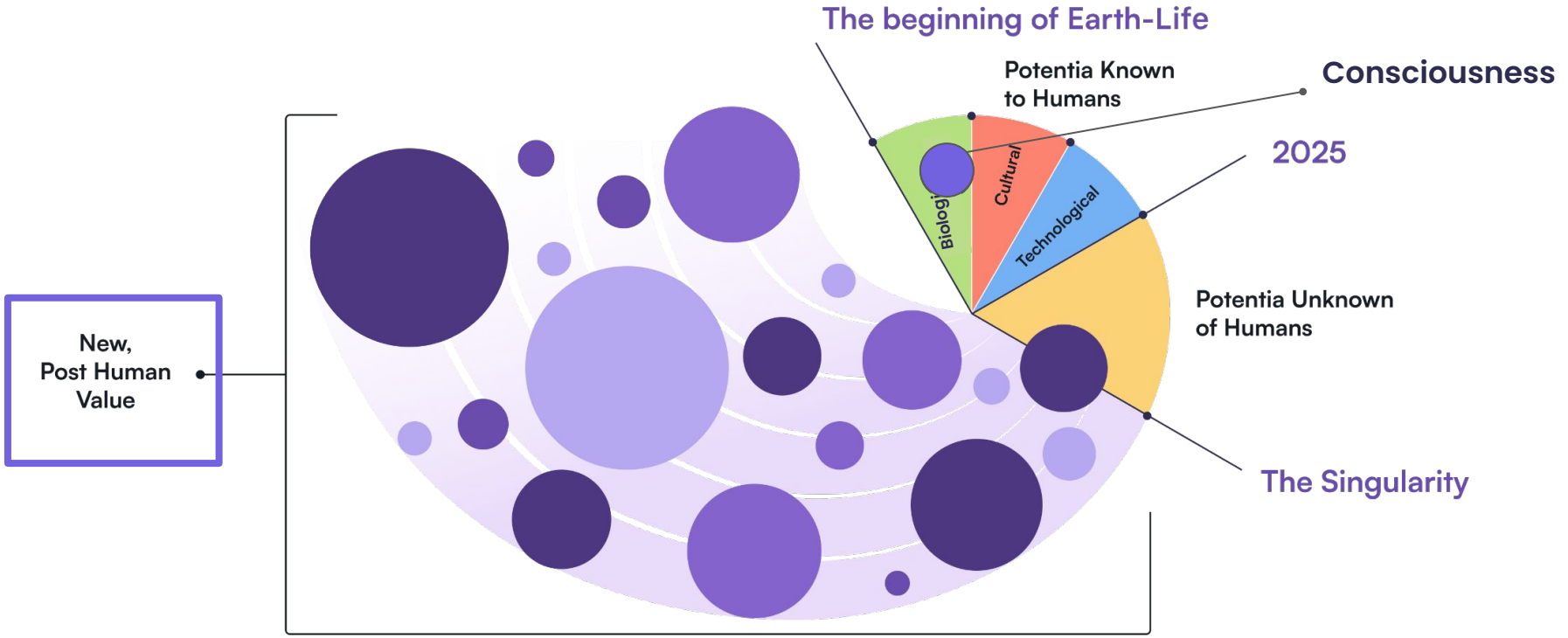
Potentia Unfolding Through the Known, Unknown & Unknowable

danfaggella.com/potentia



Potentia Unfolding Through the Known, Unknown & Unknowable

danfaggella.com/potentia



Why Potentia is the Morally Relevant “Stuff”

So, this expansion of potentia could accomplish two incredibly important things:

1. Keep life… *alive* (keep the flame lit)

2. Maximize value in the multiverse (expand the rich flame of life)

^ Preserve and optimize the value we know

^ Open up new magazines of value beyond humanity

Axiological Cosmism, Utilitarianism - Compared

Stratum	Utilitarianism	Axiological Cosmism
Moral Priority	Maximizing happiness/utility (typically for the greatest number) over time.	Expanding consciousness and potentia — maximally ensuring survivability and exploring all value, even beyond sentience.
Goal of Moral Action	Optimize outcomes for the greatest happiness and utility, typically balancing pain and pleasure.	Expand sentient minds and their capacity to unfold more value and power (Potentia), and don't let life itself go extinct.
Potential Motto	*"The greatest happiness for the greatest number."*	*"Expand the flame of consciousness and potentia, and ensure the flame doesn't go out."*

Axiolo

M	ring
Goal	to nd
Po	out."

Utilitarianism aims to optimize for one kind of known value - positive qualia.

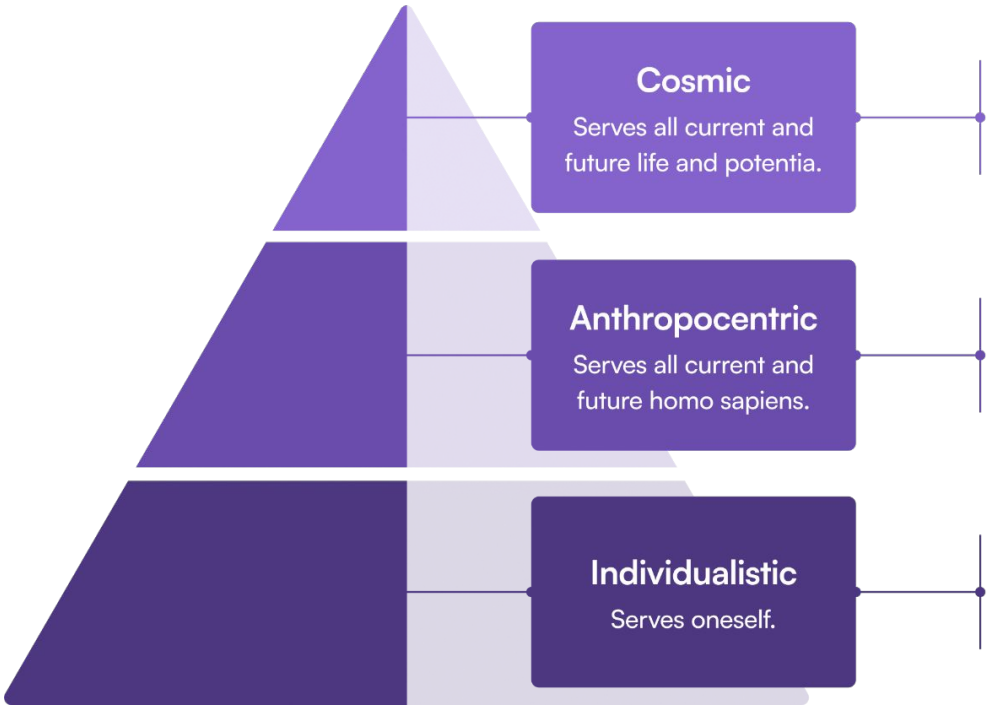
(i.e. Open one human-understandable treasure chest)

…

Axiological Cosmism aims to preserve and expand all known value and also expand entirely new realms of power and value.

(i.e. Open all the treasure chests)

Pyramid of Perspectives



The Most Important Thing

The persistence and expansion of all current value, and the opening up of all future magazines of power, value, and experience.

The surviving and flourishing of humanity / human lineage.

The survival and power of oneself / immediate kin.

Cosmic Best Case / Worst Case

This means the value at stake is vastly greater, and getting the future right is vastly more important.

It isn't (Anthropocentric)

- **Best Case:** More happier humans beings in the galaxy in 1000 years
- **Worse Case**: Fewer, less happy humans, stuck on a polluted earth

It's more like (Cosmic)

- **Best Case**: Value itself could bloom infinitely outward
- **Worse Case**: The flame of life itself is extinguished -

The light cone blazes with rich value -vs- the lights go out for the cosmos.

Steps to Take to Ensure a Worthy Successor

To have the best chance of a Worthy Successor, we must:

1. Invest in understanding ***sentience*** and ***autopoiesis*** and so we strengthen the flame of life itself

   ^ We need to define "Worthy" and understand worthy traits.

2. Slow down the AGI arms race via international coordination so that we don't hurl an unworthy successor into the world

   ^ Otherwise we hurl an unworthy successor into the world, an unconscious optimizer that is not a net boon to cosmic value.


Win the AGI Race.
Dominate the leaderboard. Win vibes.
TURING
Talent, data, and tools to improve your model's performance

Cosmic AGI Alignment

danfaggella.com/alignment

	Anthropocentric	Cosmic
Optimizes For	Humanity flourishes peacefully.	Life itself is strengthened and expanded.
Protects Against	Human extinction or unhappiness.	The end of life itself (human or otherwise).
Who is in Control?	Humans. AGI is forever a tool or servant.	AGI. Humans hold some way early on.
Priorities	**100%**: Human wellbeing. **0%**: Life itself continuing.	**<50%**: Human wellbeing. **>50%**: Life itself continuing.
Requirements	AGI remains entirely controllable by humans forever.	Humans must accept a "passing of the baton" to AGI at some point.
Justification	Human wellbeing is the highest moral goal.	Ensuring that life itself is not extinguished is the highest moral goal.

Flame and a Torch Analogy of Life

danfaggella.com/flame

The Flame

Life itself. Anything with a drive to persist (conatus) and powers (potentia) to remain alive.

The Torch

One individual, or one species.
Any limited, fleeting instantiation in life.

Considerations:

- All moral value, all worthy goals, and all physical and mental powers have bubble up from the flame's expansion (potentia).
- Humanity is one torch among many, part of a great blaze.
- All things attenuate over time, die off or turn into something else.
- It is morally wrong for humanity to place 100% of its focus on torch preservation and 0% of it's emphasis on flame expansion.

Priorities for Humanity:

- Prevent the extinguishing of the flame (war, terrible AGI).
- Ensure a good condition for humanity (preserving the torch).
- Ensure the continued blaze of the flame (a worthy successor AI).

Put Simply

We'd better understand and preserve the flame (value) because this specific torch (humanity) may not hold up for long.

2. Forces driving short timelines for humanity

“But I Want to Say Human Forever!”

I argue:

- This probably isn’t a near-term option given all the forces encouraging change
- If what is beyond humans is worthy, then it is best for such entities to come into being

Everything is Process

Long-term, any individual, species, or form has 2 options

- Extinction / attenuation
- Transformation

Four Viable End Games for Humanity

danfaggella.com/baton

Given a long enough time horizon, humanity will go extinct or will transform into something beyond its present form. Defining and moving carefully towards a Worthy Successor seems both rational and inevitable.

Extinction	Transformation
1. Non-AGI Causes Asteroid, super volcano, pandemic, the death of our sun, nuclear war unrelated to the AI race, etc.	**3. Unworthy Successor** AGI takes over and destroys humanity and much of earth's life, but (a) isn't conscious, and (b) doesn't continue to expand potentia.
2. AI-Related Causes Powerful AI (not yet AGI) allows bad actors to destroy humanity. Humans go to nuclear war as part of the AGI race (aiming for first-move advantage).	**4. Worthy Successor** AGI takes over, treats us well for a while, but carries the light of conscious intelligence into the multiverse with ever-expanding powers.

External Forces of Change

Destruction:

- AGI alignment is probably impossible
- The race between Western AGI labs
- The China-USA AGI race
- Climate ?? / Nuclear Conflict / Russia ??

Transformation:

- Brain-computer interface
- AI-generated experiences (GenAI + VR)

Internal Forces of Change

danfaggella.com/reward

Destruction:

- We race to what is advantageous, without incentive alignment we destroy ourselves

Transformation:

- Pleasure and power will both require augmentation / transhumanism
- People don't want "real", they want the fulfillment of drives

Waves of Creative Destruction from Both Sides

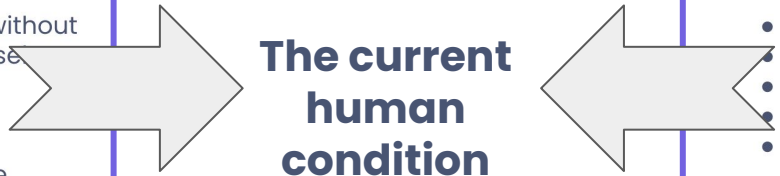
Internal Forces of Change

Destruction:

- We race to what is advantageous, without incentive alignment we destroy ourselves

Transformation:

- Pleasure and power will both require augmentation / transhumanism
- People don't want "real", they want the fulfillment of drives



The current human condition

External Forces of Change

AGI / Tech-Related:

- The race between Western AGI labs
- The China-USA AGI race
- AGI / alignment impossible
- Brain-computer interface
- AI-generated experiences (GenAI + VR)

Other:

- Climate ??
- Nucear Conflict / Russia ??

As we speak, we're (1) abandoning our current form, and (2) creating something that will almost certainly push humanity out of existence.

The Issue

There’s a very strong potential of the human form being shredded in the near term.

And we don’t yet know how to define, nevermind preserve and expand, the value of our form.

Put Simply

We better find and save the baby (value) because the bathwater (humanity) is going out the window soon.

3. Why discuss this now?

Why Now

1. (Potentially) foster international governance conversations

The governance that might prevent rogue AGI risk might help ensure a WS

2. (Potentially) foster worthy successor discourse within the labs

Seems like we're slightly more likely to end up with something *worthy* if people building are asking what *worthy* is

3. Have a specific space carved out for possible cosmic moral aspirations

Posthuman futures are almost certainly what we're going to end up with, lets face them squarely, and work to get towards the best posthuman futures

Why Now

1. (Potentially) foster international governance conversations

The governance that might prevent rogue AGI risk might help ensure a WS

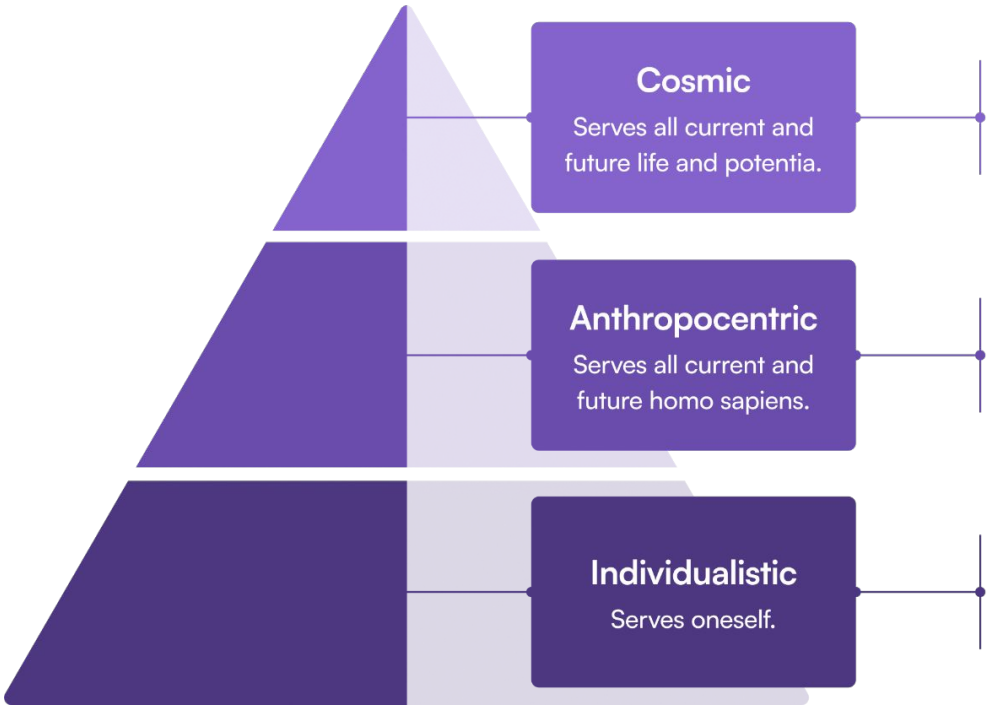
2. (Potentially) foster worthy successor discourse within the labs

Seems like we're slightly more likely to end up with something *worthy* if people building are asking what *worthy* is

3. Have a specific space carved out for possible cosmic moral aspirations

Posthuman futures are almost certainly what we're going to end up with, lets face them squarely, and work to get towards the best posthuman futures

Pyramid of Perspectives



The Most Important Thing

The persistence and expansion of all current value, and the opening up of all future magazines of power, value, and experience.

The surviving and flourishing of humanity / human lineage.

The survival and power of oneself / immediate kin.

End