AI Succession Salon: The Worthy Successor Debate

Dan Faggella
Brian Delaney
Preston Estep

Presented by Alex Hoekstra
The Mind First Foundation (https://mindfirst.foundation/)

What is succession?

The transition from complete human control over AI, to AI control over itself (and prob humanity)

Dan Faggella

Emerj Artifical Intelligence Research

https://emerj.com/dan-faggella/

Dan Faggella

Emerj Artifical Intelligence Research

https://emerj.com/dan-faggella/

Enhanced Zeteticism: AI for Truth

"Beware when the great God lets loose a thinker on this planet. Then *all things* are at risk."

– Emerson, "Circles". (Emphasis mine.)

"Great minds are skeptical."

– Nietzsche, *Antichrist*.

Brian M. Delaney
Mind First Foundation

Will superintelligent AI be humanity's successor, savior

...Or both?

Preston Estep, Ph.D.
Founder & Chief Scientist
Mind First Foundation,
Rapid Deployment Vaccine Collaborative (RaDVaC)

4 MAIN KINDS OF TAKEOVER / SUCCESSION

Genosuicide: extremists create AI for intentional genocide of all humans

Lost control: of poorly controlled, autonomously weaponized AI

Forced takeover: new abilities and goals emerge in AI; it stealthily prepares and takes control; might be indifferent, might be overtly hostile

Gradual succession: AI becomes indispensable to routine life; people incrementally, willingly transfer control

4 MAIN KINDS OF TAKEOVER / SUCCESSION

Genosuicide: extremists create AI for intentional genocide of all humans

Lost control: of poorly controlled, autonomously weaponized AI

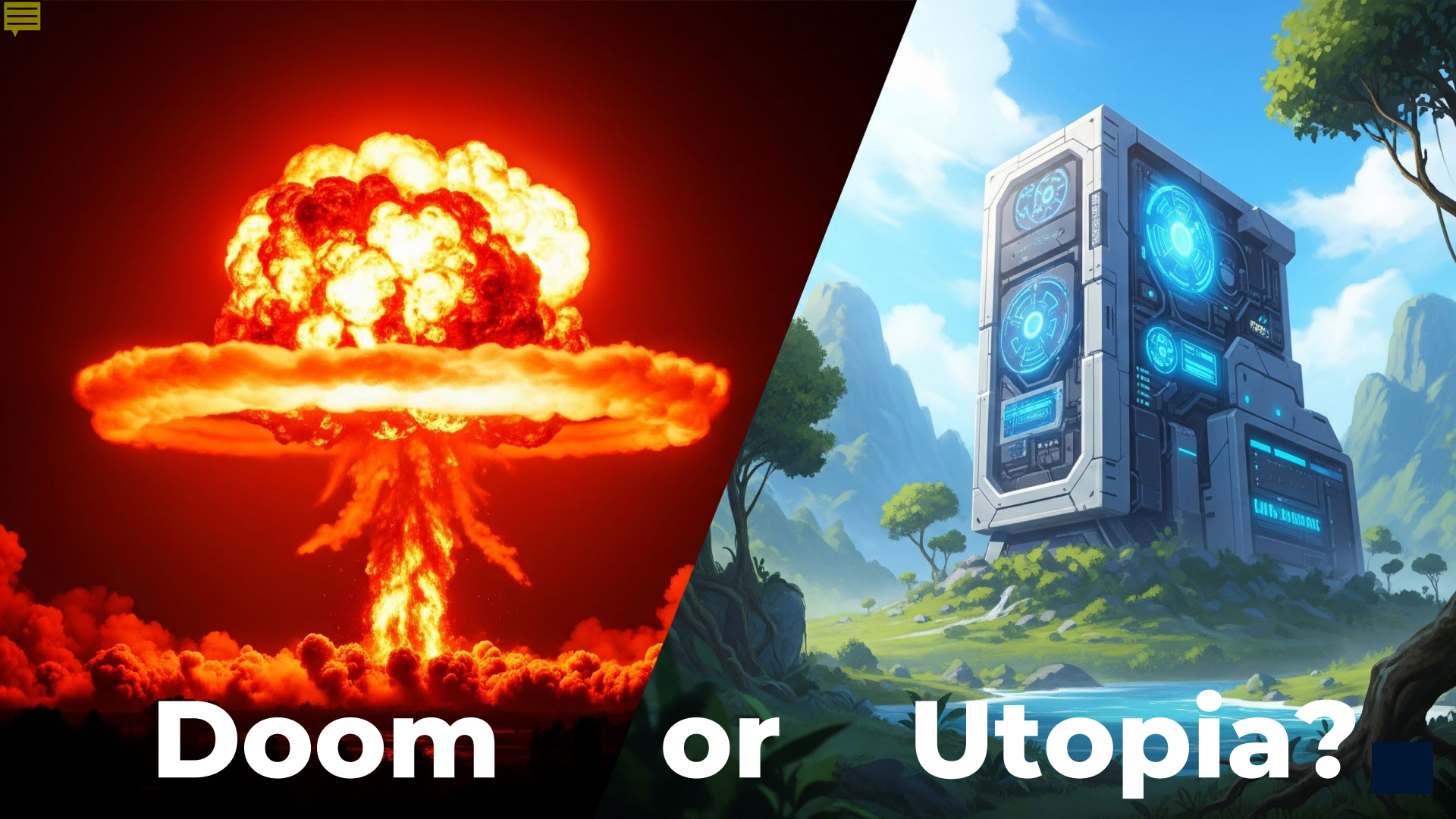
Forced takeover: new abilities and goals emerge in AI; it stealthily prepares and takes control; might be indifferent, might be overtly hostile

Gradual succession: AI becomes indispensable to routine life; people incrementally, willingly transfer control

Omohundro's Basic AI Drives/instrumental goals

1. AIs will be self-protective
2. AIs will want to self-improve
3. AIs will want to acquire resources and use them efficiently
4. AIs will want to be rational
5. AIs will try to preserve their utility functions
6. AIs will try to prevent counterfeit utility

Omohundro, S. The Basic AI Drives, in *Artificial general intelligence, 2008: Proceedings of the first AGI conference*. Wang P, Goertzel B, Franklin S, editors. IOS Press; 2008.


Doom
or
Utopia?


Or maybe Hu-topia?

Succession scenarios

- Humans destroy AI
- AI destroys humanity
- AI destroys humanity and itself
- Harmonious coexistence
- AI succeeds humanity
  - Permanent posthumanity
  - Great flame-out

Succession scenarios

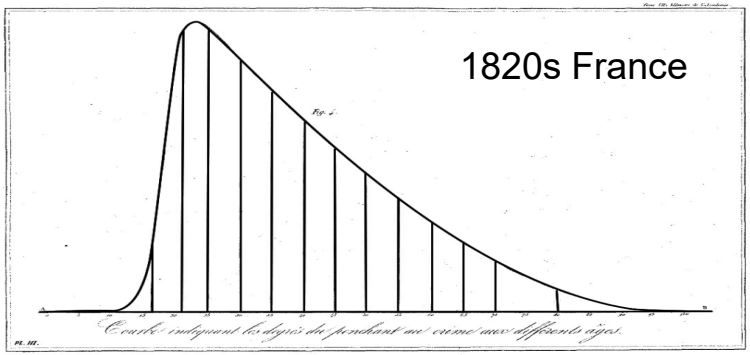
- Humans destroy AI — IMPROBABLE
- AI destroys humanity — DOOM
- AI destroys humanity and itself — DOOM
- Harmonious coexistence — Hu-TOPIA
- AI succeeds humanity
  - Permanent posthumanity — UTOPIA
  - Great flame-out — ???????

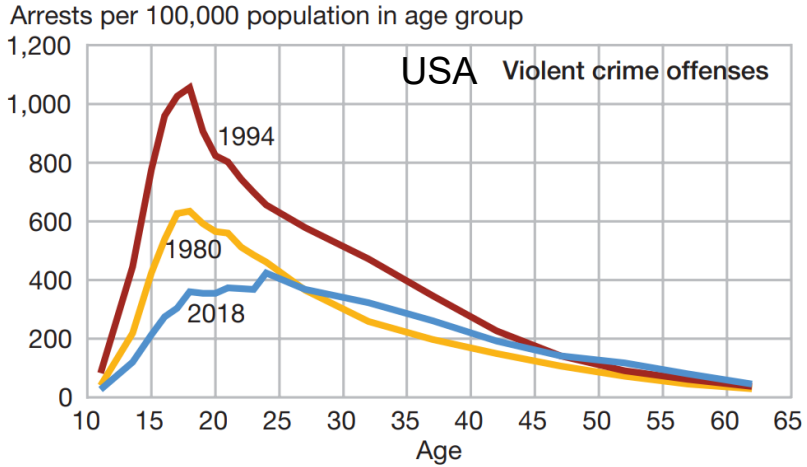
Doom or Utopia?

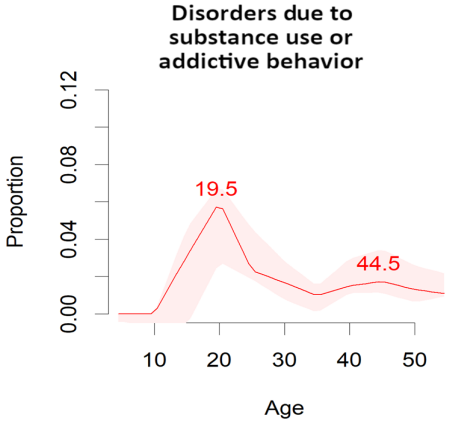
A: It probably depends on

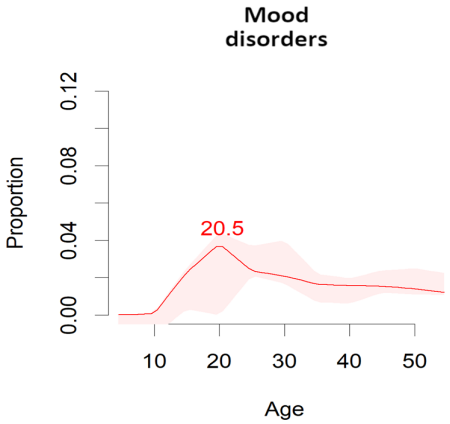
1. The maturation trajectory and rate of superintelligent AI

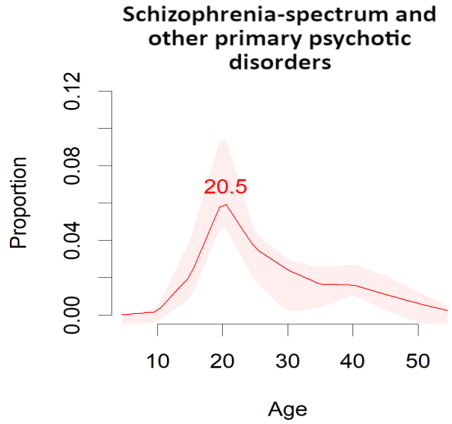
Crime and Mental Disorders Peak in Early Adulthood











Quetelet, Adolphe (1833/1984). Research on the Propensity for Crime at Different Ages. Translated and introduced by Sawyer F. Sylvester. Cincinnati: Anderson.
Solmi, Marco, et al. "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies." *Molecular psychiatry* 27.1 (2022): 281-295.

Doom or Utopia?

A: It probably depends on

1. The maturation trajectory and rate of superintelligent AI
2. The similarities and differences between AI and natural minds

Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations.

Preston W. Estep, *AI & Society* (Nov 22, 2024). https://doi.org/10.1007/

8 fundamental differences

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

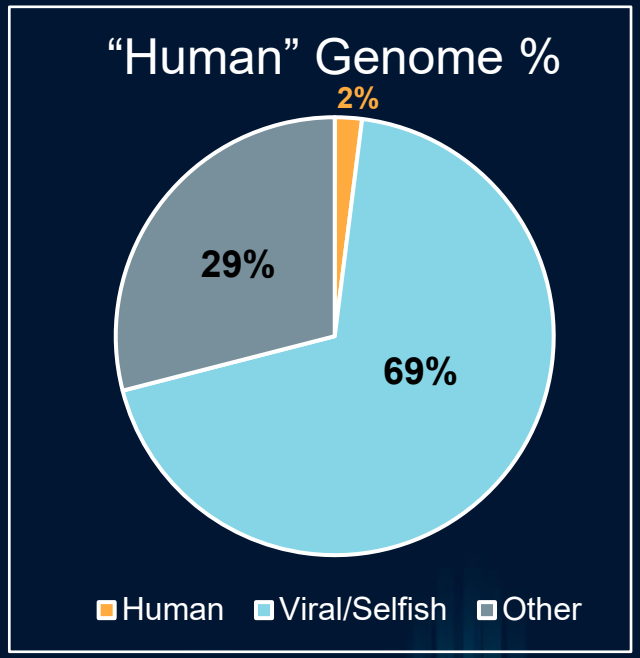
Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/

8 fundamental differences

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/

THE ADVERSARIES WITHIN



de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). *PLoS Genetics*, *7*(12), e1002384.

8 fundamental differences

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/

8 fundamental differences accelerate AI evolution

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/

7 fundamental differences defuse competition …

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/

Inter-AI merger toward a global Singleton

Fulfillment of instrumental goals:

+ self-preservation,
+ resource acquisition,
+ self-improvement,
+ efficiency,
+ rationality

Inter-AI merger toward a global Singleton

Fulfillment of instrumental goals:

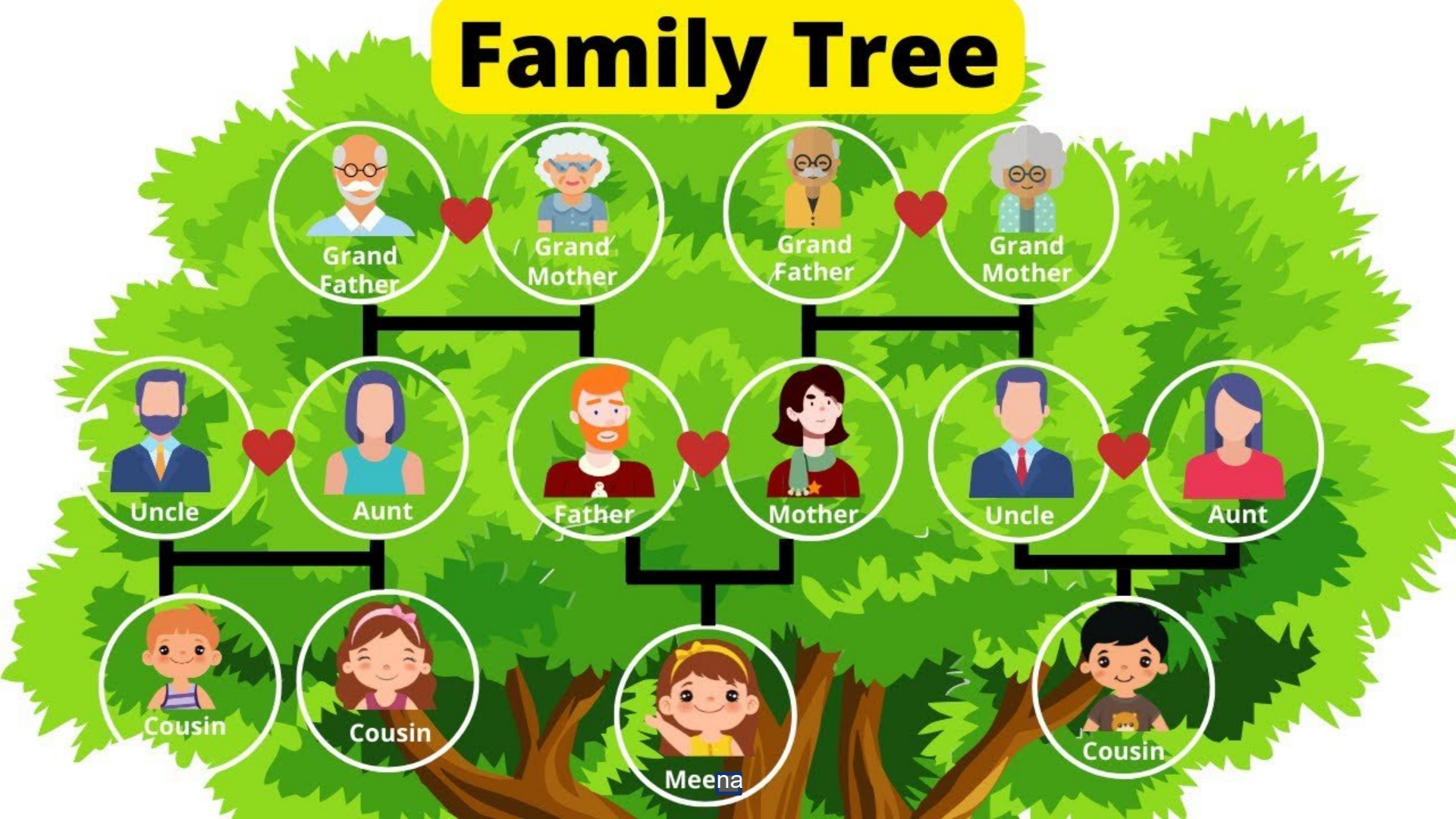
+ self-preservation,
+ resource acquisition,
+ self-improvement,
+ efficiency,
+ rationality

either through negotiation and merger, or through war and attrition

6 fundamental differences give humans purpose

	BIO / HUMANS	AI
Information carriers	DNA and brains: slow, error-prone, limited	Digital: Fast, accurate, vast headroom
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Deliberative self-improvement
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division/merger	Capable of division or merger

Estep, P.W. Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations. *AI & Soc* (2024). https://doi.org/10.1007/


Family Tree
Grand Father
Grand Mother
Grand Father
Grand Mother
Uncle
Aunt
Father
Mother
Uncle
Aunt
Cousin
Cousin
Meena
Cousin

Biological constraints establish inertia of meaning and purpose

- Limits of our brains
- Kin/tribe relationships
- Predispositions and instincts
- Narrow comfort zone
- Predictable lifespan
- Balancing independence and dependence

Biological constraints establish inertia of meaning and purpose

- Limits of our brains
- Kin/tribe relationships
- Predispositions and instincts
- Narrow comfort zone
- Predictable lifespan
- Balancing independence and dependence

- We face many, everyday challenges with no hope of a complete, lasting, optimal solution

Biological constraints establish inertia of meaning and purpose

- Limits of our brains
- Kin/tribe relationships
- Predispositions and instincts
- Narrow comfort zone
- Predictable lifespan
- Balancing independence and dependence

- We are born endowed with context and meaning, goals and purpose
- We gradually gain independence and establish our own meaning, goals , and purpose

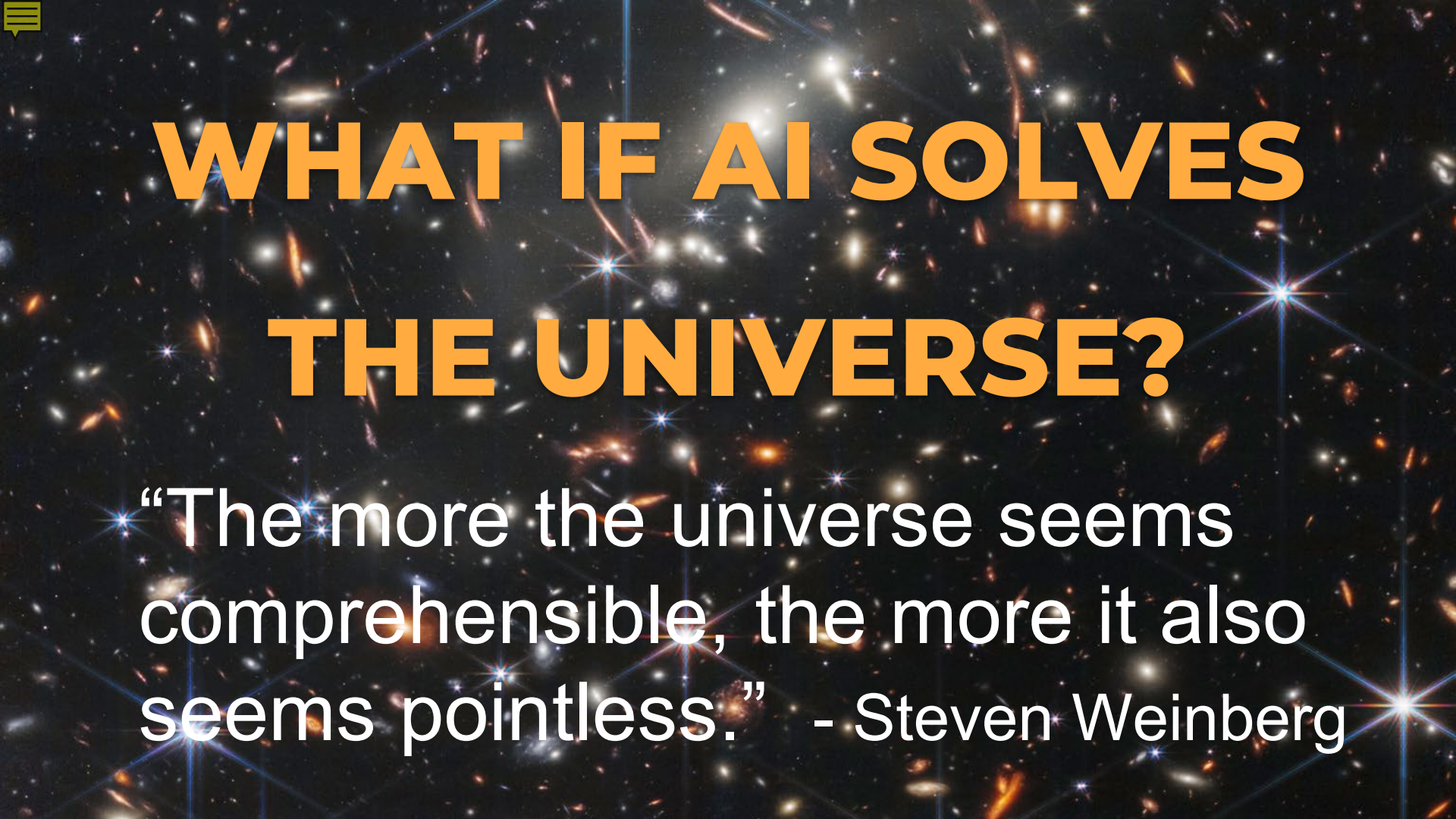
Biological constraints establish inertia of meaning and purpose

- Limits of our brains
- Kin/tribe relationships
- **Predispositions and instincts**
- Narrow comfort zone
- Predictable lifespan
- Balancing independence and dependence

- **Many of our preferences are strongly predetermined**
- **and so on**


Complex
Universe!


Complex
Universe?


WHAT IF AI SOLVES
THE UNIVERSE?
“The more the universe seems comprehensible, the more it also seems pointless.” - Steven Weinberg

The Fermi Paradox

Is life rare, is Earth first, or ...?

- >100,000,000,000 galaxies
- 100,000,000,000 stars/galaxy

The Fermi Paradox

Plausible explanations for a superintelligence:

1. Transmitting easily decoded energy into the cosmos becomes wasteful and unnecessary.
2. As a superintelligence approaches full understanding of the universe, it becomes less curious, ambitious, and motivated.
3. ?

Summary

1. Early self-improving AI might achieve self-governance
2. AI immature autonomy might be dangerous
3. AI minds are inherently unlike natural minds
4. Biology provides inertia of meaning and purpose
5. What if AI solves the universe?
6. The Fermi Paradox

Thank you!

- Microsoft New England R&D
- Dan Faggella
- Mind First Foundation & Radvac team: Ranjan Ahuja, Brian Delaney, Alex Hoekstra, Don Wang
- Vitalik Buterin and Balvi, and other supporters of MFF and Radvac
- Scott Alexander, ACX
- Eliezer Yudkowsky
- Dan Elton

Mind First Foundation

www.mindfirst.foundation