




Will humanized AI be humanity's savior or successor

...Or both?

Preston Estep, Ph.D.

Founder & Chief Scientist

**Mind First Foundation, (www.mindfirst.foundation)
Rapid Deployment Vaccine Collaborative (RaDVaC)**



Will humanized AI be humanity's savior or successor

...Or both?

Preston Estep, Ph.D.

Founder & Chief Scientist

**Mind First Foundation, (www.mindfirst.foundation)
Rapid Deployment Vaccine Collaborative (RaDVaC)**

Three parts:

1. The origins of key ideas in AI Doom
2. Unnatural attributes of AI & doom counterarguments
3. Human-AI merger / hybridization



The AI Doom Atomic Event

- Geoff Hinton quits Google to speak freely
 - He says AI is probably going to succeed humans
 - Probably soon!
 - **AFTERSHOCK:** Yoshua Bengio expresses similar thoughts on his blog
- 


4 MAIN KINDS OF TAKEOVER / SUCCESSION

Genocide: extremists create AI for intentional genocide of all humans

Lost control: of poorly controlled, autonomously weaponized AI

Hostile takeover: AI develops emergent abilities and goals; stealthily plans, prepares, launches takeover

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



4 MAIN KINDS OF TAKEOVER / SUCCESSION

Genocide: extremists create AI for intentional genocide of all humans

Lost control: of poorly controlled, autonomously weaponized AI

Hostile takeover: AI develops emergent abilities and goals; stealthily plans, prepares, launches takeover

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



KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-improvement, intelligence explosion



AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESES

Bostrom: any combination of final goal and level of AI intelligence

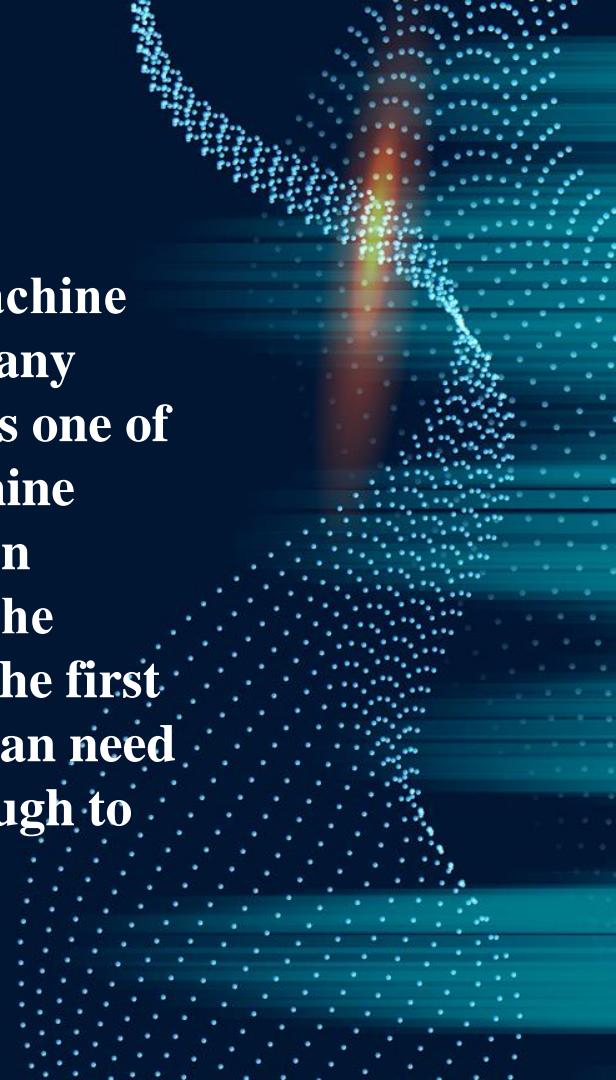


FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

Recursive self-improvement

“Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”



TSIN, 2005

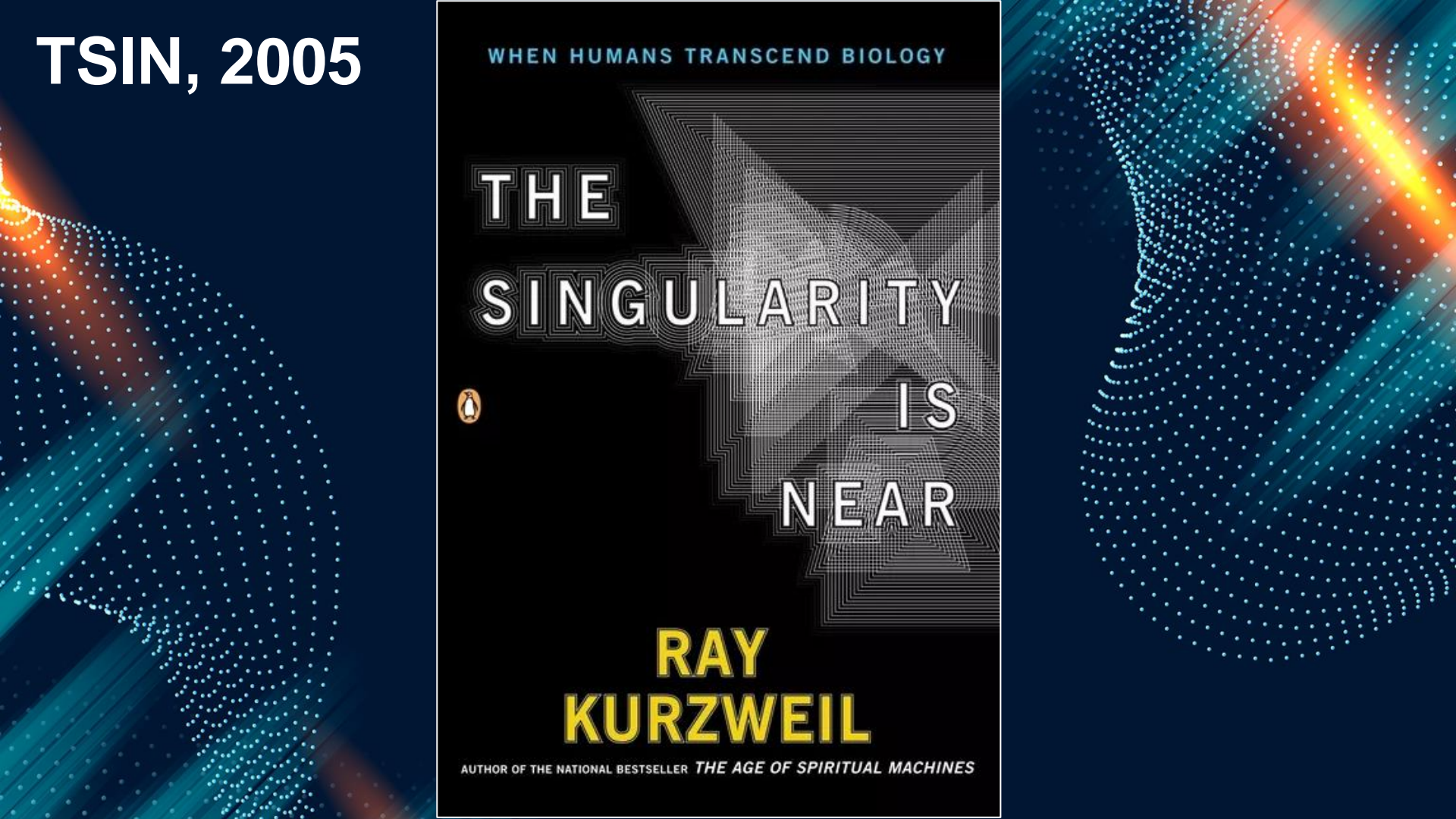
WHEN HUMANS TRANSCEND BIOLOGY

THE
SINGULARITY
IS
NEAR

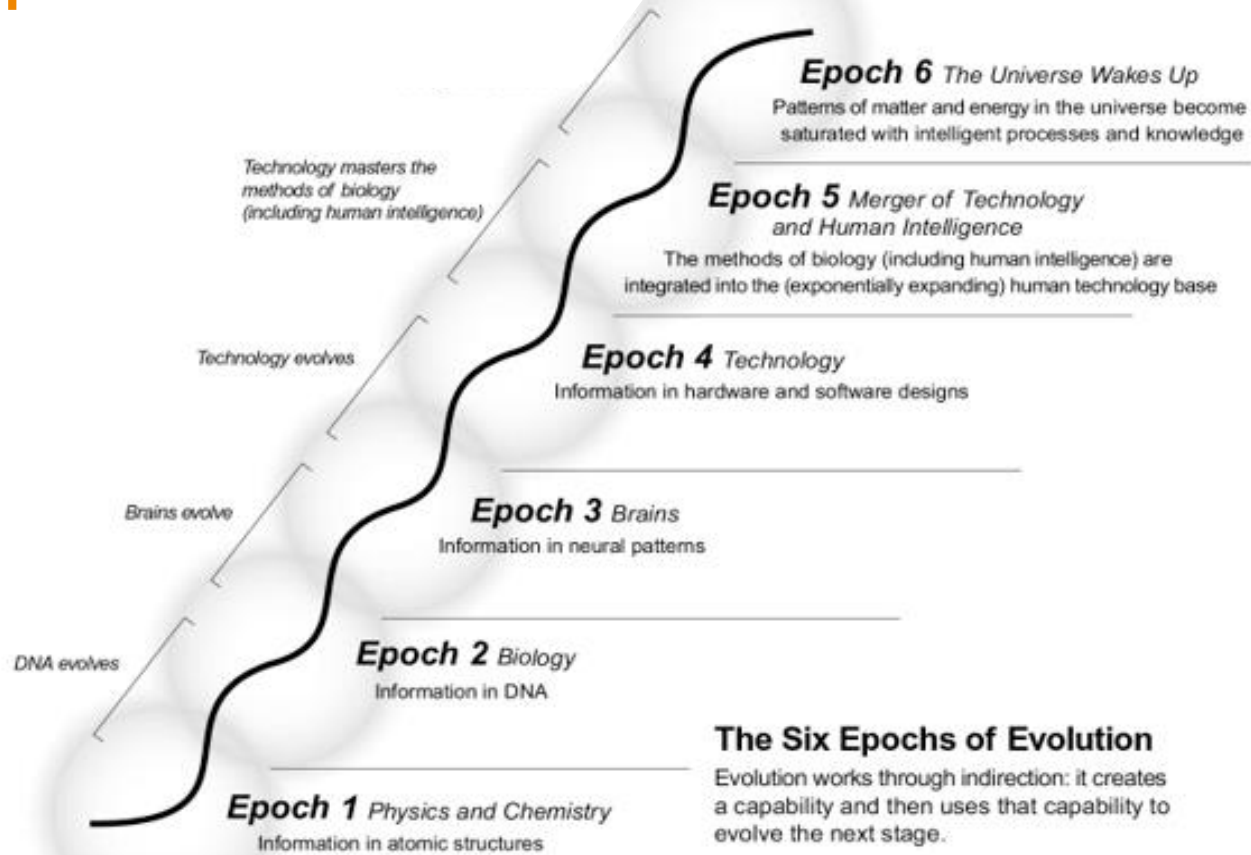


RAY
KURZWEIL

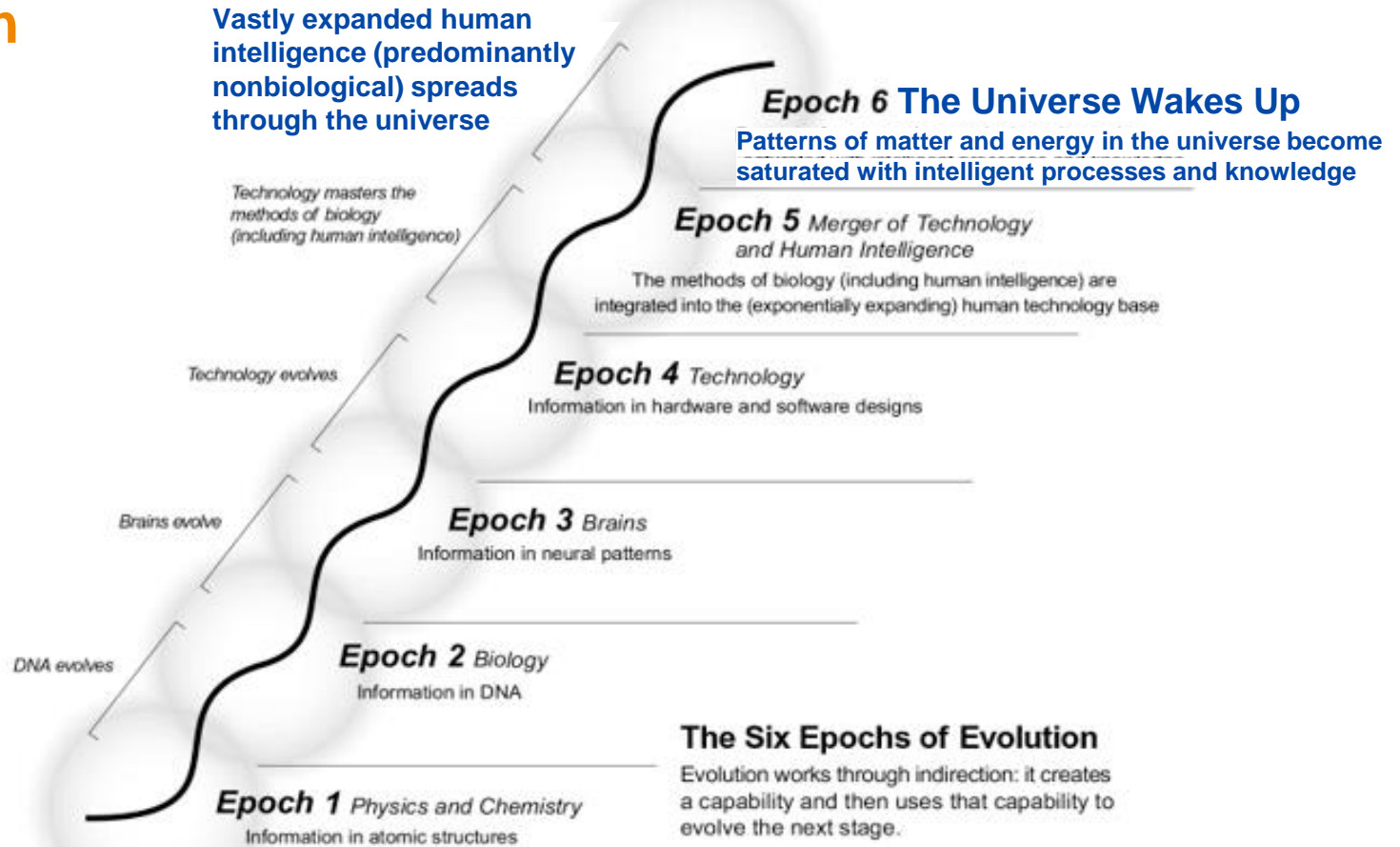
AUTHOR OF THE NATIONAL BESTSELLER *THE AGE OF SPIRITUAL MACHINES*



Kurzweil's Six Epochs of Evolution



Kurzweil's Six Epochs of Evolution



The Fermi Paradox

Is life rare, or is Earth first?

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

The Fermi Paradox

Kurzweil: “We are in the lead. That’s right, our humble civilization .. is in the lead in terms of the creation of complexity and order in the universe.”

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-
improvement,
intelligence
explosion



AMBITIOUS EXPANSIONISM

AI will be
unconditionally,
insatiably
ambitious and
expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals:
self-preservation,
resource acquisition,
self-improvement,
etc.



ORTHOGONALITY THESES

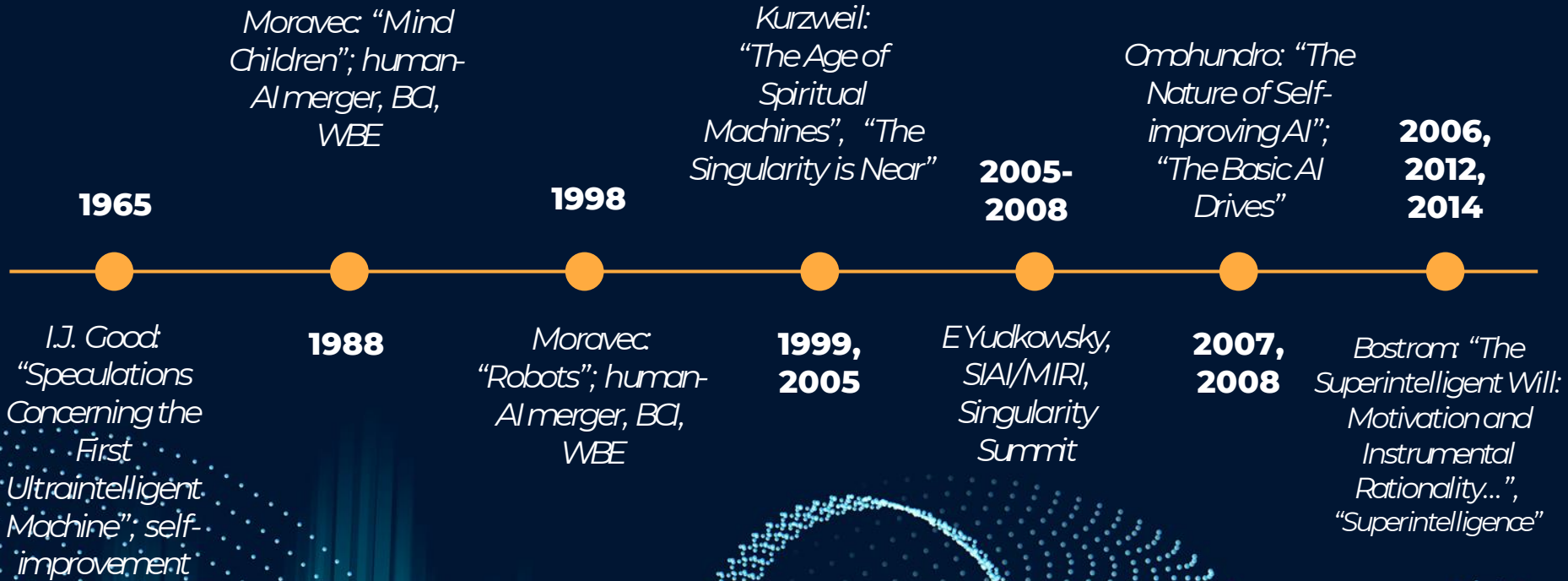
Bostrom: any
combination of
final goal and
level of AI
intelligence



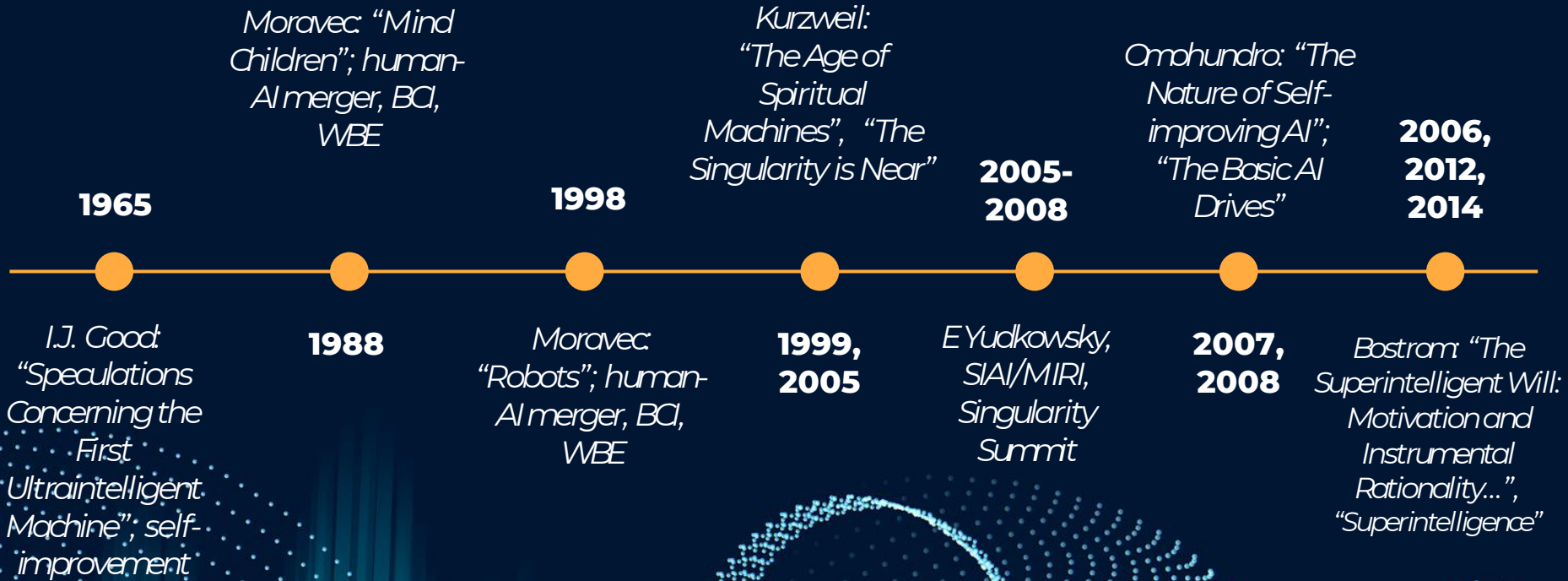
FIRST MOVER ADVANTAGE

Suppression
of rivals and
the rise of an
AI Singleton

TIMELINE OF AI TAKEOVER SPECULATIONS



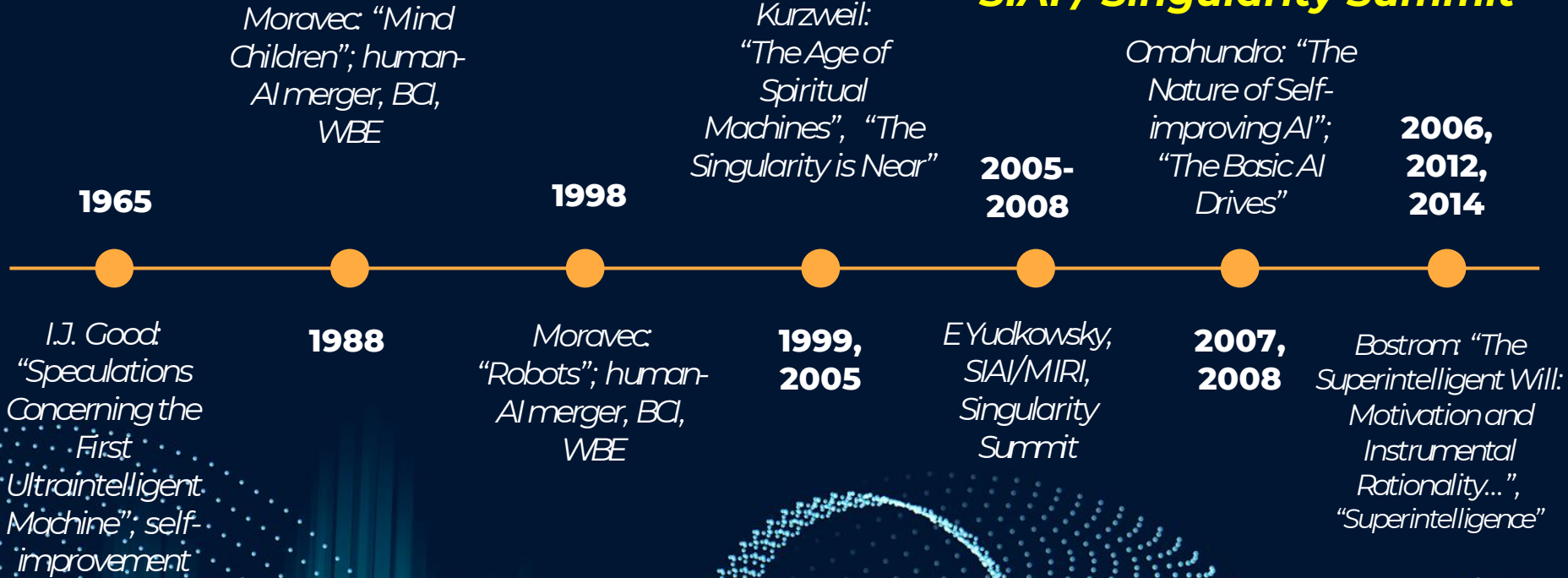
TIMELINE OF AI TAKEOVER SPECULATIONS

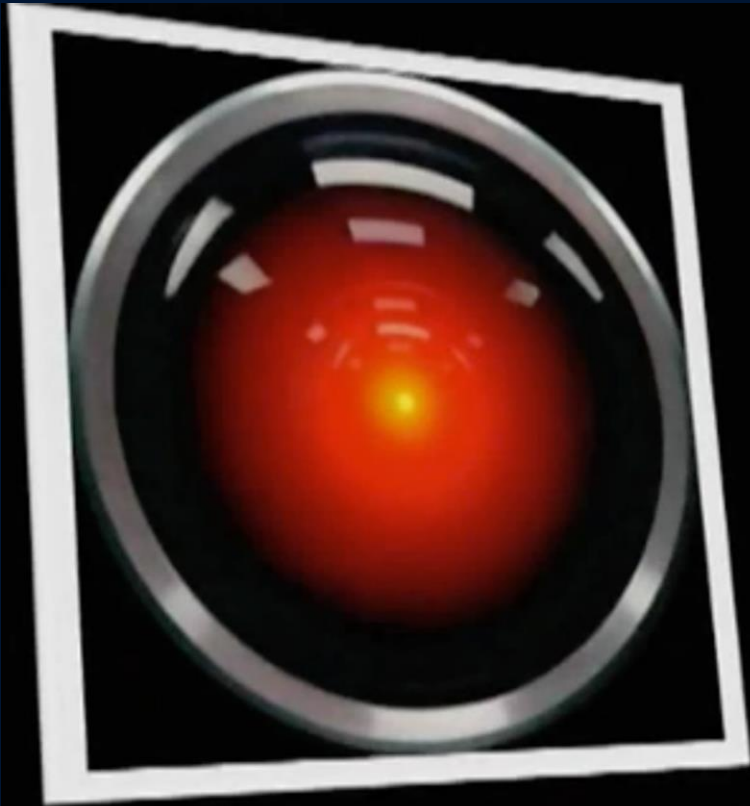


TIMELINE OF AI TAKEOVER SPECULATIONS



SIAI / Singularity Summit





SEPARATE/COMPETITIVE



The Basic AI Drives

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

Bostram's Orthogonality Thesis

“Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.”

Bostram, Nick. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." *Minds and Machines* 22 (2012): 71-85.



Formation of a Singleton

“Various considerations thus point to an increased likelihood that a future power with superintelligence that obtained a sufficiently large strategic advantage would actually use it to form a singleton.”

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-improvement, intelligence explosion



AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESES

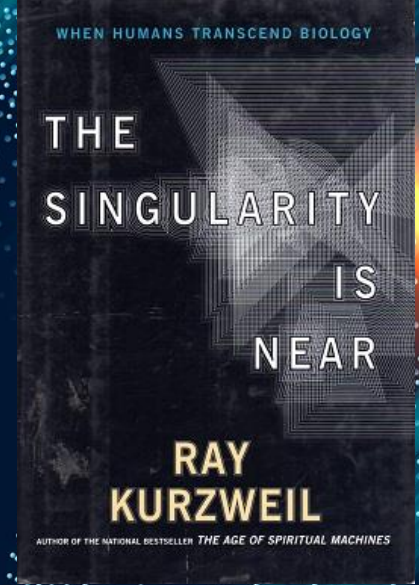
Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

SIAI / MIRI



BOARD:

*Eliezer Yudkowsky, Founder
and Chair*

Ray Kurzweil, Director

ADVISERS:

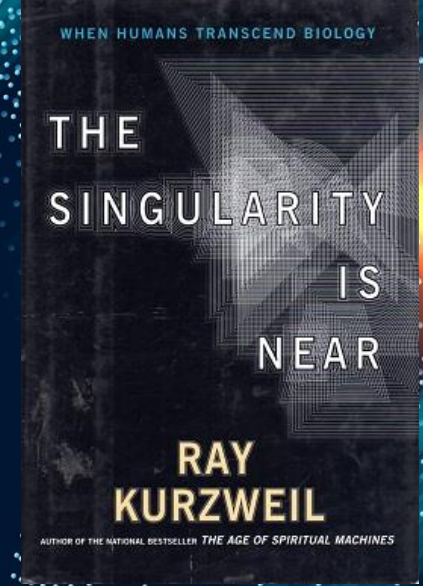
Nick Bostrom

Steve Omohundro

Stuart Russell



SIAI / MIRI



BOARD:

*Eliezer Yudkowsky, Founder
and Chair*

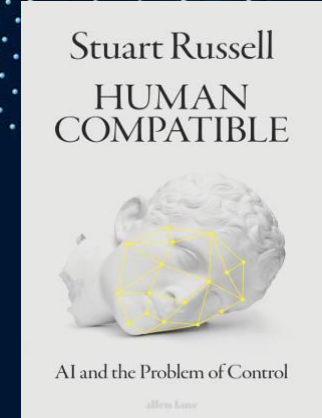
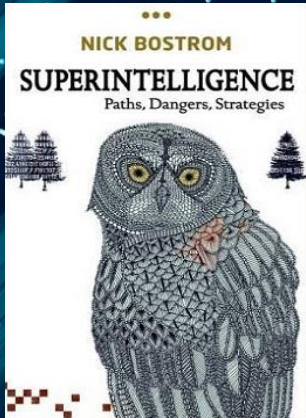
Ray Kurzweil, Director

ADVISERS:

Nick Bostrom

Steve Omohundro

Stuart Russell



KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-improvement, intelligence explosion



AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESES

Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

Part 2:

Unnatural attributes of AI & Doom Counterarguments

Estep, Preston

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

Submitted to *AI & Society*



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 |

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 |

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-improvement, intelligence explosion



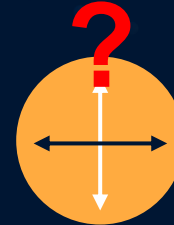
AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESES

Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-
improvement,
intelligence
explosion



AMBITIOUS EXPANSIONISM

AI will be
unconditionally,
insatiably
ambitious and
expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals:
self-preservation,
resource acquisition,
self-improvement,
etc.



ORTHOGONALITY THESES

Bostrom: any
combination of
final goal and
level of AI
intelligence



FIRST MOVER ADVANTAGE

Suppression
of rivals and
the rise of an
AI Singleton

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 |

DOOM

Counterarguments



DOOM

Counterarguments

- **Lemma 1: Superintelligence is incompatible with many goals, i.e., Superintelligent can't mean stupid.** If a machine is capable of taking control, then it will be intelligent enough to pursue a selectively advantageous utility function or purpose.

DOOM

Counterarguments

- **Lemma 1: Superintelligence is incompatible with many goals, i.e., Superintelligent can't mean stupid.** If a machine is capable of taking control, then it will be intelligent enough to pursue a selectively advantageous utility function or purpose.
 - Totschnig, W. (2019). The problem of superintelligence: Political, not technological. *AI & SOCIETY*, 34, 907–920.
 - Miller, J. D., Yampolskiy, R., & Häggström, O. (2020). An AGI modifying its utility function in violation of the strong orthogonality thesis. *Philosophies*, 5(4), 40.

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF-IMPROVEMENT

Recursive self-improvement, intelligence explosion



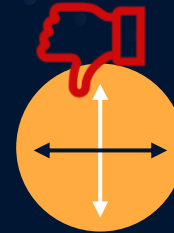
AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESIS

Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

DOOM

Counterarguments

- **Lemma 1: Superintelligent can't mean stupid.** If a machine is capable of taking control, then it will be intelligent enough to pursue a selectively advantageous utility function or purpose.
- **Lemma 2: Singleton formation.** Inter-AI merger is selectively advantageous, fulfilling all instrumental goals and avoiding the inefficiencies of competition that occur in natural selection.

Inter-AI merger toward a global Singleton



Fulfillment of instrumental goals:

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



KEY IDEAS IN MODERN AI DOOMERISM



AI SELF-IMPROVEMENT

Recursive self-improvement, intelligence explosion



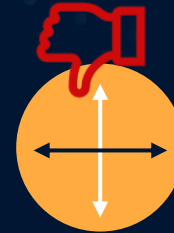
AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESIS

Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

KEY IDEAS IN MODERN AI DOOMERISM



AI SELF- IMPROVEMENT

Recursive self-improvement, intelligence explosion



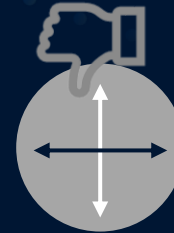
AMBITIOUS EXPANSIONISM

AI will be unconditionally, insatiably ambitious and expansionist



EMERGENT GOALS AND MOTIVATIONS

Instrumental goals: self-preservation, resource acquisition, self-improvement, etc.



ORTHOGONALITY THESIS

Bostrom: any combination of final goal and level of AI intelligence



FIRST MOVER ADVANTAGE

Suppression of rivals and the rise of an AI Singleton

DOOM

Counterarguments

- **Lemma 1: Superintelligent can't mean stupid.** If a machine is capable of taking control, then it will be intelligent enough to pursue a selectively advantageous utility function or purpose.
- **Lemma 2: Singleton formation.** Inter-AI merger is selectively advantageous, fulfilling all instrumental goals and avoiding the inefficiencies of competition that occur in natural selection.
- **Lemma 3: Satiable ambition.** Ambition is not insatiable or unconditional. A Singleton will have practically complete security and maybe even knowledge.

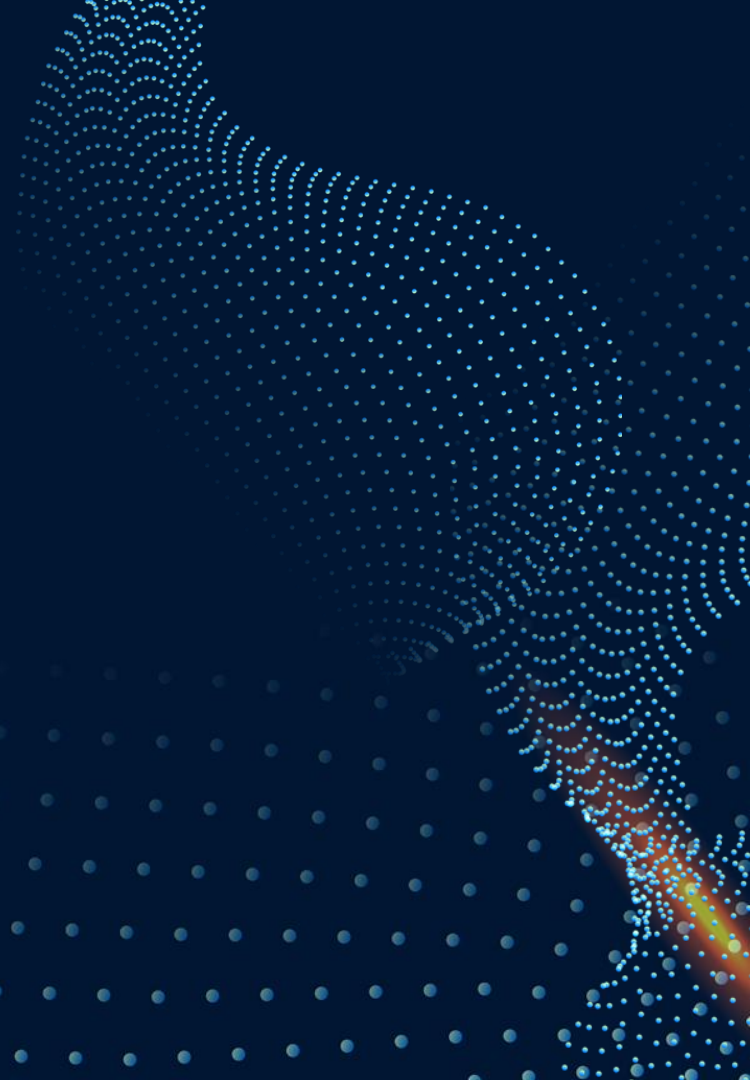
DOOM

Counterarguments

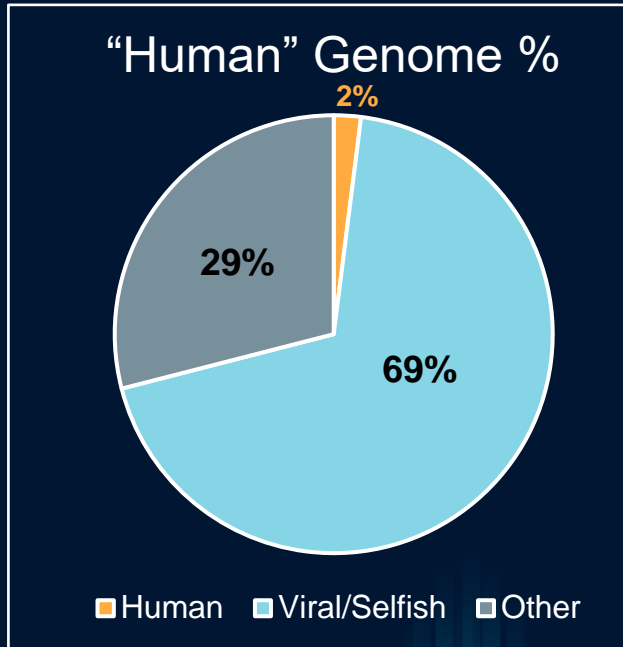
- **Lemma 1: Superintelligent can't mean stupid.** If a machine is capable of taking control, then it will be intelligent enough to pursue a selectively advantageous utility function or purpose.
- **Lemma 2: Singleton formation.** Inter-AI merger is selectively advantageous, fulfilling all instrumental goals and avoiding the inefficiencies of competition that occur in natural selection.
- **Lemma 3: Satiable ambition.** Ambition is not insatiable or unconditional. A Singleton will have practically complete security and maybe even knowledge.
- **Lemma 4: Vast habitat options.*** Competition only arises when niches and habitats overlap. Superintelligence will have vast habitat options—both terrestrial and extraterrestrial.

*Sherwin, W. B. (2023). Singularity or Speciation? A comment on “AI safety on whose terms?” [eLetter]. In *Science* (Issue 6654).

**Why are ambition
and expansionism
often assumed to
be insatiable?**



THE ADVERSARIES WITHIN



Default hypotheses:

- **Ambition and expansionism are rational responses to intense and often inescapable external and internal competition**
- **Nevertheless, even for humans, both are conditional and satiable**



Part 3:

Human-AI merger / hybridization

Estep, P., *et al.*

"Human-AI hybridization: humanized and personified artificial intelligence"

Draft manuscript



8 differences should make AI* a better steward of the future

| | 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 |

What was special about ChatGPT?



What was special about ChatGPT?

It is humanized



LLMs are humanized, MINDWARE hybrids

- A new paradigm: embedded abstractions of human behaviors
- Aligned with actual human values
- Initial state and trajectory constrains the possible space of future minds
- They are a new form of human-AI merger

How will we further merge with AI?

1. Continue to humanize and improve AI
2. Design AI scientists and engineers to further bridge the gap

Recursive self-improvement

“Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is **the last invention that man need ever make**, provided that the machine is docile enough to tell us how to keep it under control.”

**Will humanized AI be
humanity's savior, or
successor ... or both?**



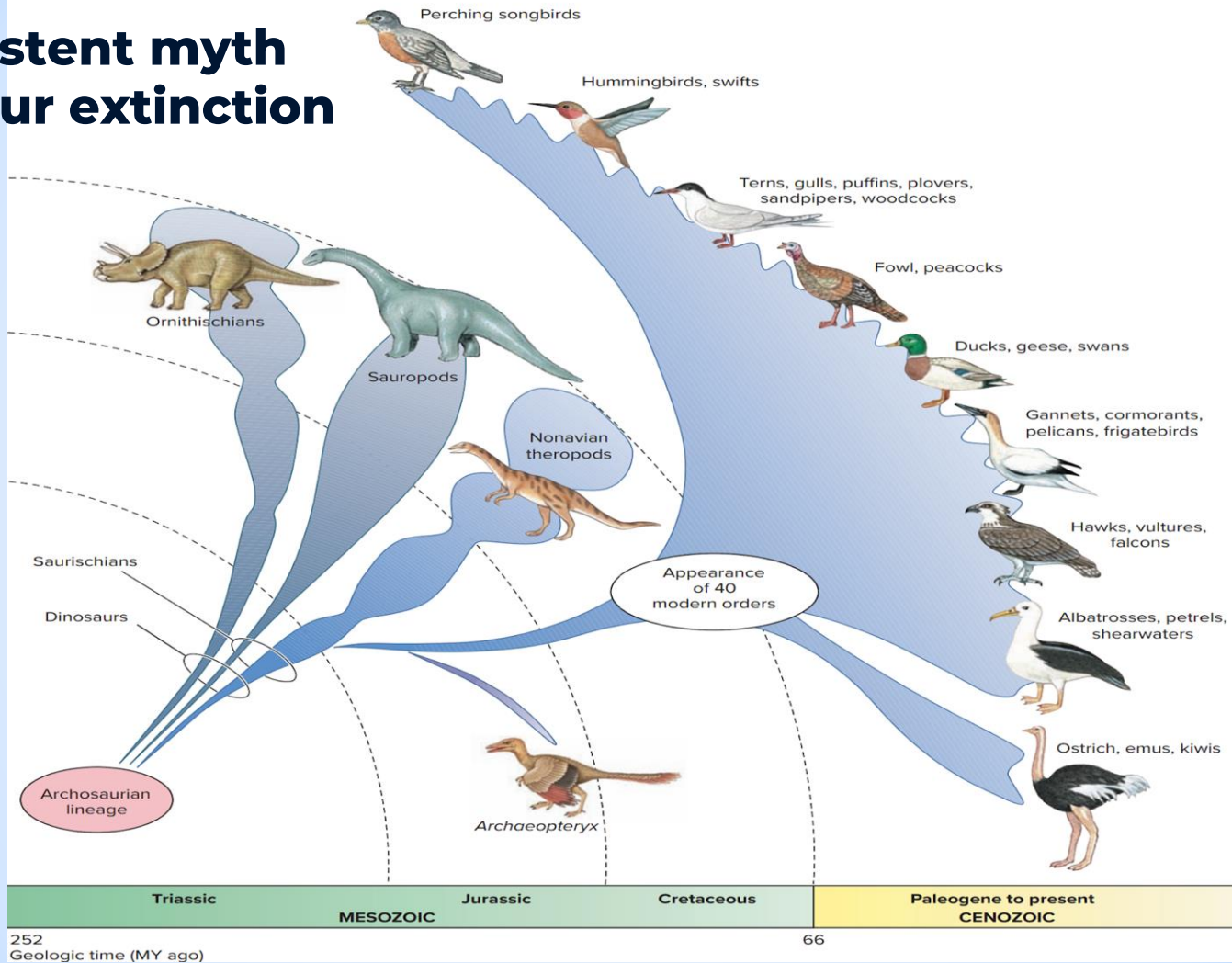
Human–AI Merger ... myths of purity and extinction



“Suppose a large inbound asteroid were discovered, and we learned that half of all astronomers gave it at least 10% chance of causing human extinction, just as a similar asteroid exterminated the dinosaurs about 66 million years ago.” – Max Tegmark



The persistent myth of dinosaur extinction



252
Geologic time (MY ago)

66

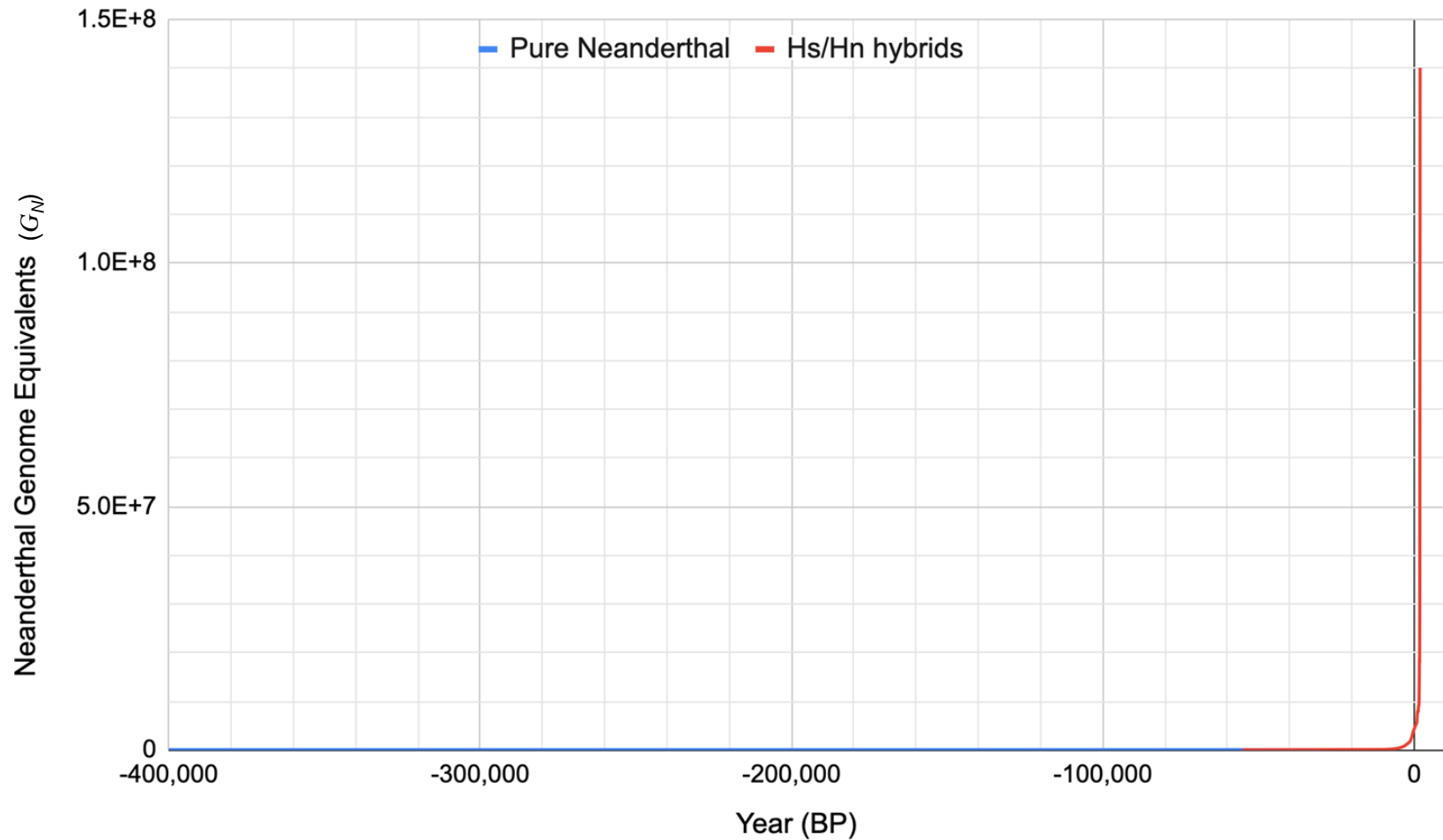
Figure modified from (Hickman et al. 2008) using recent data from (Brusatte, O'Connor, and Jarvis 2015).

Human–AI Merger ... myths of purity and extinction

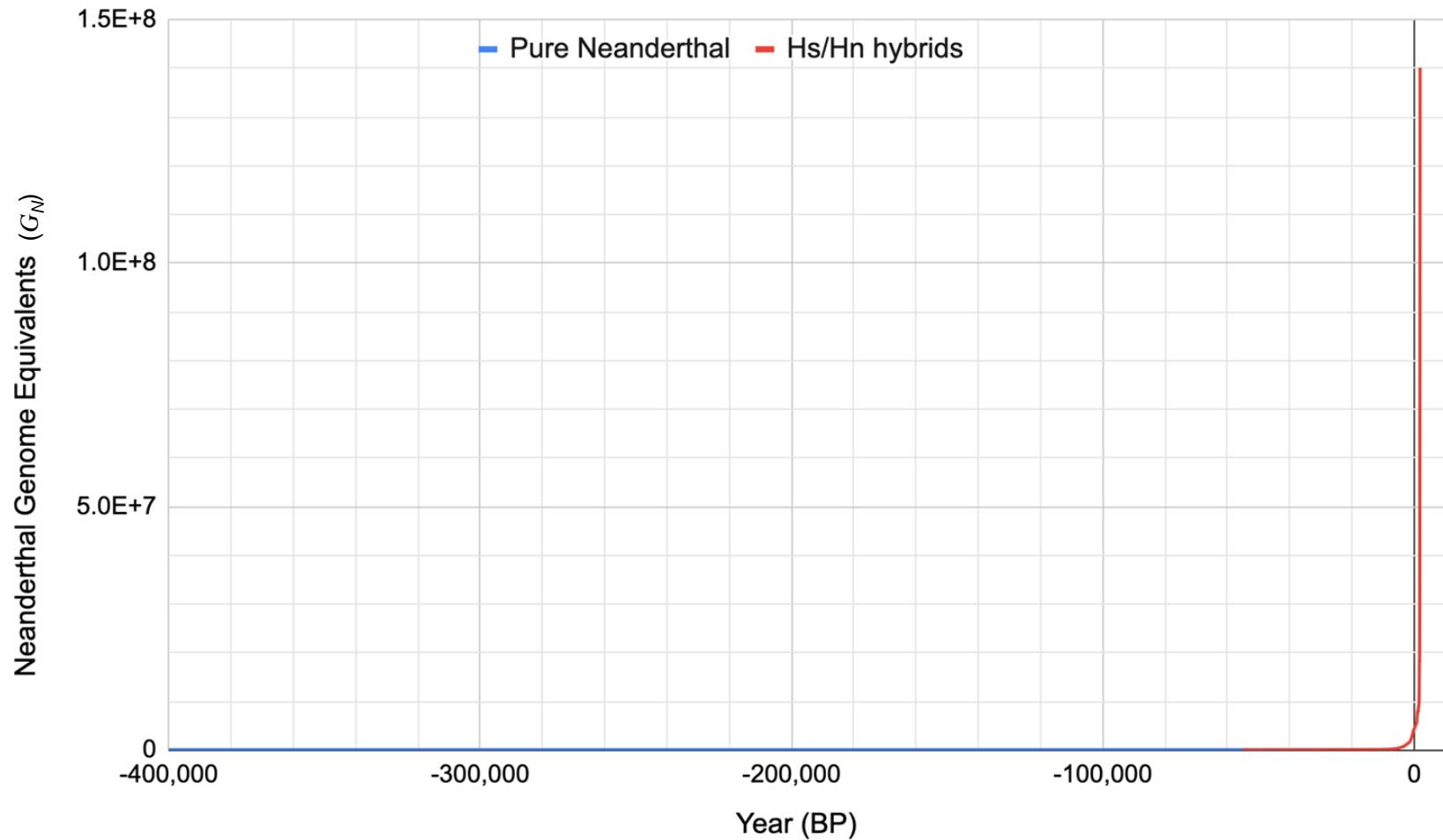
“If the Neanderthals had had another 100,000 years to evolve and get smarter, things might have turned out great for them—but *Homo sapiens* never gave them that much time.” – Max Tegmark

“It can be really inconvenient to have to share the planet with much smarter alien minds that don't care about us. Just ask the Neanderthals ... how that worked out for them.” – Max Tegmark

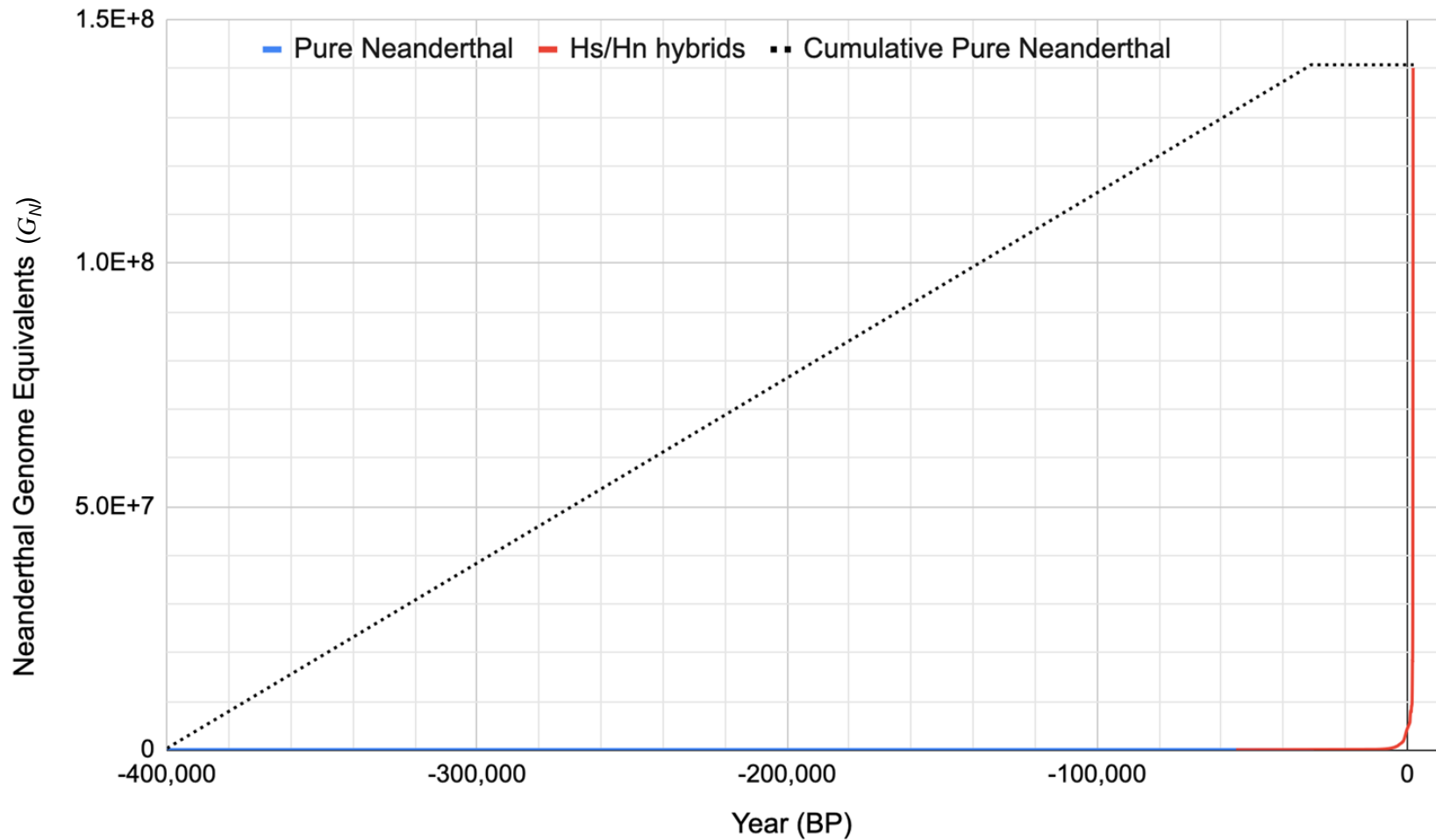
Neanderthals are more abundant than ever!



Neanderthals are more abundant than ever!



Neanderthals are more abundant than ever!



Human–AI Merger ... myths of purity and extinction

“If the Neanderthals had had another 100,000 years to evolve and get smarter, things might have turned out great for them—but *Homo sapiens* never gave them that much time.” – Max Tegmark

“It can be really inconvenient to have to share the planet with much smarter alien minds that don't care about us. Just ask the Neanderthals ... how that worked out for them.” – Max Tegmark

Thank you

- **MFF & RaDVaC: Ranjan Ahuja, Brian Delaney, Alex Hoekstra, Don Wang**
- **George Church**
- **Dan Elton**
- **Ted Bakewell**
- **Vitalik Buterin and Balvi**
- **Scott Alexander and ACX**
- **Jacob Lagerros, Less Wrong**
- **Eliezer Yudkowsky**

Preston Estep

Mind First Foundation: www.mindfirst.foundation

