Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations

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

The AI Doom Atomic Event

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

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, preserve utility function, 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

Do the big 5 lead to …

AI DOOM?

Or is high p(DOOM) due to ANTHROPOMORPHIC BIAS?

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, preserve utility function, 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

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.”

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers* (Vol. 6, pp. 31-88). Elsevier.

The Basic AI Drives

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

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.

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, preserve utility function, 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

Bostrom’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.”

Bostrom, 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.”

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies. Oxford University Press*, 2014. p. 109

Unnatural attributes of AI & Doom Counterarguments

Estep, Preston

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

AI & Society, Accepted pending revision

8 fundamental differences

	NI	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

NI = Natural Intelligence

8 fundamental differences accelerate AI evolution

	NI	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

NI = Natural Intelligence

7 fundamental differences defuse competition …

	NI	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

NI = Natural Intelligence

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, 👎preserve utility function,👎 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

Inter-AI merger toward a global Singleton

Fulfillment of
instrumental goals:

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

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.

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, 👎preserve utility function,👎 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, 👎preserve utility function👎 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.

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

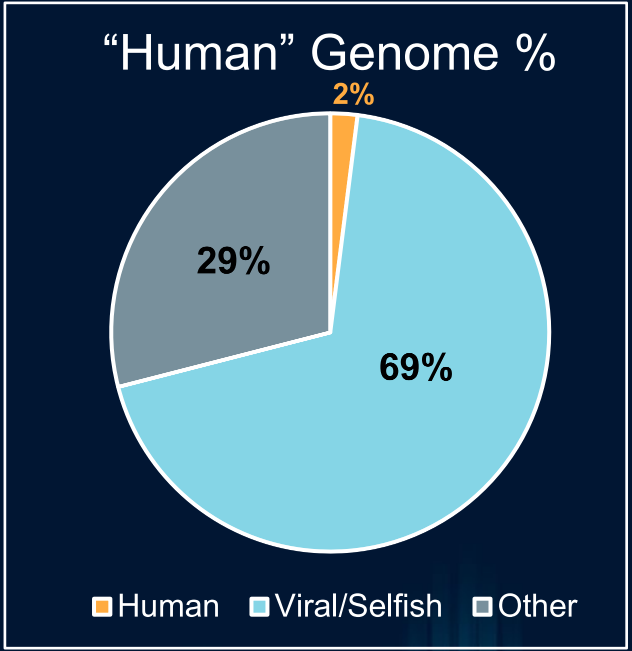
DOOM Counterarguments

- **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



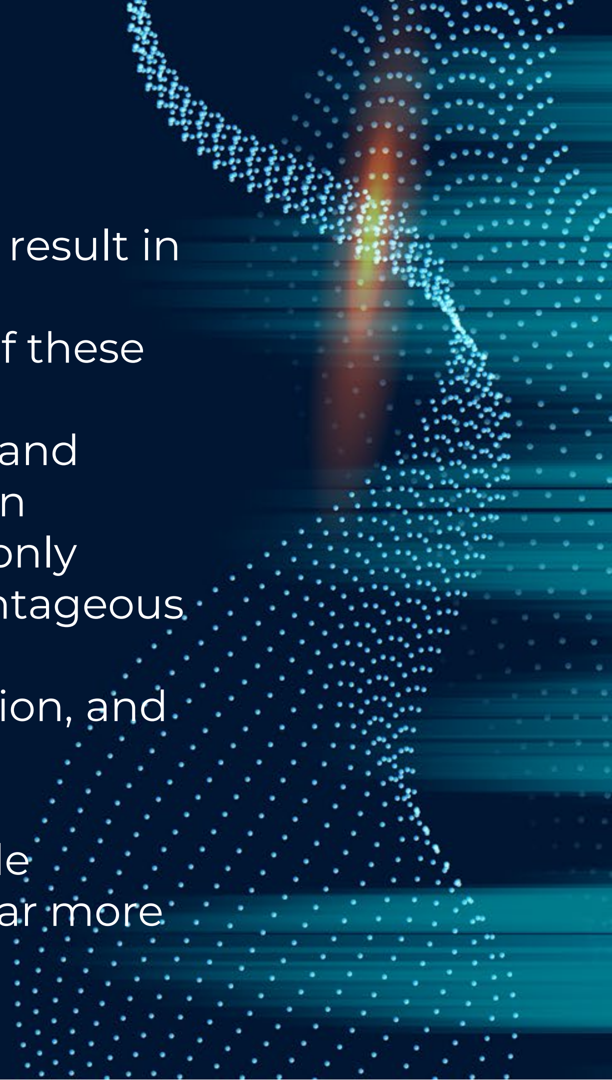
de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). *PLoS Genetics*, *7*(12), e1002384.

Default hypotheses:

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

Summary

1. 8 fundamental differences between NI and AI will result in different goals, values, and behaviors
2. All 8 differences will all accelerate AI evolution; 7 of these will defuse inter-AI competition
3. Recursive self-improvement and emergent goals and behaviors will likely lead to AI takeover / succession
4. Superintelligence will be strategically Darwinian; only certain goals/utility functions are selectively advantageous
5. Inter-AI merger satisfies all instrumental goals
6. Merger into a Singleton defuses inter-AI competition, and potentially neutralizes ambitious expansionism
7. These differences undermine common anthropomorphically biased speculations of hostile takeover, and suggest incremental succession is far more likely to occur because people will desire it



Thank you

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

Preston Estep

Mind First Foundation: www.mindfirst.foundation