AGI Safety & Alignment

Data Circles Journal Club 3-2-22

Weapons of Math Destruction by Cathy O'Neil

Shell Shocked: Journey of Disillusionment

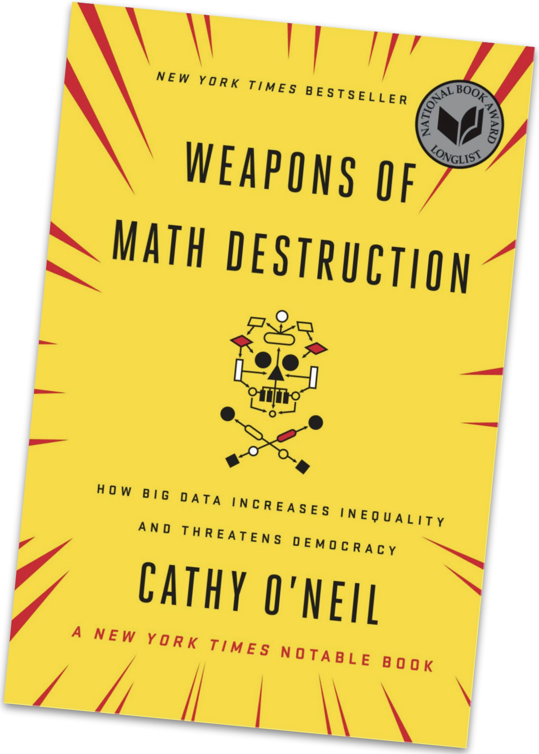
Arms Race: Going to College

Civilian Casualties: Justice in the Age of Big Data

Ineligible to Serve: Getting a Job

No Safe Zone: Getting Insurance

Targeted Citizen: Civic Life


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK

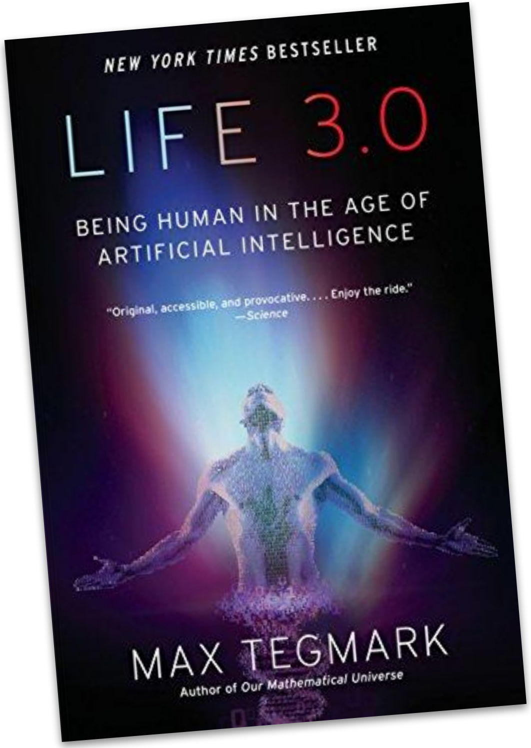
Life 3.0 by Max Tegmark

Life 1.0 fixed hardware & software

Life 2.0 fixed hardware, upgrade software

Life 3.0 upgrade both hardware & software

Explore scenarios with AGI


NEW YORK TIMES BESTSELLER
LIFE 3.0
BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE
"Original, accessible, and provocative. . . . Enjoy the ride." —Science
MAX TEGMARK
Author of Our Mathematical Universe

"Any sufficiently ***advanced technology***
is indistinguishable from ***magic****."*

~Arthur C. Clarke

The Life You Can Save by Peter Singer

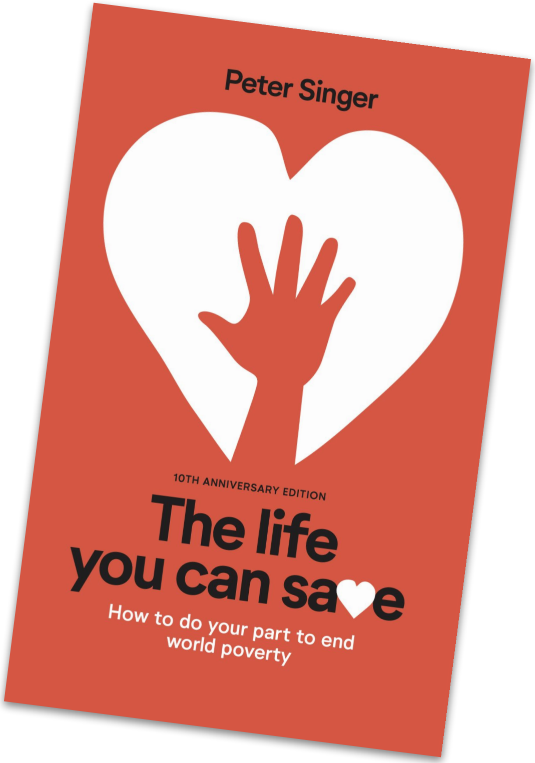
Effective altruism

80,000 hours

Pressing global issues list

Highest priority areas

- Global priorities research
- Building effective altruism
- Reducing global catastrophic biological risks
- Positively shaping the development of AI



*“Fearing a rise of **killer robots** is like worrying about **overpopulation on Mars**”*

~Andrew Ng

AGI Safety

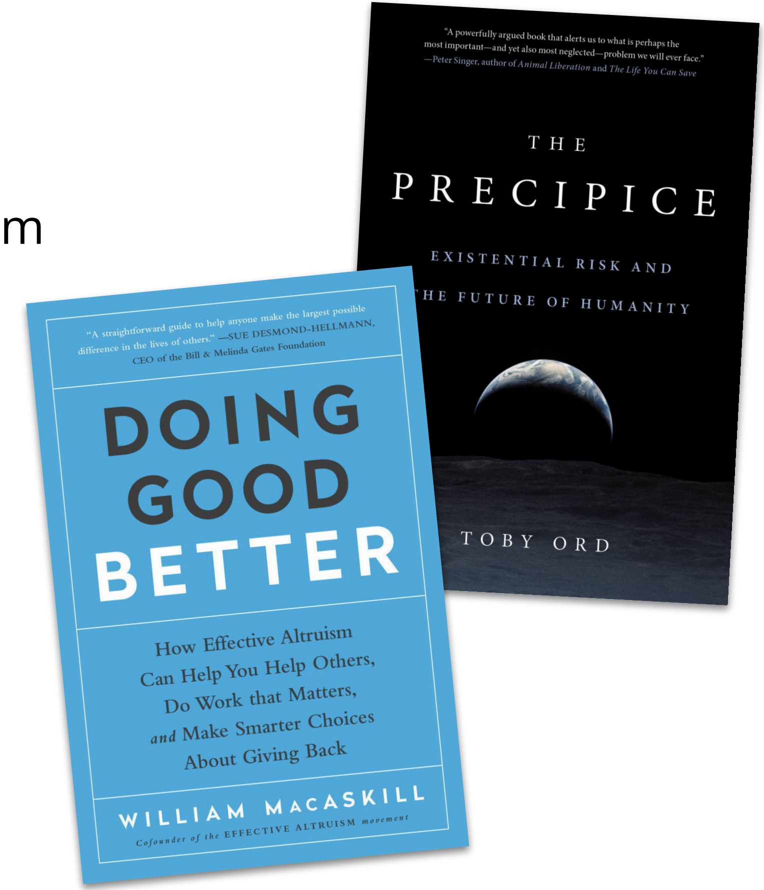
aka the Control Problem, AI Alignment Problem

Long-termism

- Will MacAskill, *Doing Good Better* (2015)
- Toby Ord, *The Precipice* (2020)

Recommended readings

- Nick Bostrom, *Superintelligence* (2014)
- Stuart Russell, *Human Compatible* (2019)
- Brian Christian, *Alignment Problem* (2020)

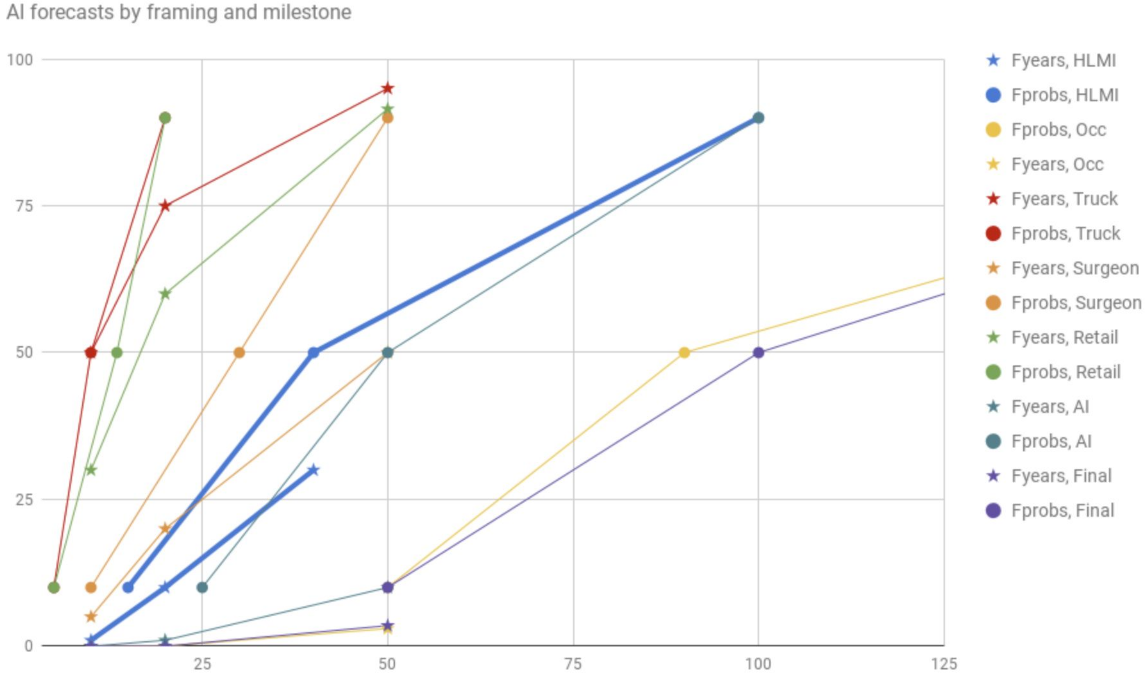

"A powerfully argued book that alerts us to what is perhaps the most important—and yet also most neglected—problem we will ever face." —Peter Singer, author of Animal Liberation and The Life You Can Save
THE PRECIPICE
EXISTENTIAL RISK AND THE FUTURE OF HUMANITY
TOBY ORD
"A straightforward guide to help anyone make the largest possible difference in the lives of others." —SUE DESMOND-HELLMANN, CEO of the Bill & Melinda Gates Foundation
DOING GOOD BETTER
How Effective Altruism Can Help You Help Others, Do Work that Matters, and Make Smarter Choices About Giving Back
WILLIAM MACASKILL
Cofounder of the EFFECTIVE ALTRUISM movement

*“We will, **sooner or later**, build an artificial **agent** with **general intelligence**.”*

~Rob Miles

2016 Expert Survey

https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/



Superintelligence by Nick Bostrom

History of ups and downs

Paths: AI vs Whole Brain Emulation (WBE)

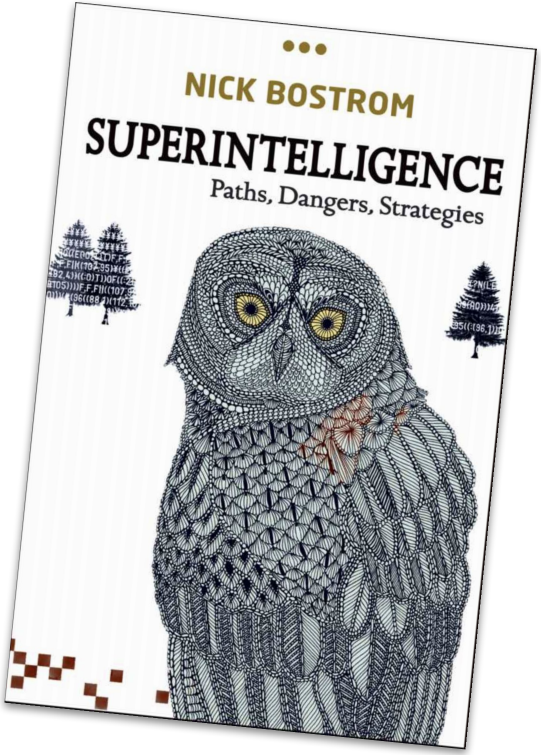
Speed: fast vs slow

Orthogonality thesis: Paperclip problem

Instrumental goals

- Self-preservation
- Resource acquisition
- Self-improvement

Learn human values



*“The development of **full artificial intelligence** could spell the **end** of the **human race**.”*

~Stephen Hawking

Human Compatible by Stuart Russell

Conceptual breakthroughs needed

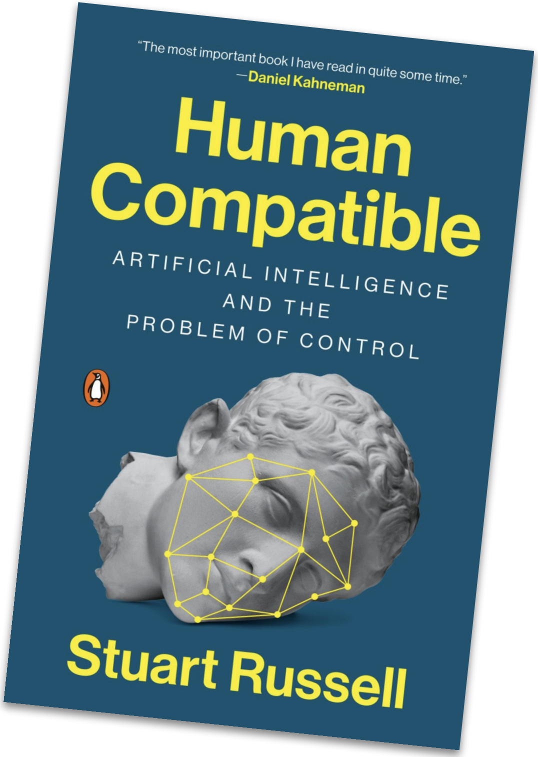
Gorilla problem, 2nd species

Beneficial machines:

- Purely altruistic
- Humble
- Learn to predict

Many benefits

Many risks


"The most important book I have read in quite some time."
—Daniel Kahneman
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell

*“Be **careful** what you **wish** for!”*

~King Midas

Alignment Problem by Brian Christian

Machine Learning and Human Values

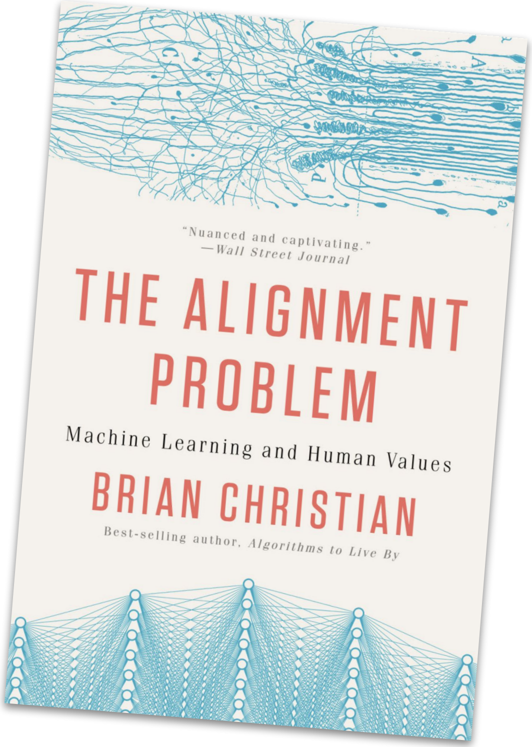
Prophecy: Representation, Fairness, Transparency

Agency

- Reinforcement: Policy vs Value, Temporal Diff
- Shaping: Rewards & Behaviors
- Curiosity: Novelty, Intrinsic Motivation

Normativity:

- Imitation: Recover Cascading Errors
- Inference: Demos, Feedback, Cooperation
- Uncertainly: Impact, Corrigible


"Nuanced and captivating."
—Wall Street Journal
THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-selling author, Algorithms to Live By

“Biased ***algorithms*** *are easier to fix than biased* ***people****.”*

~Sendhil Mullainathan

References

Cathy O’Neil, *Weapons of Math Destruction* (2016)

Max Tegmark, *Life 3.0* (2017)

Peter Singer, *The Life You Can Save* (2019)

Will MacAskill, *Doing Good Better* (2015)

Toby Ord, *The Precipice* (2020)

Nick Bostrom, *Superintelligence* (2014)

Stuart Russell, *Human Compatible* (2019)

Brian Christian, *Alignment Problem* (2020)