Title: The Alignment Problem

URL: https://en.wikipedia.org/wiki/The_Alignment_Problem

PageID: 69438830

Categories: Category:2020 non-fiction books, Category:Books about effective altruism, Category:Books about existential risk, Category:English-language non-fiction books, Category:English non-fiction books, Category:Existential risk from artificial intelligence, Category:Futurology books, Category:Non-fiction books about artificial intelligence, Category:W. W. Norton & Company books

---

The Alignment Problem: Machine Learning and Human Values is a 2020 non-fiction book by the American writer Brian Christian . It is based on numerous interviews with experts trying to build artificial intelligence systems, particularly machine learning systems, that are aligned with human values.

Summary

The book is divided into three sections: Prophecy, Agency, and Normativity. Each section covers researchers and engineers working on different challenges in the alignment of artificial intelligence with human values.

Prophecy

In the first section, Christian interweaves discussions of the history of artificial intelligence research, particularly the machine learning approach of artificial neural networks such as the Perceptron and AlexNet , with examples of how AI systems can have unintended behavior. He tells the story of Julia Angwin , a journalist whose ProPublica investigation of the COMPAS algorithm, a tool for predicting recidivism among criminal defendants, led to widespread criticism of its accuracy and bias towards certain demographics. One of AI's main alignment challenges is its black box nature (inputs and outputs are identifiable but the transformation process in between is undetermined). The lack of transparency makes it difficult to know where the system is going right and where it is going wrong.

Agency

In the second section, Christian similarly interweaves the history of the psychological study of reward, such as behaviorism and dopamine , with the computer science of reinforcement learning , in which AI systems need to develop policy ("what to do") in the face of a value function ("what rewards or punishment to expect"). He calls the DeepMind AlphaGo and AlphaZero systems "perhaps the single most impressive achievement in automated curriculum design." He also highlights the importance of curiosity, in which reinforcement learners are intrinsically motivated to explore their environment, rather than exclusively seeking the external reward.

Normativity

The third section covers training AI through the imitation of human or machine behavior, as well as philosophical debates such as between possibilism and actualism that imply different ideal behavior for AI systems. Of particular importance is inverse reinforcement learning , a broad approach for machines to learn the objective function of a human or another agent. Christian discusses the normative challenges associated with effective altruism and existential risk , including the work of philosophers Toby Ord and William MacAskill who are trying to devise human and machine strategies for navigating the alignment problem as effectively as possible.

Reception

The book received positive reviews from critics. The Wall Street Journal 's David A. Shaywitz emphasized the frequent problems when applying algorithms to real-world problems, describing the book as "a nuanced and captivating exploration of this white-hot topic." Publishers Weekly praised

the book for its writing and extensive research.

Kirkus Reviews gave the book a positive review, calling it "technically rich but accessible", and "an intriguing exploration of AI." Writing for Nature , Virginia Dignum gave the book a positive review, favorably comparing it to Kate Crawford 's Atlas of AI .

In 2021, journalist Ezra Klein had Christian on his podcast, The Ezra Klein Show, writing in The New York Times , " The Alignment Problem is the best book on the key technical and moral questions of A.I. that I've read." Later that year, the book was listed in a Fast Company feature, "5 books that inspired Microsoft CEO Satya Nadella this year".

In 2022, the book won the Eric and Wendy Schmidt Award for Excellence in Science Communication , given by The National Academies of Sciences, Engineering, and Medicine in partnership with Schmidt Futures .

In 2024, The New York Times placed The Alignment Problem first in its list of the "5 Best Books About Artificial Intelligence," saying: "If you're going to read one book on artificial intelligence, this is the one."

See also

Effective altruism

Global catastrophic risk

Human Compatible: Artificial Intelligence and the Problem of Control

Superintelligence: Paths, Dangers, Strategies

References

v

t

e

Aid effectiveness

Charity assessment

Demandingness objection

Disability-adjusted life year

Disease burden

Distributional cost-effectiveness analysis

Earning to give

Equal consideration of interests

Incremental cost-effectiveness ratio

Longtermism

Marginal utility

Moral circle expansion

Psychological barriers to effective altruism

Quality-adjusted life year

Utilitarianism

Venture philanthropy

Sam Bankman-Fried

Liv Boeree

Nick Bostrom
Hilary Greaves
Holden Karnofsky
William MacAskill
Dustin Moskovitz
Yew-Kwang Ng
Toby Ord
Derek Parfit
Kelsey Piper
Peter Singer
Brian Tomasik
Cari Tuna
Eliezer Yudkowsky
80,000 Hours
Against Malaria Foundation
Animal Charity Evaluators
Animal Ethics
Centre for Effective Altruism
Centre for Enabling EA Learning & Research
Center for High Impact Philanthropy
Centre for the Study of Existential Risk
Development Media International
Evidence Action
Faunalytics
Fistula Foundation
Future of Humanity Institute
Future of Life Institute
Founders Pledge
GiveDirectly
GiveWell
Giving Multiplier
Giving What We Can
Good Food Fund
The Good Food Institute
Good Ventures
The Humane League
Mercy for Animals
Machine Intelligence Research Institute

Malaria Consortium
Open Philanthropy
Raising for Effective Giving
Sentience Institute
Unlimit Health
Wild Animal Initiative
Biotechnology risk
Climate change
Cultured meat
Economic stability
Existential risk from artificial general intelligence
Global catastrophic risk
Global health
Global poverty
Intensive animal farming
Land use reform
Life extension
Malaria prevention
Mass deworming
Neglected tropical diseases
Risk of astronomical suffering
Wild animal suffering
Doing Good Better
The End of Animal Farming
Famine, Affluence, and Morality
The Life You Can Save
Living High and Letting Die
The Most Good You Can Do
Practical Ethics
The Precipice
Superintelligence: Paths, Dangers, Strategies
What We Owe the Future
Effective Altruism Global