

The Alignment Problem: Machine Learning and Human Values

Description

By Brian Christian. 334 pages. 2020 Norton. [Author's web page here](#)

[Tweet](#)

Brian Christian [talking about his book on YouTube](#)

RD comment: This is one of the most interesting and informative books I have read in the last few years. Totally relevant for evaluators thinking about the present and about future trends

CONTENTS

PROLOGUE	1
INTRODUCTION	5
I. Prophecy	
1 REPRESENTATION	17
2 FAIRNESS	51
3 TRANSPARENCY	82
II. Agency	
4 REINFORCEMENT	121
5 SHAPING	152
6 CURIOSITY	181
III. Normativity	
7 IMITATION	213
8 INFERENCE	251
9 UNCERTAINTY	277
CONCLUSION	311

“Users have a right to see and to alter any preference model that a site or app or advertiser has about them. It is worth considering regulation to this effect: to say, in essence, I have the right to my own models. I have the right to say, “That’s not who I am”. [Or,](#)

aspirationally, “This is who I want to be. This is the person in whose interest you must work” p275

"This is the delicacy of our present moment. Our digital butlers are watching closely. They see our private as well as our public lives, our best and worst selves, without necessarily knowing which is which or making a distinction at all. They by a large reside in a kind of uncanny valley of sophistication: able to infer sophisticated models of our desires from our behaviour, but unable to be taught, and disinclined to cooperate. They are thinking hard about what we're going to do next, about how they might make their next commission, but they don't seem to understand what we want, much less who we hope to become"

Category

1. Uncategorized

Date

13/12/2025

Date Created

16/11/2020

Author

admin