

Rebecca Raper

Raising Robots to be Good



A Practical Foray
into the Art and Science
of Machine Ethics



Springer

Raising Robots to be Good

Rebecca Raper

Raising Robots to be Good

A Practical Foray into the Art and Science
of Machine Ethics

 Springer

Rebecca Raper
Centre for Robotics and Assembly
Cranfield University
Coventry, UK

ISBN 978-3-031-75035-9 ISBN 978-3-031-75036-6 (eBook)
<https://doi.org/10.1007/978-3-031-75036-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

*Never let your sense of morals prevent you
from doing what is right.*

– Isaac Asimov (1960)

*Everything's got a moral, if only you can
find it.*

– Alice from 'Alice's Adventures in
Wonderland' (Lewis Carroll, 1865)

Inspired by and made for Lilly and Jack.

Preface and Acknowledgements

Though the outcome of my PhD research, this book has been the culmination of 7 years intense thinking around how we might create machines with morals. I was first introduced to the topic back in 2017 when I saw a PhD advertisement to work on the topic of ‘Robots and Kindness’. Up until this point I had only studied philosophy (I had a particular interest in *logic* and *the philosophy of mind*) and a bit of psychology, and knew I wanted to apply my experience to solving problems in Artificial Intelligence but didn’t know how. When I saw the project scope asking for a philosopher to look at how we might create *kind robots*, I thought the project was ambitious and crazy enough for me to try out. Little did I know that it would become such a big part of my life, and that 7 years later I would be writing up my ideas into a book, and that it would ultimately give me a whole new career in Robotics.

I was fortunate enough to be given the chance to spend my time working on this area, therefore, after completion of my PhD I decided I wanted to write a book—using the same material—but for a wide audience, so that everyone can enjoy thinking about questions in this area. Though a significant number of the later chapters represent my own thinking on how we might create moral machines, a large portion of the book is devoted to introducing those unfamiliar to the area to the concepts and language to be able to have informed conversations about it. It is my belief that science benefits from a diversity of ideas, therefore, by opening up this area to as many people as possible, science will benefit, and we can move closer (together) to solving one of society’s most pressing and interesting challenges.

I would have not succeeded in the past 7 years, and in creating this book, if it were not for the individuals that have supported me on this journey, whether that be academically, personally or in terms of offering more practical support during periods of ill health. First and foremost, I want to thank my family (that’s my mum, Gail, brother, Mathew, and sister, Toni) who have provided critical support during the hardest times, and who have endured many dinnertime conversations about *The Trolley Problem*, *Moral Agency* and *Robot Rights*, particularly my mum for reading through the drafts of this book and for giving me ‘non expert’ insight into whether my writing was understandable to a broad audience. I want to thank my PhD supervisors, Nigel Crook and Matthias Rolf, who introduced me to this topic and also

endured many heated debates surrounding how we might create robots with morals, alongside the Machine Ethics reading group at Oxford Brookes University and other colleagues who have contributed to my thinking in this area, with special mention to Oliver Bridge, Nicola Strong and Phil Harvey for their out of hours discussions. Also, my PhD examiners, David Gunkel and Alan Winfield were invaluable in rigorously dissecting my PhD thesis and providing insight to help shape my ideas to what they eventually are in this book. Finally, I want to thank Springer Nature and the publishing team for sponsoring this project, particularly Susan Grove and Arun Siva Shanmugam for answering many questions as I embarked on this writing journey.

Coventry, UK

Rebecca Raper

Contents

- 1 Introduction: *The Pursuit for Moral Machines* 1**
 - Reference 5
- 2 Background: ‘Morals’ and ‘Machines’ 7**
 - 2.1 Morals 7
 - 2.2 Machines 12
 - 2.3 Moral Machines 15
 - References. 17
- 3 A Survey of Machine Ethics. 19**
 - 3.1 Ethical Machines vs Ethical Decision Machines. 20
 - 3.2 Building Morality into Machines. 21
 - 3.3 Moral Machines Behaving Badly?. 26
 - 3.4 Testing Morality. 28
 - 3.5 AI Alignment 30
 - References. 32
- 4 Why We Need Moral Machines 35**
 - 4.1 Social Moral Machines 36
 - 4.2 Unpicking Machine Ethics. 39
 - 4.3 Assuring Moral AI. 41
 - References. 42
- 5 A Framework and Approach 43**
 - 5.1 The Problem Statement 44
 - 5.2 Enabling Moral Agency (Rather Than Constraining Immoral Behaviour) 46
 - 5.3 Moral Cognitive Requirements 48
 - 5.4 Testing Morality. 50
 - Reference 51