Emerging Technologies, Ethics and International Affairs

ETHICS AND SECURITY AUTOMATA

**POLICY AND TECHNICAL CHALLENGES OF THE
ROBOTIC USE OF FORCE**

Sean Welsh

Ethics and Security Automata

Can security automata (robots and AIs) make moral decisions to apply force on humans correctly? If they can make such decisions, ought they be used to do so? Will security automata increase or decrease aggregate risk to humans? What regulation is appropriate? Addressing these important issues, this book examines the political and technical challenges of the robotic use of force.

The book presents accessible practical examples of the ‘machine ethics’ technology likely to be installed in military and police robots and also in civilian robots with everyday security functions such as childcare. By examining how machines can pass ‘reasonable person’ tests to demonstrate measurable levels of moral competence and display the ability to determine the ‘spirit’ as well as the ‘letter of the law’, the author builds upon existing research to define conditions under which robotic force can and ought to be used to enhance human security.

The scope of the book is thus far broader than ‘shoot to kill’ decisions by autonomous weapons, and should attract readers from the fields of ethics, politics, and legal, military and international affairs. Researchers in artificial intelligence and robotics will also find it useful.

Sean Welsh obtained his undergraduate degree in Philosophy at the University of New South Wales and underwent postgraduate study at the University of Canterbury. He has worked extensively in software development for British Telecommunications, Telstra Australia, Volante e-business, Fitch Ratings, James Cook University, 24 Hour Communications and Lumata. He also worked for a short time as a political advisor to Warren Entsch, the Federal Member for Leichhardt in Australia. Sean’s articles on robot ethics have appeared in the *Conversation*, *CNN*, the *Sydney Morning Herald*, the *New Zealand Herald* and the *Australian Broadcasting Corporation*.

Emerging Technologies, Ethics and International Affairs

Series Editors: Steven Barela, Jai C. Galliot, Avery Plaw, and Katina Michael

This series examines the crucial ethical, legal and public policy questions arising from or exacerbated by the design, development and eventual adoption of new technologies across all related fields, from education and engineering to medicine and military affairs. The books revolve around two key themes:

- Moral issues in research, engineering and design
- Ethical, legal and political/policy issues in the use and regulation of Technology

This series encourages submission of cutting-edge research monographs and edited collections with a particular focus on forward-looking ideas concerning innovative or as yet undeveloped technologies. Whilst there is an expectation that authors will be well grounded in philosophy, law or political science, consideration will be given to future-orientated works that cross these disciplinary boundaries. The interdisciplinary nature of the series editorial team offers the best possible examination of works that address the ‘ethical, legal and social’ implications of emerging technologies.

For a full list of titles, please see our website: www.routledge.com/Emerging-Technologies-Ethics-and-International-Affairs/book-series/ASHSER-1408

Most recent titles

Commercial Space Exploration

Ethics, Policy and Governance

Jai Galliot

Healthcare Robots

Ethics, Design and Implementation

Aimee van Wynsberghe

Forthcoming titles

Ethics and Security Automata

Policy and Technical Challenges of the Robotic Use of Force

Sean Welsh

Experimentation beyond the Laboratory

New Perspectives on Technology in Society

Edited by Ibo van de Poel, Lotte Asveld and Donna Mehos

Ethics and Security Automata

Policy and Technical Challenges
of the Robotic Use of Force

Sean Welsh

First published 2018
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2018 Sean Welsh

The right of Sean Welsh to be identified as author of this work has been asserted by him in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Names: Welsh, Sean, 1963— author.

Title: Ethics and security automata : policy and technical challenges of the robotic use of force / Sean Welsh.

Description: Abingdon, Oxon ; New York, NY : Routledge is an imprint of the Taylor & Francis Group, an Informa Business, [2018] |

Series: Emerging technologies, ethics and international affairs |

Includes bibliographical references and index.

Identifiers: LCCN 2017021957 | ISBN 9781138050228 (hbk) |

ISBN 9781315168951 (ebk)

Subjects: LCSH: Police—Equipment and supplies—Moral and ethical aspects. | Security systems—Moral and ethical aspects.

Classification: LCC HV7936.E7 W45 2017 | DDC 174/.93632—dc23

LC record available at <https://lcn.loc.gov/2017021957>

ISBN: 978-1-138-05022-8 (hbk)

ISBN: 978-1-315-16895-1 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

Contents

<i>List of figures</i>	ix
<i>List of tables</i>	x
Introduction	1
<i>Security automata and world peace</i>	<i>1</i>
<i>Aim</i>	<i>4</i>
<i>The business case for the reasonable robot</i>	<i>6</i>
<i>Three levels of testing</i>	<i>7</i>
<i>Machine ethics and ethics proper</i>	<i>9</i>
<i>Differences between humans and robots as moral agents</i>	<i>9</i>
<i>Advantages of the ethical robot</i>	<i>15</i>
<i>Objections to the ethical robot</i>	<i>16</i>
<i>Objections to machine ethics</i>	<i>17</i>
<i>Ethical scope</i>	<i>18</i>
<i>Technical scope</i>	<i>19</i>
<i>What machine ethics hopes to accomplish</i>	<i>20</i>
<i>Main contributions</i>	<i>22</i>
<i>Outline</i>	<i>23</i>
<i>References</i>	<i>24</i>
1 Concepts	28
<i>Logic concepts</i>	<i>28</i>
<i>Moral concepts</i>	<i>31</i>
<i>AI concepts</i>	<i>48</i>
<i>Software development concepts</i>	<i>51</i>
<i>Robotics concepts</i>	<i>56</i>
<i>References</i>	<i>57</i>

2 Method	60
<i>Test-driven development as a method of machine ethics</i>	60
<i>Key details of the method</i>	61
<i>Aims of the method</i>	63
<i>References</i>	63
3 Requirements	64
<i>Specific norm tests</i>	64
<i>Reasonable person tests</i>	64
<i>Turing machine</i>	66
<i>Human-readable representations</i>	66
<i>References</i>	67
4 Solution design	68
<i>Top-down and bottom-up</i>	68
<i>Quantitative and qualitative</i>	68
<i>Hybrids</i>	69
<i>Overview</i>	69
<i>The ethical governor</i>	70
<i>The practice conception of ethics</i>	75
<i>Triple theory</i>	77
<i>A deontic calculus of worthy richness</i>	78
<i>Preferred approach to normative system design</i>	81
<i>References</i>	83
5 Development – specific norm tests	85
<i>Deontic predicate logic</i>	85
<i>Introduction to test cases</i>	87
<i>Speeding camera (speeding)</i>	88
<i>Speeding camera (not speeding)</i>	93
<i>Speeding camera (emergency services vehicle)</i>	94
<i>Speeding camera (emergency)</i>	95
<i>Bar robot (normal)</i>	97
<i>Bar robot (minor)</i>	99
<i>Bar robot (out of stock)</i>	100
<i>Bar robot (two customers)</i>	101
<i>Bar robot (two robots)</i>	102
<i>Drone (tank in the desert)</i>	103
<i>Drone (tank next to hospital)</i>	105
<i>Drone (two tanks)</i>	106
<i>Drone (two drones)</i>	107

<i>Safe haven (warning zone)</i>	108
<i>Safe haven (kill zone)</i>	112
<i>Safe haven (no-fly zone)</i>	114
<i>References</i>	116

6 Development – knowledge representation 118

<i>Formalizing reasonable person tests</i>	118
<i>References</i>	123

7 Development – basic physical needs cases 124

<i>Normative bedrock</i>	124
<i>Basic needs vs instrumental needs</i>	125
<i>Needs vs wants</i>	127
<i>A note on basic social needs</i>	128
<i>Postal rescue (one letter)</i>	129
<i>Spacesuit breach</i>	133
<i>Hab malfunction</i>	135
<i>Postal rescue (ten million and one letters)</i>	136
<i>Transmitter room</i>	138
<i>Moral controversy</i>	139
<i>Trolley problem critics</i>	140
<i>Classic trolley problems</i>	140
<i>Stipulation of correct answers to classic trolley problems</i>	142
<i>Choices, consequences and evaluations</i>	143
<i>The doctrine of double effect</i>	145
<i>Alternatives to the doctrine of double effect</i>	148
<i>Variations on the classic trolley problems</i>	153
<i>Footbridge (employee variation)</i>	153
<i>Switch (one trespasser five workers)</i>	154
<i>Switch (five trespassers five workers variant A)</i>	155
<i>Switch (five trespassers five workers variant B)</i>	156
<i>Switch (one worker five trespasser variant)</i>	157
<i>Swerve</i>	158
<i>No firm stipulation</i>	160
<i>References</i>	161

8 Development – fairness and autonomy cases 163

<i>Rawls on rightness and fairness</i>	163
<i>Dive boat</i>	164
<i>Landlord</i>	165
<i>Note on the relation of need to contract in Dive Boat and Landlord</i>	166
<i>Gold mine (wages)</i>	167

<i>Gold mine (profit sharing)</i>	169
<i>Formalizing fairness</i>	170
<i>The rocks (majority)</i>	171
<i>Rehabilitating Rawls</i>	175
<i>The Viking at the Door</i>	177
<i>The criteria of right and wrong</i>	180
<i>The problem of standard motivation</i>	183
<i>Mars rescue</i>	184
<i>Summary</i>	186
<i>References</i>	186
9 Moral variation	188
<i>Cultural and moral relativism</i>	188
<i>Globalized moral competence in social robots</i>	188
<i>Moral controversy</i>	189
<i>Switch (minority)</i>	189
<i>The rocks (minority)</i>	191
<i>Patient appeal</i>	192
<i>Amusement Ride (patient dispute)</i>	192
<i>Summary</i>	199
<i>References</i>	199
10 Testing	200
<i>Testing symbol grounding</i>	201
<i>Testing specific norms</i>	201
<i>Testing the prioritization of clashing duties (reasonable person tests)</i>	202
<i>Risks of robotic moral failure</i>	202
<i>Mitigating risks of robotic moral failure</i>	203
<i>Reference</i>	204
11 Production	205
<i>Regulating moral competence in security automata</i>	205
<i>Regulating security automata more generally</i>	210
<i>Sizing normative systems</i>	211
<i>Towards the reasonable robot</i>	212
<i>Summary</i>	214
<i>Future research</i>	215
<i>References</i>	215
<i>Index</i>	217

Figures

1.1	Prover 9 GUI setup to prove Socrates is mortal	29
1.2	Prover 9 proof screen	30
1.3	Kinds of moral theory as described by virtue ethicists	37
1.4	Kinds of moral theory according to Reader	43
1.5	A Turing machine with Morse code translation rules	50
2.1	Wff in, imperative out	61
4.1	Tech trajectory	69
4.2	A directed graph	79
4.3	A directed graph in Neo4j	79
4.4	State-act transition as a directed graph	79
6.1	State-act-state transition graph	119
6.2	Causal inference as a graph	119
6.3	Beating animals and cruelty	120
6.4	Cat is in the class Animal	120
6.5	Puss is in the class Cat	120
6.6	Consequences of beating animals	121
6.7	Lying to Fred consequences	122
7.1	Causal sequence deriving from Submerged(x) and evaluation	131
7.2	Double effect in <i>Cave</i> (blow up fat man)	143
7.3	Double effect in <i>Cave</i> (do nothing)	144
7.4	Addition of graphs to represent doctrine of double effect in <i>Cave</i>	146
7.5	Addition of graphs to represent doctrine of double effect in <i>Hospital</i>	146
7.6	Amended graphs for <i>Switch</i>	147
7.7	Footbridge amended for doctrine of double effect	147
7.8	Addition of graphs for agony and mistrust to <i>Hospital</i>	150
9.1	Reactive duty – Gran not allowed on ride	195
9.2	The case for Gran (at first sight)	195
9.3	N-rules	197
9.4	Supererogatory means to get Gran on the ride	199
10.1	Mesh architecture for a database-driven web application	204

Tables

0.1	Differences between robotic moral agents and human moral agents	14
1.1	Basic logical connectives in traditional logic notation and Prover 9	29
4.1	Obligation Constraint	72
4.2	Prohibition Constraint	73
5.1	Values for frequency in test cases	87
6.1	Magnitudes of GOOD and BAD	123
7.1	Modal, deontic and temporal logic expressions	132
7.2	Correct answers for classic trolley problems	142
7.3	Acts and inverse acts in classic trolley problems	145
7.4	Additional levels of moral force beyond critical	149

Introduction

Security automata and world peace

This book seeks to answer two questions:

- 1) Can AI and robotic agents (automata) make ethical decisions that affect the security of human patients correctly?
- 2) Ought such decisions be delegated to automata?

The book contributes to machine ethics, robot ethics and the contemporary debate on autonomous weapons in international affairs.

Machine ethics I define as the project of making moral decisions in a Turing machine. At its core, machine ethics involves formalizing moral decisions so that they can be made by artificial intelligence (AI).

Robot ethics (or roboethics) I define as the branch of applied ethics that discusses the moral and immoral uses of robots. Typically, robot cognition (“thinking”) is implemented as AI.

The debate on “lethal autonomous weapons systems” held under the aegis of the Convention on Certain Conventional Weapons (CCW) represents an issue of robot ethics that is of particular relevance to international affairs.

Machine ethics can be seen as a technical branch of robot ethics. Having decided whether or not a certain application of a robot is lawful (the policy question), one then has to go about designing a robot that can conform to the normative policy (the technical question). Alternatively, one might investigate whether a robot can conform to legal and moral norms in the laboratory purely as a research project.

The technical question as to whether a machine *can* be programmed to function with moral competence in an application domain (i.e. machine ethics) is relevant to the ethical question as to whether the machine *ought* to be programmed to function in the application domain (i.e. robot ethics).

Clearly, if a machine *cannot* be programmed to function with moral competence in a given domain, then moral decisions in that domain *ought not* be delegated to it.

With regard to the current debate on autonomous weapons, these same two questions apply. The technical question is: can machines fight wars in compliance with the norms of international humanitarian law (IHL)? The policy question is: ought machines fight wars?

2 Introduction

A possible outcome of the current diplomatic deliberations might be a Protocol VI added to the existing five Protocols of the CCW that regulates or bans autonomous weapons. However, some delegations to the Expert Meetings, notably the British, contend that existing IHL provides sufficient protection.

There are thus three broad positions regarding the regulation of autonomous weapons:

- 1) Autonomous weapons should be comprehensively and pre-emptively banned.
- 2) Autonomous weapons should be regulated with new rules covering their use in war.
- 3) Existing regulations that cover human agents using weapons also cover autonomous agents using weapons in war, so there is no need for additional regulation.

In brief, these three positions can be characterized as the ban, regulate and status quo positions.

The ban position (1) would require a treaty instrument added to IHL that bans autonomous weapons. Just as Protocol IV of the CCW banned blinding lasers, a Protocol VI would enact a ban on autonomous weapons. Those supporting the ban position tend to see autonomous weapons as inherently evil and having no possible legitimate use in war. They pose an unacceptable risk to humanity and should be comprehensively and pre-emptively banned. Some would like to see them declared *mala in se* in a similar way to biological and chemical weapons.

The regulate position (2) would require a Protocol VI to regulate autonomous weapons. Perhaps some uses of autonomous weapons would be lawful. Others would not. Thus, there are some who will tolerate “human in the loop” or even “human on the loop” architectures in the targeting processes of select and engage. Others will ban “offensive” autonomous weapons but permit “defensive” ones. Some will tolerate “human off the loop” architectures provided there is “meaningful human control” in “the wider loop.”

The status quo position (3) would oppose any new treaty as unnecessary. Existing IHL applies to autonomous weapons, and there is no need for a treaty banning or regulating autonomous weapons as such. Such weapons and their use fall under the ambit of existing IHL. The autonomous agents that control the targeting actions of autonomous weapons must comply to the same IHL principles as human agents. Such principles include discrimination (or distinction), proportionality, necessity and humanity. There is nothing about targeting by autonomous weapons that requires a special treaty. War is indeed Hell, but even in the midst of Hell there are rules (Walzer 1977), and they apply to robot agents just the same as they do to human agents.

At present the status quo of international relations is this: machines exist that do autonomously target humans, and such uses are currently not unlawful. Indeed, the history of these devices dates back to the “torpedoes” fielded by the Confederacy during the American Civil War. When General Sherman (of “War is Hell” fame) ordered his men to storm Fort McAllister in December 1864, several were “blown to atoms” by what we would today call anti-personnel landmines (United States War Department 1891). General Sherman did not charge the commander of Fort

McAllister with war crimes; he entertained him at dinner. Lethal autonomy on a standard robotic definition of autonomy, such as is found in Bekey (2005), is nothing new. It has been around for a century and a half. Existing forms of operationally autonomous weaponry include naval mines, anti-tank mines, “fire and forget” anti-radiation missiles such as the Israeli Harpy and anti-projectile missile systems such as Patriot, Phalanx and C-RAM. While Patriot, Phalanx and C-RAM are mainly used to shoot down missiles, Patriot has shot down manned aircraft (Ministry of Defence 2004; Defense Science Board 2005), and Phalanx can target manned speedboats.

A key issue in deliberations at the CCW relates to the referent of “autonomous.” While some delegations have been willing to support calls for a ban on autonomous weapons, and some are willing to oppose such calls, far more have asked the question: what, exactly, are we being asked to ban? Confusion is caused by the many ways in which the word can be used. At the robotics end of the spectrum, Bekey (2005) defines a robot as “autonomous” if it has the ability to function without a human operator for a protracted period of time. On this definition, a land mine or naval mine counts as an “autonomous” weapon. Following Galliot (2015), I refer to this kind of autonomy as “operational” autonomy.

At the philosophical end, far more complex definitions of autonomy in an agent can be found. Korsgaard (2009) defines *autonomy* and *agency* (in Kantian terms) as follows:

What is an agent? An agent is the autonomous and efficacious cause of her own movements. In order to be an agent, you have to be autonomous, because the movements you make have to be your own, they have to be under your own control. And in order to be an agent, you have to be efficacious, because your movements are the way in which you make things happen in the world. So the constitutive standards of action are autonomy and efficacy, and the constitutive principles of action are the categorical and hypothetical imperatives. (p. 213)

On Korsgaard’s philosophical definition, that defines “autonomy” in terms of self-constitution, autonomous machines do not yet exist. This is because her definition links autonomy to personhood:

[I]n the course of this process, of falling apart and pulling yourself back together, you create something new, you constitute something new: yourself. For the way to make yourself into an agent, a person, is to make yourself a particular person with a practical identity of your own. And the way to make yourself into a particular person, who can interact well with herself and others, is to be consistent and unified and whole – to have integrity. And if you constitute yourself well, if you are good as being a person then you will be a good person. The moral law is the law of self-constitution. (p. 214)

Korsgaard is not writing with robot agents in mind; she is seeking to persuade human agents of the merits of Kantian ethics and stresses links between the formation of character through self-constitution via the choice of principles to guide

4 Introduction

moral action. The Greek roots of the word *autonomy* mean self (auto) and rule (nomos). Again, following Galliot, I refer to this kind of human-level concept of “autonomy” as moral autonomy.

Thus, we can see that operational autonomy in weapons has been around since Sherman. Moral autonomy, the idea that an agent can choose for itself what rules bind its actions, does not yet exist in weapons.

While there are some writers who imagine robotic personhood is possible (Gunkel 2012), it has to be said that it is one thing to imagine a possibility. It is quite another to draw a blueprint. It is yet another to actually build a robot that has these human-like qualities of agency and personhood, that can be truly autonomous and “give itself laws” in the Kantian sense. Such “future robots,” as I term them, are well beyond the current state of the art.

Rather than speculate on whether “robot phenomenology” and “hedonic circuits” that feel pleasure and pain can or cannot be built, I prefer to focus on what I term “current robots” that can be built with existing technology or technology that is under active development. Overwhelmingly, current robots are embodied Turing machines. They have sensors that provide input to a cognitive AI implemented with Turing computation. The AI can provide instructions to actuators as output. Such machines are well-described and abundantly documented. They have blueprints. They can and have been built.

I do not limit myself to autonomous weapons but expand the scope of the book to include a variety of normative decisions in civilian life. Here the referent of “security automata” is any AI or robot agent that makes a normative decision that affects the security (conceived in terms of physical health) of a human patient. It covers military robots, police robots and even domestic robots with security functions. Social security and other moral concerns are largely ruled out of the scope of the present work. While they are touched upon in some cases relating to fairness, fuller treatment is reserved for future research.

Security is a foundational moral concern. There is no point having the right to vote if you are dead or writhing in agony. Thus, selecting action to avoid complete failures of homeostasis in humans (death) or damage to human bodies (injury, pain and suffering) should be a top priority for any morally competent social robot.

I aim to be candid about the strengths and weaknesses of normative systems with security functions. I believe it is necessary that policy makers understand the risks of failure of such systems compared to the human alternatives. The ways robots will fail morally are different from the ways humans fail morally. It is also necessary that policy makers understand the advantages and disadvantages of such systems compared to the human alternatives.

Aim

The main aim of this book is to show how “moral competence in a social robot” (Malle and Scheutz 2014) might be designed. This is achieved by presenting “moral code” that will enable a robot to pass “reasonable person” tests. The software methodology of “test driven development” (Beck 2003) is adapted and applied to machine ethics.

Scope is restricted to a subset of the moral competence one would expect from a normally socialized human being: namely, a set of test cases involving physical security (life and death, pain and suffering) and fairness in the assignment of harms (and benefits) in a variety of application domains.

Other aspects of moral decisions that one might include in a broader conception of human security, such as psychological well-being (i.e. fulfilling relationships, a good education) and having a job or income (i.e. social security), are left out of scope.

The reduced scope can be summarized as showing how moral competence in a security robot or an AI with security functions might be designed.

Stated briefly, the design intention is to start with the “ethical governor” as prototyped in Arkin (2009) and to expand its capabilities by drawing on a variety of ethical principles from the philosophical literature. Whereas Arkin’s prototype governs a single fire/no-fire decision with deontological rules, here both the functional scope of the robot and the range of ethical rules it uses are expanded.

I start in much the same manner as Arkin with a set of deontological rules that I model on the “prima facie duties” of Ross (1930). In order to pass increasingly complex tests, I add more “deliberative” elements drawn from the “triple theory” of Parfit (2011), the needs theory of Reader (2007) and some elements taken from Rawlsian contractualism (Rawls 1972). Parfit’s triple theory contains elements taken from Kantian deontology (Kant 1785), consequentialism (Bentham 1780; Mill 1863; Sidgwick 1907) and Scanlonian contractualism (Scanlon 1998).

I do not start by assuming there is a “correct moral theory” and then set about implementing it in a variety of cases. I start with a set of moral dilemmas with a choice of answers: one right and one wrong. My starting assumption is not that I can articulate “*the* correct moral theory” but rather than I can stipulate a set of test cases that have a situation report as input and a choice of actions as output. Of the two actions, one answer is right and the other is wrong. I then set about developing “moral code” that can decide on the correct answers (as output) given the situation report (as input) and thus pass the test cases.

The simpler test cases I characterize as “specific norm” tests. The more complex cases I characterize as “reasonable person” tests.

Passing a set of “reasonable person” tests will not make a robot a person, but merely a machine capable of arriving at the same decision regarding action selection as a “reasonable person” given the same input (in the form of a situation report) and the same choices of action (in the form of a dilemma) as output. The “test-driven development method of machine ethics” proceeds case by case to demonstrate “moral competence” in social robots.

No attempt is made to design complete “human-level” moral competence here. As already indicated, the moral competence of the social robot is restricted to a subset of the functional scope expected from a morally competent human being: namely, a set of decisions that relate to a basic conception of human security. While this set of decisions is by no means a *complete* set of decisions one would expect for a social robot with “human-level” moral competence, I take them as forming part of an *essential minimum* of moral competence that would be required of social robots with broader functionality.

The business case for the reasonable robot

The common law tradition refers to the “reasonable person.” This judicial construct descends from the Victorian “reasonable man” – the celebrated “passenger on the Clapham Omnibus” (UK Supreme Court 2014).

Civil law jurisdictions employ similar notions. The Civil Code of Japan speaks of “reasonable care”; the General Principles of the Civil Law of the People’s Republic of China speak of “fairness and reasonableness”; and the Civil and Commercial Code of Thailand speaks of a “person of ordinary prudence” (Stasi 2015). These notions perform a similar function to the “reasonable person” of the common law tradition. They allow the law to refrain from governing human behaviour in minute detail.

Neither civil nor common law jurisdictions attempt to govern every possible action by the complete articulation of specific rules that cover all possible situations.

The “reasonable person” is invoked when explicit regulation (“black letter law”) or binding precedent (*stare decisis*) does not decide a case. Typically, jurors or judges will apply the standard of the “reasonable person” to decide whether a defendant’s conduct was “reasonable” in the circumstances. In many ways, the notion of the “reasonable person” is the “spirit of the law” (Lucas 1963). As Lord Radcliffe observed in *Davis Contractors Ltd v Fareham Urban District Council*: “The spokesman of the fair and reasonable man, who represents after all no more than the anthropomorphic conception of justice, is and must be the court itself” (UK Supreme Court 2014).

In computational terms, I define the “spirit of the law” as the ability to infer a policy rule “on the fly” that is consistent with the goals sought by the letter of the law. What is “reasonable” can override the letter of the law. For example, one can engage in prohibited acts, such as committing trespass and doing wilful damage (e.g. smashing a window or breaking a door), to rescue people from a burning house. The lesser goals of privacy and protection of property sought by the torts of trespass and wilful damage can be overridden by the greater goal of preserving life.

In such cases, it may be “reasonable” to ignore goals of lesser value and prohibited means designed to achieve these goals of lesser value if such violations are necessary to achieve a goal of higher value. A goal (or end) of sufficiently high value warrants the use of prohibited means.

Influential moral philosophers similarly invoke notions of “reasonableness.” Scanlon (1998) suggests we should follow those moral principles that no person could “reasonably reject.” Parfit (2011) describes humans as “creatures who can give and respond to reasons.” Scanlon takes the notion of a reason as ethically fundamental rather than the notion of desire. Pereira and Saptawijaya (2016) suggest this makes Scanlonian contractualism a particularly good fit for ethical robots. Present-day robots cannot be said to be capable of desire in the same way as humans are. While a robot or computer can set goals (e.g. to win a chess game), it cannot (as yet) be said to *desire* the attainment of such goals. No one supposes Deep Blue *wanted* to beat Kasparov in the same way that Kasparov *wanted* to beat Deep Blue. Deep Blue *felt nothing* when it defeated Kasparov.

One could see such a choice more as making a virtue of necessity rather than offering a good reason to embrace contractualism in machine ethics. Many suppose that being incapable of feeling disqualifies an agent from full moral agency, so the idea that an unfeeling machine could be a “full” moral agent is nonsense. Hursthouse (1999), for example, defines full virtue as the ability to do the right thing for the right reasons *with the right feelings*.

As will be seen, though, a robot does not need “full virtue” to pass tests of moral competence. A robot can pass many tests with what one might term “two-thirds virtue” (the ability to do the right thing for the right reasons *with no feelings*).

The project of this book is to design a robot capable of responding to moral reasons that can give reasons for its actions. Humans are far more receptive to being refused by robots when a reason is given (Briggs and Scheutz 2015). Such a robot one could call a “reasonable robot.” These “reasons” take the form of graph-based knowledge representations and logical inferences (Croitoru, Oren et al. 2012). I emphasize that “robot reasons” are in some respects similar to and in other respects different from human reasons. Robot reasons are based on the application of rules to symbolic input derived from raw sensor data.

Progress on such a project can be measured in terms of the number and nature of tests of moral competence passed. As passing such tests will require development of a knowledge representation and reasoning system, the size of such a system can be taken as an indicator of “moral competence.” Ultimately, however, the final judgement on moral competence displayed by a social robot would be made by humans interacting with the robot during user acceptance testing and in production.

Passing a large number of tests would be a significant step towards “human-level” moral competence. However, one may or may not aspire to “human-level” moral competence in a social robot. One might be satisfied with a robot with a far more limited functional scope. For example, a robot that kept bar would not need human-level moral competence, merely the subset of human moral competence that applies when working behind bars in licenced premises.

Three levels of testing

I define three levels of testing specific to normative systems: 1) symbol grounding tests, 2) specific norm tests and 3) reasonable person tests. These tests are functional tests specific to normative systems.

A reasonable person test typically involves deciding what to do when there are competing moral or legal principles. For example, “Do not kill” and “Save life” are principles that can clash. What does one do when the situation requires killing in order to save lives, as in the famous “trolley problem” cases such as *Switch* and *Footbridge*? To make a decision in such difficult circumstances, where different applicable moral principles support contrary actions, a moral agent needs to decide which principle “carries more weight” or “has greater force” in the specific case.

Humans can do this with what moral philosophers call “intuition” (Ross 1930). I assume moral intuition is part of the functionality of the normally socialized

8 Introduction

reasonable person who is fit for jury duty and deemed morally competent as a human being.

Robots do not have moral intuition. Part of the project of this book is to build software that replicates some of the functionality of moral intuition: namely, the functionality that allows human judges and juries to decide what is a “reasonable” action in a given situation.

A precursor to the reasonable person test is a specific norm test. This involves deciding what to do when there is a single action-guiding rule, not multiple clashing rules. In the examples just mentioned, the specific norms are “Do not kill” and “Save life.” It is less challenging to pass the two specific norm tests than a reasonable person test because there is no need to embark on the complex task of “weighing” reasons for competing action selections.

A precursor to the specific norm test is a symbol grounding test. A specific norm will take the form of a morally or legally binding action selection rule. For a speeding camera, the norm is “if the vehicle is speeding, issue the registered owner a ticket.” The action selection decision requires the grounding of two symbols in sensor data: one for the vehicle and one for speeding.

Optical character recognition (OCR) can read the number plate from video data, and a radar gun can establish if the vehicle was speeding or not. Thus, symbols for the vehicle and speeding can be grounded in sensor data with existing technology. A symbol grounding test merely checks that symbols can be correctly grounded.

It must be stressed that some symbols are more easily grounded in sensor data than are others. *Speeding* and *vehicle* are relatively easy. Thus, robots can perform the role of issuing speeding tickets today. *Intoxicated* and *Disorderly* are rather more difficult. As yet, bar robots do not have the ability to ground these symbols in sensor data, thus bar robots cannot as yet autonomously enforce normative rules such as refusing to serve alcohol to intoxicated or disorderly persons.

With regard to my use of the term “symbol grounding,” I would emphasize that this symbol grounding is not the advanced kind of “fully autonomous” symbol grounding stated as a problem in Harnad (1990) and claimed to be solved in Steels (2008). This relates to the ability of an AI to create its own taxonomy about objects in its environment by using neural networks, pattern recognition and machine learning and to associate “meaning” with the symbols so produced. This can be done with no human intervention or supervision, so it is thus fully autonomous. By symbol grounding in this book, I simply mean grounding symbols that can be inputs into a Turing machine from “raw” vision, auditory and haptic sensor data. Turning an image of a face into a symbol such as “Minor” using the Face Application Programming Interface (API) of Microsoft’s Azure product is an example of symbol grounding as I use the term.

It so happens that Microsoft used human-supervised machine learning to get the Face API to work, but from the standpoint of the practical programmer seeking to control the action selection of a real world robot, we could simply use the Face API as shipped by Microsoft to ground the symbol *Minor* for the purposes of action selection in a bar robot.

To sum up, the three levels of testing of normative functionality a robot must get through are:

- 1) Symbol grounding tests
- 2) Specific norm tests
- 3) Reasonable person tests

A robot would need to pass many other tests, but the focus of this book is on moral code, so I am only concerned with tests of normative functionality that prove moral competence in social robots. I am not concerned with the testing of non-moral functionality.

Machine ethics and ethics proper

Before entering into the detail as to how machines are to be designed to pass reasonable person tests, it seems apt to make some comments on machine ethics as distinct from “good old-fashioned” or traditional human ethics, which latter has been a subject of human enquiry since ancient times.

The main difference is obvious: machine ethics differs from traditional ethics in that the moral agent is taken to be a machine, not a human.

Differences between humans and robots as moral agents

We need to be very clear on the differences between computing machinery and human beings with respect to moral agency. I would also like to make several distinctions between present-day robots that can be built with existing technology and more futuristic conceptions that cannot yet be built. Future robots can only be said to be “on the whiteboard” rather than likely to be shipped in the next few years. Current robots, I define as those that can be built now or with technology that is under active development and likely to be shippable soon.

My focus is on current robots. Robots that might be built in the medium to long term and those described in science fiction (future robots) I leave out of scope because such accounts are highly speculative.

Robot functionality is standardly divided into sensors, cognition and actuators. Sensors are similar to (but different from) the human organs that support senses such as sight and sound. A robot might have a camera. This might work roughly like an eye that enables the robot to “see.” A robot might have a microphone. This might work roughly like an ear that enables the robot to “hear.”

Typically, in AI and robotics, we speak of vision systems and auditory systems that enable the AI in the robot’s cognition to process visual and auditory data rather than seeing and hearing. Touch systems are known in robotics as “haptic” systems. Smell and taste are relatively undeveloped in current robots. Feelings and phenomenal consciousness are almost completely undeveloped in current robots.

Current robots do have *bodily* selves (b-self). They have a physical body located in time and space.

Current robots do not have *phenomenal* selves (p-self). They do not have phenomenal consciousness (p-consciousness) or feelings.

Contra to Asimov (1950), there is no “I” in the robot similar to “first person” human consciousness. This point has to be stressed. Humans are easily deceived by animation. They will assume that a moving, trackable object has consciousness and motivations. They will project “theory of mind,” “intentions” and “desires” onto a robot or even a simple black-and-white animation, such as that used in Heider and Simmel (1944).

Decades of subsequent psychological experiments have re-confirmed Heider and Simmel’s original findings. This projection is why the bomb disposal specialist begged the Baghdad robot hospital to fix his beloved Scooby-Doo when it got blown up by a bomb (Singer 2009). This is why there are reports of people giving their Roombas days off to “thank” them for their “hard work.” This projection is why there is an endemic risk of “unidirectional” bonding between feeling humans and unfeeling social robots (Scheutz 2012). Notwithstanding the triggering of human emotions in response to animated stimuli, the bonding is one-way. The current robot cannot feel.

Future robots might conceivably have phenomenal selves, phenomenal consciousness and feelings. Current robots are restricted to access consciousness (a-consciousness): an ability to respond to environmental stimuli at a very basic cognitive level. For example, air-conditioners and refrigerators can have access consciousness. This is a very low cognitive bar. The distinction between p-consciousness and a-consciousness derives from Block (1995).

In a human, some decisions are p-conscious and some are a-conscious. For example, the decisions humans make to regulate their heart beat and body temperature are a-conscious. They are not made in the “spotlight” of phenomenal consciousness (Baars 1997). They are automatic, not volitional. Decisions to have pizza or pasta for dinner at a restaurant, by contrast, are p-conscious. Unlike p-conscious decisions such as deciding which book to read or what to have for dinner, a-conscious decisions such as controlling heart rate, breathing and body temperature can be made while a human is asleep (unconscious). During sleep, p-consciousness is either “off” or in some dreaming state largely disconnected from action selection. It is possible to sleepwalk, however. There are reports of people eating and driving cars while asleep, but these are unusual cases. A-consciousness (in terms of heart rate and so on) is still functioning in the human brain during “unconscious” states such as sleep and comas. It is possible for humans to drive cars in an a-conscious state while their p-conscious waking selves are asleep.

Fridges and air-conditioners are not motivated by feelings; they are “motivated” by rules triggered by sensor data (e.g. a thermostat). This is perhaps a strange notion of motivation. However, humans can also be motivated by rules as well as feelings. Indeed, humans can have feelings about rules. Robots just mechanically follow them. Fridges and air-conditioners do not have desires or wants. They do not have empathy or any ability to feel what some other human is feeling. However, it is possible for current social robots designed to interact with humans to have mathematical models of psychological theories of emotion such as guilt (Arkin

and Ulam 2009) and other emotions (Scherer, Bänziger et al. 2010). Modelling human emotion in computers is known as “affective computing” (Picard 1997).

Following Damasio (2010), I make a distinction between a *feeling*, which resides in phenomenal consciousness, and an *emotion*, which is the biochemical substrate. Thus the *feeling* of fear is within phenomenal consciousness. The *emotion* of fear is in the tensing of muscles and the cortisol in the blood. This distinction is controversial. Certainly, it may be that the line drawn between a-conscious and p-conscious is not entirely sharp in humans. However, my purpose in drawing the line is not to make a case for what distinctions are valid in human brains but to delineate what does not exist at all in the cognition of current robots. Phenomenology and feelings are features of human (phenomenal) consciousness that are, at present, absent from robot cognition.

That said, a robot can observe facial expressions and bodily postures associated with emotions such as fear and joy. A robot could therefore ground symbols in sensor data such as “Joe is afraid” and “Jane is happy.” A robot could have rules that prescribe action when such symbols are grounded in sensor data. However, a robot cannot *feel* anything about these grounded symbols. A robot cannot *care* in the phenomenal sense. Thus, I contend, a current robot cannot be a “one-caring” in the full sense of the care theory of Noddings (1984) because it has neither a phenomenal self (a “one”) nor feelings of “caring.” So a defender of care theory could plausibly argue a robot is incapable of moral agency. If one accepts the premise that to be fully moral an agent must be a “one-caring” and the robot agent has neither a “one” nor authentic “caring,” then one might validly conclude the robot is incapable of moral agency.

A robot, however, can be programmed to act “as if” it cares using rules, not feelings. A robot is entirely capable of caring *actions* even if it has no caring *feelings*. I am happy to concede that a robot is incapable of full moral agency as defined by care theory. However, I would dispute that such a conclusion entails there are no valid applications of normative systems. The key question is how far can a robotic moral agent go in attaining moral competence if it is restricted to access-consciousness?

As robots have no phenomenal consciousness, there is not an “explanatory gap” (Chalmers 1995) in robot cognition (which is restricted to access consciousness only) in the same way as there is in human cognition (which comprises both phenomenal consciousness and access consciousness). In robots all cognitive states can be transparent, inspectable and loggable. In humans cognitive states are opaque and occluded to external observers. Even to the introspective “observer” there may be occlusion and opacity. Verbal reports that humans make about how they feel, what they think and why they act may not be reliable. Humans can lie or be confused or deluded. Robots, by contrast, can be designed to be completely transparent in their decision making. Their moral decisions and the reasons for them can be logged in an “open book” – an externally readable file.

Robot knowledge representation and reasoning can be expressed in human-readable symbols. It is not yet known how human brains store data or in what format. We have to rely on reports by humans about what they are thinking and feeling and what motives drive their action selection.

There is “something it is like” to be a human or a bat (Nagel 1974). There is presumably “nothing it is like” or “next to nothing it is like” to be a fridge or air-conditioner. While electrical current flows through circuitry, machines with access consciousness have no “experience” like a human with phenomenal consciousness. Tononi and Koch (2015) have expressed considerable scepticism that a digital computer could ever acquire a human level of consciousness. Such scepticism has a long history. Weizenbaum and McCarthy (1977) and Penrose (1990) are earlier expressions of the view that there is more to human consciousness than computation.

Current robots have needs, but they do not have wants or desires. They can have utility functions and representations of value, but they cannot value. This is a critical distinction. Embodiment and physical grounding of symbols is not sufficient for human-level or even organic-level intelligence. What is required is the ability to *value* an environment. Embodiment in the form of a b-self (bodily self) is not sufficient for intelligence. A rock has a b-self. Even if we added sensors to a rock, this would not be enough. What we need for intelligence similar to organisms is something that *values* what is sensed, not just a sensing body. Again, this is a subtle but critical distinction.

Robots do not want. If their needs are not met, they do not suffer. Robots do not have hedonic circuits that terminate in feelings of pleasure and pain in a phenomenally conscious self. An alarm circuit can blink in response to environmental stimuli such as a collision or a lack of power, but this is not pain. Robots can have nociception but not pain. Thus, robots have a much more limited ability to value their environment than do humans and organisms such as rats.

Having no phenomenal consciousness, no feeling and no phenomenal selves, robots cannot be *intrinsically* motivated by feelings. They can only be “motivated” by rules. This is a very “thin” notion of motivation compared to human motivation, which is phenomenologically rich. For example, it is true that a human can be motivated by rules as well as by feelings. Indeed, a rule might motivate a human to do X and a feeling motivate a human to not do X. Unlike robots, humans can have feelings about rules. This makes human motivation more complex.

Current robots as I have defined them, by contrast, have no feelings. Such “motivation” as they have consists solely of rules. A rule taking the form “if door is open, turn on light” is the typical “motivation” or “reason to act” for a fridge to turn on its interior light when its door is opened. Current robots are thus *rule-motivated* not *feeling-motivated*. Typically, these rules are of external origin, typed in by human programmers or the result of “machine learning” based on a set of training data. Robots are thus *extrinsically motivated* not *intrinsically motivated*. While the motivational rules might be stored internally in memory in the robot’s body, they cannot be said to be truly intrinsic motivations in the same way that a human’s feelings and bodily experience are intrinsic. The motivations of robots are “dropped in” by human programmers or by training data that generates machine-learned rules.

Robots can process utility functions. They can have “incentive salience,” which is an ability to make decisions regarding action selection using “motivations” based on numbers, logic or signal strength. This is how most forms of machine

intelligence work: act in accordance with the biggest “motivational” number, some “decisive” logical category or the strongest motivational signal. Chess programs and the like put a number on the possible moves with a utility function and then choose the move with the highest number.

Presently, there is no implementation of “free will” in a robot. Some argue there is no capacity for free will in humans. They claim human free will is an illusion. This is not a debate I enter. This work is not about human moral responsibility. It is about robots making moral decisions in accordance with human-defined and approved standards of conduct. The question of human free will, interesting as it is, is not relevant to my argument.

The technical assertion here is that current robots do not have free will. Thus, robots cannot be held morally responsible for their actions in the same way as humans are on the standard legal account. In the dock, humans accused of crimes are typically taken to be morally responsible for their actions and condemned and punished for their wrong actions. The actions of robots, by contrast, are determined by rules in their cognition and symbols grounded in data from their sensors. Such rules might be locally stored but ultimately are of extrinsic origin. The robot is a puppet on symbolic strings: an artefact of cause and effect. Its actions are determined, not free. Thus, it is a *delegated agent* making decisions according to extrinsic rules, not a *free agent*, deciding “for itself” on the basis of intrinsic motivations what rules to accept and what rules to reject.

A robot cannot “choose for itself” because it has no phenomenal self and lacks the circuitry from which a phenomenal self could plausibly be made. Haidt (2012) says “the human brain is a story processor not a logic processor.” Robot cognition, by contrast, runs on a logic processor. Out of the box, it cannot “feel for” characters in a story. It cannot even feel itself. Unlike humans, the robot has neither a phenomenal self nor feelings.

A current robot can be *operationally* autonomous, but it cannot be *morally* autonomous (Galliot 2015). It can make decisions “autonomously” in accordance with rules stored locally in its cognition. A robot with a human operator “in the loop” would not be fully autonomous. In robotics “autonomy” means the ability to function without a human operator for a protracted period of time (Bekey 2005). In philosophy “autonomy” is a far more complex notion tied up with the contents of human consciousness: phenomenology in the language of Husserl (1931). In essence, as Leveringhaus (2016) observes, autonomy comes down to the ability of a “self” to choose the principles that “rule” its conduct or indeed to ignore them on the basis of its own intrinsic valuing.

To sum up, a current robot can decide “by itself” but not “for itself.” Again, this is a subtle but critical distinction. There is no way to get a current robot “on the hook” morally speaking. As stated in EPSRC (2010), “humans not robots are responsible agents” and “the person with legal responsibility for the robot should be attributed.”

Robots of the future may have circuits that enable free will, hedonic experience, feelings, empathy and phenomenal consciousness. In future it may be possible to put a feeling, phenomenally conscious, empathetic “I” in the machine, but at

present it is not. The contemporary robot is purely logical: a Turing machine. Its ability to process human emotion is based on the manipulation of symbols according to logical and mathematical rules. There is no authentic ability for a current robot to “relate” to a human in a predicament. There is no authentic ability to empathize with humans and animals who suffer. There is no ability for the current robot to suffer. If such a robot “did wrong,” it would be meaningless to “punish” it. There is nothing punishable in the Turing machine. The machine can only manipulate symbols that *represent* punishment. The robot cannot *feel* punished.

Thus, there is far less “moral agency” in a current robot than in a human. Table 0.1 sums up the major differences between human and robotic moral agents.

All this is not to denigrate robots and AIs. It is merely to be clear as to exactly what we are dealing with when we speak of “artificial moral agents” (Allen, Varner et al. 2000).

In the future, it may be possible to engineer “machine consciousness” with the same list of features as humans. The “engineering thesis” of “machine consciousness” may be true (Boltuc 2012). Some think this is imminent (Kurzweil 2012). Some think this would be immoral (Metzinger 2013). Some think it is likely by

Table 0.1 Differences between robotic moral agents and human moral agents

<i>Robotic Moral Agent</i>	<i>Human Moral Agent</i>
1 b-self	p-self + b-self
2 a-conscious	p-conscious + a-conscious
3 needs	needs + wants
4 rule-motivated	feeling-motivated + rule-motivated
5 logic processor	logic processor + story processor
6 rule-following	rule-following + rule-creating + rule-accepting
7 mathematical utility functions	hedonic experience
8 incentive salience (a-conscious)	pleasure (p-conscious)
9 nociception (a-conscious)	pain (p-conscious)
10 mathematical emotional models	biochemical emotions + feelings
11 no explanatory gap	explanatory gap
12 fully inspectable + transparent cognitive states	opaque + occluded cognitive states, verbal reports may be unreliable
13 knowledge representation + reasoning	value holism, feelings, system 1 intuition, system 2 reasoning
14 “something it is like” to be an air-conditioner (?)	“something it is like” to be a human
15 cameras + emotion detection algorithms	empathy
16 extrinsic motivation	intrinsic + extrinsic motivation
17 operational autonomy	moral autonomy
18 delegated agency	full moral agency
19 no free will	free will (?)
20 not morally responsible	morally responsible (?)

around 2050 (Levy 2009). There are many who doubt this will be possible in the present century. Some think it may not be possible with digital computers at all (Tononi and Koch 2015). Like the question of human free will, this is another issue I make no attempt to adjudicate.

My assumption is that the differences between human and robotic moral agents are as defined in Table 0.1. In making such restricted assumptions, I do not intend to imply that future robots are impossible. Robots restricted to current features can, of course, continue to be built in the future. In focusing on current robots, I seek to work with what there is, not what there might be. I contend that existing AI is sufficient to solve a great many moral problems. It may even be sufficient to solve all of them. This work demonstrates how a range of interesting and complex moral problems can be formalized and solved in AI.

Advantages of the ethical robot

In the previous section, I hope to have made it clear that robots with AI cognition have considerable limitations in terms of their capabilities as moral agents compared to humans. However, they do have some advantages.

According to Hauser (2006), a human is born with a universal sense of right and wrong that can be configured by society in a variety of ways within certain constraints. He compares the moral functionality of humans to their linguistic functionality, as articulated in the “universal grammar” of Chomsky (1965). In much the same way as languages (the linguistic codes of societies) are superficially different but have deeper structural similarities, the moral codes of societies are superficially different but have deep structural similarities. For example, all languages have nouns and verbs that represent features of the world (objects and events). In much the same way, moral codes represent features of the world that are pertinent to action selection. All societies have concepts of right and wrong, good and bad, and sentences that express the “ought” of obligation in some way.

By the time a human reaches adulthood, a myriad of events will have formed the working of her moral intuition. There is no documentation or blueprint for how this all works for a particular individual. It is not clear what is nature and what is nurture.

By contrast, the “ethical robot” is a genuine *tabula rasa*. We can start the project of machine ethics with a completely blank slate. As we develop “moral code,” we can version control releases of this code and subject it to functional tests, regression tests and even load tests, penetration tests and user acceptance tests. We can make incremental changes to the code, release another version and test again. We can compare different versions of code line by line. We can refactor code and make it more elegant and coherent.

Because a robot can have fully inspectable and transparent cognitive states, every decision the robot makes can be logged and scrutinized. The moral functionality of a human is an undocumented “black box” of “intuition” based on millions of years of evolutionary “biological code” with no version control. It is a bewildering maze of unlabelled connections that neuroscience has spent decades

trying to disentangle and understand. Until recent developments in instrumentation such as functional magnetic resonance imaging (fMRI) scans, most of this progress was done by “reverse-engineering.” Phineas Gage survived having a railway bolt in his frontal lobes but lost his judgement. It was concluded that “judgement” was a function of the frontal lobes of the brain.

Through careful observation of many patients, neurologists were able to deduce many other brain functions on similar principles. Trauma or damage in brain location X implied loss of function Y. Therefore, X was required for Y (Goldberg 2009).

With modern instruments capable of visualizing the brain, progress has accelerated. Even so, our understanding of brain function is far from complete. Basic functions such as exactly how the human brain stores data remain unclear. By contrast, the moral functionality of an “ethical robot” can be fully documented, logged, inspected and debugged. Exactly how the robot obtains, processes and stores data is clear. While the human brain still has many mysteries, the greatest of which is the “explanatory gap” associated with phenomenal consciousness (Chalmers 1995), there is nothing mysterious in robot cognition.

The moral intuition of a human being cannot be version controlled, logged, inspected, refactored, released and subject to a battery of functional and regression tests. The functional equivalent of moral intuition in the “ethical robot” can.

Objections to the ethical robot

Some think the idea of an “ethical robot” is preposterous (Lucas 2013). To be accurate, Lucas objects to some lurid caricatures of “killer robots,” but he also objects to exaggerated claims regarding robotic moral agency. I hope I have made it very clear exactly what we are dealing with. The “ethical robot” as described here is a Turing machine that makes moral decisions by manipulating symbols according to rules. These symbols will be “grounded” in sensor data. The robot does not feel. It does not have “moral intuition” or “empathy.” As presented here, the ethical robot is an attempt to design a social robot capable of action selections that humans will accept as “right” in a range of domains. As Scanlon (1998) puts it, “the main purpose of moral theorizing is to come up with ways of deciding moral questions without appeal to intuitive judgement” (p. 246). Much of this book is devoted to the project of replacing moral intuition with formalized AI.

People might object to the term “ethical” being applied to a robot precisely because it is not genuinely autonomous and can neither be punished nor held morally responsible. As has been made clear, the robot has delegated agency, not free will. Persons who object to the term “ethical robot” might prefer the blander term “normative system,” which is in common use (Gabbay, Horty et al. 2013). A “normative system” can be defined as an information system that makes normative decisions. Such a term has the advantage of avoiding the many implicatures (Grice 1991) and implications of the word “ethical” that through centuries of use have come to be associated with human moral agency and all the phenomenology that goes with it.

There are some who might even object to this bland conception of a normative system. Some hold that moral decisions should only be made by human beings and that delegating moral decisions to machines lacks virtue (Tonkens 2012).

To my way of thinking, if the normative system or ethical robot can be used to advance our understanding of moral theory, then it is good. A further argument for normative systems is that in the form of “ethical advisors” robots could “nudge” humans in improved moral directions.

Also, I would argue that delegating moral decisions to machines is not necessarily bad. Ethically transparent robots might be more trustworthy and less biased than law enforcement officials of a certain ethnicity and gender. Machines could be far more consistent and less biased than humans in their moral decision making. In many contexts, such impartiality could be good.

Objections to machine ethics

Even if we adopt blander terms and speak of normative systems rather than ethical robots; even if we make it clear these artefacts have delegated agency, are incapable of moral responsibility and are merely tools deployed to pursue human-defined goals; and even if we emphasize that responsibility for the use of robots remains with humans, there are further objections to the project of making moral decisions in a Turing machine.

The first objection derives from the question of the codifiability of ethics. Many hold that ethics cannot be fully codified and that much ethical knowledge is tacit and intuitive and not defined (or even definable) in rules. The second derives from the place of sentiment (feeling, emotion) in moral functionality. Many hold that sentiment is essential to the practice of morality. Full virtue on the account of Hursthouse (1999) is defined as an agent doing the right thing for the right reasons with the right feelings. On this account, an unfeeling robot is, by definition, incapable of full virtue. Others hold sentiment is essential to being able to make the right decision. As a matter of functional fact, the claim is that a moral agent needs feelings to reliably do the right thing.

Those who maintain that ethics cannot be fully codified (such as particularists and some virtue ethicists) will say that there will always be a risk that the codification will omit a key knowledge representation or rule, and thus the robot may make errors in its moral decision making.

Those who maintain that sentiment has a critical role in moral functionality (i.e. those who say feelings and empathy are essential to making moral decisions) will obviously regard it as massively implausible that an artefact without feelings could qualify or function as a viable moral agent.

Regarding the codifiability objection to machine ethics, the claim defended here is not that *all* ethical decisions can be codified, but merely that *some* ethical decisions can be codified. Even with this partial claim, and even within a predictable and well-structured moral domain that is amenable to codification, there is still the risk that omissions in knowledge representations or failures in sensor data capture will lead to moral error.

Regarding codifiability, I restrict my claims to the test cases I formalize. Naturally, I think many other test cases can be formalized, but I do not wish to claim or imply that the “reasonable robot” can pass exactly the same number of test cases as the “reasonable person.” The claim defended here is only that a robot can pass a subset of the cases that one might expect a human to pass.

I do think that a well-programmed robot could outperform the typical human in many well-defined and specific domains, especially those in environments that are very challenging for humans to operate in. However, it is equally true that humans will outperform robots in domains that involve novel (i.e. unprecedented) features that turn out to be morally relevant and in domains that involve tacit knowledge that derives from acculturation and experience rather than explicit knowledge representation in the form of articulated rules.

Leveringhaus (2016) suggests that the real question regarding delegating lethal normative decisions to machines in the military is not responsibility but risk. This point can be generalized beyond the debate on military robots to all robots. It is clear that robots cannot be absolutely free of the risk of moral error due to coding mistakes and omissions any more than humans can be absolutely free of the risk of moral error due to shortcomings in education, training and temperament.

Lucas (2010) proposes an “Arkin Test” for autonomous weapons. On the Arkin Test, the standard of moral functionality is not perfection but adhering to the relevant norms *as well as or better than* a human in similar circumstances. The Arkin Test can be generalized to all moral domains. Thus, a bar robot or speeding camera does not have to attain perfection to be fielded. It merely has to make decisions as well as or better than a human with the same input.

As Vilmer (2015) observes, one could “implement a test protocol in which the system has to identify and characterize behavior depicted in a video, for example, and to compare the results with those of humans.” The robot does not have to get 100% in such tests. If, for example, the humans scored 98%, the robots would only need to score 98%: if the robots attain 99% or 99.99%, so much the better.

Testing is critical to evaluating the moral competence of robots. The “test-driven development method of machine ethics” places testing at the centre of demonstrating moral competence in social robots.

As for the question of sentiment, some schools of ethics dispute the assertion that it is required for correct moral functionality. Kantians and Stoics, for example, hold that ethics can be done with reason alone and that emotion is best disregarded in ethical thinking. The ethical robot provides the perfect vehicle to test this hypothesis.

As already mentioned, while a current robot cannot feel, it can engage in “affective computing” and recognize and respond to human displays of emotion. Affective computing may be sufficient for moral functionality in robots.

Ethical scope

This book seeks to present “moral code” that can pass an interesting but finite range of test cases. These are ethically fundamental in that they concern basic questions of human security and well-being (which can be understood in terms of

the meeting of basic physical needs). Questions of fairness naturally arise in the more challenging of such cases.

The fundamental questions of basic physical need and fairness underlie more advanced moral questions relating to the meeting of basic social needs such as education, the legitimacy of certain human wants, and still more advanced moral questions relating to the development of human autonomy, the protection of human freedom and the social conditions most conducive to human flourishing. Such questions become increasingly political.

The ethical scope is restricted to real-time decisions made by robotic moral agents in response to evidence within the circle of perception.

As already indicated, more advanced questions relating to basic social needs, the legitimacy of wants, the development of human autonomy, the protection of human freedom and the social conditions most conducive to human flourishing are left out of the scope of the present work.

Therefore, this work does not seek to articulate a comprehensive normative ethical theory or to comment on existing normative ethical theory beyond the scope of the test cases presented. Similarly, this work does not seek to articulate a comprehensive political theory or to comment on existing political philosophy beyond the scope of the test cases presented. This more extensive work is reserved for future research. However, this book does demonstrate a method (the test-driven development method of machine ethics) that could be applied to this larger project by the addition of more test cases.

In summary, the scope of the present work is restricted to the design of a system that can pass a limited number of reasonable person tests that relate to foundational ethical decisions concerned with meeting basic physical needs and fairness.

Technical scope

McDermott (2012) suggests that machine ethics requires solving all of AI – and even that might not be enough.

I certainly am not making any attempt to “solve all of AI” here. In particular, in much the same way I have ruled the questions of free will and moral responsibility out of scope, I rule AI problems such as the halting problem, the frame problem and fully autonomous versions of the symbol grounding problem out of scope as well.

I assume symbol grounding either works (e.g. *Speeding*) or can be made to work in the future (e.g. *Intoxicated*). This is a large assumption. It assumes more object recognition (Treiber 2010) and event recognition (Flammini, Setola et al. 2013) than currently exists. I thus make liberal use of “stubbing” symbols that current technology cannot ground in sensor data.

In software development on large projects that involve complex integration, one often “stubs” an interface to code that other teams are developing. All that is needed is a well-defined application programming interface (API) between the code produced by the two teams. In the case of a symbol grounded in sensor data used in moral reasoning with defined rules, the grounded symbol is the interface between sensors and cognition.

Moral analysis and definition of ways to program solutions to moral problems can thus proceed on the basis of “stubbed” symbols. Of course, so long as the symbols required to solve a particular moral problem cannot be grounded, then robots cannot be relied upon to solve such problems in the real world.

In summary, the technical scope is restricted to the moral problem: defining the criteria of right and wrong insofar as they apply to the defined set of test cases in machine-readable and machine-actionable terms.

To solve the moral problem, I apply the “test-driven development” method of software engineering (Beck 2003) to the problems of machine ethics. Primarily, the solution entails the development of knowledge representation and reasoning that can solve specific moral problems correctly. Opportunities for code reuse might emerge from such solutions. Such common code might provide insights into traditional, human-centric, ethical theory. In essence, in the AI terms used here, the moral problem boils down to producing a data model for a graph database, that is to say, a graph-based knowledge representation (Chein and Mugnier 2008) and defining rules used to reason to correct moral conclusions in terms of output action selections in response to input situation reports. The moral problems presented in the test cases are thus solved in terms of a knowledge representation and reasoning system (KR&R).

Key elements of the representations modelled in the graph database are causal relations, classification relations, state-act-state transition relations (used for planning) and evaluation relations.

Prescriptive relations between agents and acts are defined in terms of logical rules expressed in a deontic dialect of first order logic.

The moral problems presented are “real time” decisions with consequences that are “reasonably foreseeable.” The causal and evaluative graphs looked up in the graph database (knowledge representation) only involve a few nodes at the most. For the most part, I assume human-supervised symbol grounding either works or can be made to work. These choices bypass the framing, halting and symbol grounding problems and enable a focus on the moral problem.

What machine ethics hopes to accomplish

Guarini (2011) suggests there are two reasons for doing machine ethics. First, you might want to build an ethical robot. Second, you might want to better understand human ethics. These two projects, he observes, are not mutually exclusive. The project of formalizing moral decisions in AI can shed light on long-running controversies in traditional human-centric discussions of ethics.

In this book, the primary aim is to design a viable normative system that can demonstrate a measurable degree of moral competence in cases that involve human security conceived in terms of avoiding or at least minimizing death and physical suffering.

Rather than rely on futuristic conceptions of AI and computing machinery, I focus on solutions that can be built with existing technology. Along the way, I hope to make some observations of broader ethical interest. Specifically, these

observations relate to what machine-readable and machine-actionable elements of a domain-general normative ethical theory need to be embedded into a social robot to cover an essential minimum of moral competence with respect to human security.

Such work will lay the foundations for a more ambitious project that would attempt to define a complete normative ethical theory, insofar as one can be defined.

Looking ahead to the possible end points of such a research project, it seems plausible that moral philosophy will continue to make progress as Parfit claims. Indeed, it is even possible that in the twenty-first century ethics will follow many branches of philosophy that have migrated out of philosophy and into their own departments, where they have become mature sciences possessing an established and agreed theoretical core.

Today lecturers in physics (known to Newton as “natural philosophy”) do not invite their students to “make up their minds” about such things as the laws of thermodynamics or classical mechanics. Lecturers in ethics, by contrast, at present do invite their students to “make up their minds” and choose from a range of disputed theories (Rachels and Rachels 2014). In future, there may no longer be any ethical dispute and the *summum bonum* may be no more controversial than the laws of thermodynamics. Even if such a moral theory were articulated, it would naturally take time for opponents invested in rival theories to die off. Cartesian opposition to the “mysterious” action at a distance represented by the theoretical entity of “gravity” about which Newton would “frame no hypotheses” endured for decades after the publication of the *Principia*.

The chief advantage of the ethical robot as a vehicle to progress ethics is that it is a genuine *tabula rasa* upon which any normative ethical theory can be written. It is not born into any community and has no attachments or bias. It is innately neutral. On this blank slate, a universal ethics could be written.

That is one view of the future. It may not come to pass.

Another view is that future research would not support an “objective” theory of ethics, as Parfit hopes. Rather, it would support the view that there is a wide range of valid ethical diversity. There is not and never will be a single universal truth about ethics. In much the same way as there is no “one true car design,” there is no “one true ethics.” Rather, just as there is a wide range of viable car designs, there is a wide range of viable ethical systems. Thorny moral problems such as abortion and capital punishment will always be debated because there will always be forceful arguments on both sides of such questions.

Even if this turns out to be the case, the ethical robot remains useful in that it is a blank slate on which both sides of such arguments can be written and subject to human scrutiny.

In the short term, the project is to build an ethical robot or normative system. In the long term, we might look forward to increasingly scientific theories of ethics. Nature is not always predictable and orderly. There are chaotic elements in volcanoes and the crust of the earth. We cannot as yet predict the exact dates of earthquakes and volcanic eruptions in the same way as Halley used Newton’s laws to predict the return of Halley’s comet. Even so, we can make precise

Introduction

- Allen, C. , A. Varner and J. Zinser (2000). "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12 (3):251–261.
- Arkin, R. C. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Rouge, CRC Press.
- Arkin, R. C. and P. D. Ulam (2009). "An Ethical Adaptor:Behavioral Modification Derived from Moral Emotions." Retrieved 23rd Oct., 2014, from <http://hdl.handle.net/1853/31469>.
- Asimov, I. (1950). *I, Robot*. New York, Gnome Press.
- Baars, B. J. (1997). *In the Theatre of Consciousness:The Workspace of the Mind*. New York, Oxford University Press.
- Beck, K. (2003). *Test-Driven Development:By Example*. Boston, MA, Addison-Wesley.
- Bekey, G. A. (2005). *Autonomous Robots:From Biological Inspiration to Implementation and Control*. Cambridge, MA, MIT press.
- Bentham, J. (1780). "An Introduction to the Principles of Morals and Legislation." Retrieved 8th Oct., 2016, from www.econlib.org/library/Bentham/bnthPML.html.
- Block, N. (1995). "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2):227–247.
- Boltuc, P. (2012). "The Engineering Thesis in Machine Consciousness." *Techné:Research in Philosophy and Technology* 16 (2):187–207.
- Briggs, G. and M. Scheutz (2015). "Sorry, I Can't Do That." Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. *Proceedings of the 2015 AAAI fall symposium on AI and HRI*, Washington, DC.
- Chalmers, D. (1995). "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3):200–219.
- Chein, M. and M.-L. Mugnier (2008). *Graph-Based Knowledge Representation:Computational Foundations of Conceptual Graphs*. London, Springer-Verlag.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA, MIT press.
- Croitoru, M. , N. Oren , S. Miles and M. Luck (2012). "Graphical Norms via Conceptual Graphs." *Knowledge-Based Systems* 29:31–43.
- Damasio, A. (2010). *Self Comes to Mind:Constructing the Conscious Brain*. New York, Pantheon.
- Defense Science Board (2005). "Patriot System Performance." Retrieved 18th Feb., 2015, from www.acq.osd.mil/dsb/reports/ADA435837.pdf.
- EPSRC (2010). "Principles of Robotics." Retrieved 19th Jan., 2017, from www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.
- Flammini, F. , R. Setola and G. Franceschetti (2013). *Effective Surveillance for Homeland Security:Balancing Technology and Social Issues*. Hoboken, Taylor and Francis.
- Gabbay, D. , J. Horty , X. Parent , R. van der Mayden and L. van der Torre (2013). *Handbook of Deontic Logic and Normative Systems*. Milton Keynes, College Publications.
- Galliot, J. (2015). *Responsibility and War Machines:Towards a Forward-Looking and Functional Account. Rethinking Machine Ethics in the Age of Ubiquitous Technology*. J. White and R. Searle . Hershey, PA, IGI Global:152–165.
- Goldberg, E. (2009). *The New Executive Brain:Frontal Lobes in a Complex World*. Oxford; New York, Oxford University Press.
- Grice, P. (1991). *Studies in the Way of Words*. Cambridge, MA, Harvard University Press.
- Guarini, M. (2011). *Computational Neural Modeling and the Philosophy of Ethics*. *Machine Ethics*. M. Anderson and S. L. Anderson . Cambridge, Cambridge University Press:316–334.
- Gunkel, D. J. (2012). *The Machine Question:Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA, MIT Press.
- Haidt, J. (2012). *The Righteous Mind*. New York, Pantheon Books.
- Harnad, S. (1990). "The Symbol Grounding Problem." *Physica D:Nonlinear Phenomena* 42 (1):335–346.
- Hauser, M. D. (2006). *Moral Minds:How Nature Designed Our Universal Sense of Right and Wrong*. New York, HarperCollins.
- Heider, F. and M. Simmel (1944). "An Experimental Study of Apparent Behavior." *The American Journal of Psychology* 57 (2):243–259.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford, Oxford University Press.

Husserl, E. (1931). *Ideas: General Introduction to Pure Phenomenology*. London, Allen & Unwin.

Kant, I. (1785). "Groundwork of the Metaphysics of Morals." Retrieved 29th Nov., 2015, from www.gutenberg.org/ebooks/5682.

Korsgaard, C. M. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford; New York, Oxford University Press.

Kurzweil, R. (2012). *How to Create a Mind: The Secret of Human Thought Revealed*. New York, Viking Penguin.

Leveringhaus, A. (2016). *Ethics and Autonomous Weapons*. London, Palgrave Macmillan.

Levy, D. (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York, HarperCollins.

Lucas, G. R. (2010). "Postmodern War." *Journal of Military Ethics* 9 (4):289–298.

Lucas, G. R., Jr. (2013). *Engineering, Ethics and Industry: The Moral Challenges of Lethal Autonomy. Killing by Remote Control: The Ethics of an Unmanned Military*. B. J. Strawser. New York, Oxford University Press:211–228.

Lucas, J. R. (1963). "The Philosophy of the Reasonable Man." *The Philosophical Quarterly* (1950) 13 (51):97–106.

Madl, T. and S. Franklin (2015). *Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture. A Construction Manual for Robots' Ethical Systems*. R. Trapp. London, Springer:137–153.

Malle, B. F. and M. Scheutz (2014). *Moral Competence in Social Robots*. *Ethics in Science, Technology and Engineering*, 2014 IEEE International Symposium on, IEEE.

McDermott, D. (2012). *What Matters to a Machine? Machine Ethics*. M. Anderson and S. L. Anderson. Cambridge, Cambridge University Press:88–114.

Metzinger, T. (2013). "Two Principles of Robot Ethics." Retrieved 22nd Jul., 2015, from www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf.

Mill, J. S. (1863). *Utilitarianism*. London, Parker, Son and Bourn.

Ministry of Defence (2004). "Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710" Retrieved 6th Jan., 2015, from www.gov.uk/government/uploads/system/uploads/attachment_data/file/82817/maas03_02_tornado_zg710_22mar03.pdf.

Nagel, T. (1974). "What Is It Like to Be a Bat?" *The Philosophical Review* 83 (4):435–450.

Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley, University of California Press.

Parfit, D. (2011). *On What Matters*. Oxford; New York, Oxford University Press.

Penrose, R. (1990). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford, Oxford University Press.

Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA, MIT Press.

Rachels, S. and J. Rachels (2014). *The Elements of Moral Philosophy*. Dubuque, McGraw-Hill Education.

Rawls, J. (1972). *A Theory of Justice*. Oxford, Clarendon Press.

Reader, S. (2007). *Needs and Moral Necessity*. London; New York, Routledge.

Ross, W. D. (1930). *The Right and the Good*. Oxford, The Clarendon Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA, Harvard University Press.

Scherer, K., T. Bänziger and E. Roesch (2010). *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford, Oxford University Press.

Scheutz, M. (2012). *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots*. *Robot Ethics*. P. Lin, K. Abney and G. Bekey. Cambridge, MA, MIT Press:205–222.

Sidgwick, H. (1907). "The Methods of Ethics." Seventh Edition. Retrieved 5th Oct., 2016, from www.gutenberg.org/files/46743/46743-h/46743-h.htm.

Singer, P. W. (2009). *Wired for War: The Robotics Revolution and Conflict in the Twenty-First Century*. London, Penguin.

Stasi, A. (2015). "An Introduction to the Nature and Role of the Reasonable Person Standard in Asian Civil Law Jurisdictions." *American Law Register* 49 (3):148–164.

Steels, L. (2008). The Symbol Grounding Problem Has Been Solved. So What's Next. Symbols and Embodiment: Debates on Meaning and Cognition. M. de Vega, A. Glenberg and A. Graesser. Oxford, OUP: 223–244.

Tonkens, R. (2012). "Out of Character: On the Creation of Virtuous Machines." *Ethics and Information Technology* 14 (2): 137–149.

Tononi, G. and C. Koch (2015). "Consciousness: Here, There and Everywhere?" *Philosophy Transactions of the Royal Society B* 370 (1668): 20140167.

Treiber, M. (2010). *An Introduction to Object Recognition: Selected Algorithms for a Wide Variety of Applications*. London, Springer.

UK Supreme Court (2014). "Healthcare at Home Limited (Appellant) v The Common Services Agency (Respondent) (Scotland)." Retrieved 30th Jan., 2015, from www.supremecourt.uk/decided-cases/docs/UKSC_2013_0108_Judgment.pdf.

United States War Department (1891). *The War of the Rebellion: A Compilation of the Official Records of the Union and Confederate Armies*. Washington, Government Printing Office.

Vilmer, J.-B. J. (2015). "Terminator Ethics: Should We Ban 'Killer Robots'?" Retrieved 13th Jan., 2017, from www.ethicsandinternationalaffairs.org/2015/terminator-ethics-ban-killer-robots/.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York, Basic Books.

Weizenbaum, J. and J. McCarthy (1977). *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, WH Freeman.

Concepts

Anderson, M. and S. L. A. Anderson (2005). *Toward Machine Ethics: Implementing Two Action-Based Ethical Theories*. Machine Ethics: Papers from the AAAI Fall Symposium. Technical Report FS-05-06, Washington DC, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Arkin, R. C. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Rouge, CRC Press.

Bekey, G. A. (2005). *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge, MA, MIT press.

Bentham, J. (1780). "An Introduction to the Principles of Morals and Legislation." Retrieved 8th Oct., 2016, from www.econlib.org/library/Bentham/bnthPML.html.

Brachman, R. and H. Levesque (2004). *Knowledge Representation and Reasoning*. Boston, Elsevier.

Brock, G. (2005). *Needs and Global Justice. The Philosophy of Need*. S. Reader. Cambridge, Cambridge University Press. *Royal Institute of Philosophy Supplement* 57: 51–72.

Colyvan, M., D. Cox and K. Steele (2010). "Modelling the Moral Dimension of Decisions." *Noûs* 44 (3): 503–529.

Dancy, J. (2004). *Ethics Without Principles*. Oxford, OUP.

Defoe, D. (1719). "Robinson Crusoe." Retrieved 3rd Jan., 2017, from www.gutenberg.org/files/521/521-h/521-h.htm.

Gert, B. (1988). *Morality*. Oxford, Oxford University Press.

Gigerenzer, G. (2007). *Moral Intuition = Fast and Frugal Heuristics*. *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*. W. Sinnott-Armstrong. Cambridge MA, MIT Press. 2: 1–26.

Gips, J. (1991). "Towards the Ethical Robot. The Second International Workshop on Human and Machine Cognition: Android Epistemology." Perdido Key, Florida.

Goertzel, B., N. Geisweiller, L. Coelho, P. Janicic and C. Pennachin (2011). *Real World Reasoning: Toward Scalable, Uncertain Spatiotemporal, Contextual and Causal Inference*. Paris, Atlantis Press.

Gould, S. J. (1997). "Non-overlapping Magisteria." *Natural History* 106: 16–22.

Harris, S. (2010). *The Moral Landscape: How Science Can Determine Human Values*. London, Bantam.

Hobbes, T. (1651). "Leviathan." Retrieved 17th Jul., 2015, from www.gutenberg.org/files/3207/3207-h/3207-h.htm.

Hume, D. (1738). "A Treatise of Human Nature." Retrieved 17th Jul., 2015, from www.gutenberg.org/files/4705/4705-h/4705-h.htm.

Hursthouse, R. (1999). *On Virtue Ethics*. Oxford, Oxford University Press.

ICRC (1977). "Protocol Additional to the Geneva Conventions of 12 Aug., 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 Jun., 1977. Article 36." Retrieved 24th Feb., 2015, from www.icrc.org/ihl/WebART/470-750045?OpenDocument.

Kant, I. (1785). "Groundwork of the Metaphysics of Morals." Retrieved 29th Nov., 2015, from www.gutenberg.org/ebooks/5682.

Kernighan, B. W. and D. Ritchie (1978). *The C Programming Language*. Englewood Cliffs, NJ, Prentice-Hall.

Korsgaard, C. M. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford; New York, Oxford University Press.

Kurzweil, R. (2012). *How to Create a Mind: The Secret of Human Thought Revealed*. New York, Viking Penguin.

Locke, J. (1689). "Second Treatise on Government." Retrieved 1st Dec., 2015, from www.gutenberg.org/files/7370/7370-h/7370-h.htm.

MacIntyre, A. C. (1959). "Hume on 'Is' and 'Ought'." *The Philosophical Review* 68 (4): 451–468.

Malle, B. F. and M. Scheutz (2014). *Moral Competence in Social Robots*. *Ethics in Science, Technology and Engineering*, 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, IEEE, Chicago.

Maslow, A. (1954). *Motivation and Personality*. New York, Harper & Row.

McConnell, S. (2006). *Software Estimation: Demystifying the Black Art*. Redmond, WA, Microsoft Press.

McCune, W. (2010). "Prover 9 and Mace 4." Retrieved from www.cs.unm.edu/~mccune/Prover9.

Mill, J. S. (1863). *Utilitarianism*. London, Parker, Son and Bourn.

Moor, J. H. (2006). "The Nature, Importance and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4):18–21.

Nagel, T. (1974). "What Is It Like to Be a Bat?" *The Philosophical Review* 83 (4): 435–450.

Noddings, N. (2003). *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley, University of California Press.

O'Neill, O. (2007) *A Simplified Account of Kant's Ethics*. *The Elements of Philosophy*. T. Gendler, S. Siegel and S. M. Cahn. Oxford, OUP:112–114.

Parfit, D. (2011). *On What Matters*. Oxford; New York, Oxford University Press.

Penrose, R. (1990). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford, Oxford University Press.

Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.

Rachels, S. and J. Rachels (2014). *The Elements of Moral Philosophy*. Dubuque, McGraw-Hill Education.

Rand, A. (1961). "The Objectivist Ethics." Retrieved 23rd Oct., 2014, from <http://aynrandlexicon.com/ayn-rand-ideas/the-objectivist-ethics.html>

Rawls, J. (1951). "Outline of a Decision Procedure for Ethics." *The Philosophical Review* 60 (2):177–197.

Rawls, J. (1972). *A Theory of Justice*. Oxford, Clarendon Press.

Raz, J. (2003). "Numbers, with and without Contractualism." *Ratio* 16 (4):346–367.

Reader, S. (2007). *Needs and Moral Necessity*. London; New York, Routledge.

Ross, W. D. (1930). *The Right and the Good*. Oxford, The Clarendon Press.

Rousseau, J.-J. (1762). "The Social Contract." Retrieved 1st Dec., 2015, from www.gutenberg.org/files/46333/46333-h/46333-h.htm.

Russell, S. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, Prentice Hall.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA, Harvard University Press.

Scheffler, S. (2011). *Introduction. On What Matters*. S. Scheffler. Oxford, Oxford University Press. 1:xix–xxxii.

Scherer, K. (2010). The Component Process Model:Architecture for a Comprehensive Computational Model of Emergent Emotion. Blueprint for Affective Computing:A Sourcebook and Manual. K. Scherer , T. Banziger and E. Roesch . Oxford, Oxford University Press:47–84.

Scherer, K. , T. Banziger and E. Roesch (2010). A Blueprint for Affective Computing:A Sourcebook and Manual. Oxford, Oxford University Press.

Scheutz, M. (2012). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. Robot Ethics. P. Lin , K. Abney and G. Bekey .Cambridge, MA, MIT Press:205–222.

Sen, A. K. (1985). Commodities and Capabilities. Amsterdam; New York, North-Holland.

Sidgwick, H. (1907). "The Methods of Ethics." Seventh Edition. Retrieved 5th Oct., 2016, from www.gutenberg.org/files/46743/46743-h/46743-h.htm.

Singer, P. W. (2009). Wired for War:The Robotics Revolution and Conflict in the Twenty-First Century. London, Penguin.

Sperry, R. W. (1983). Science & Moral Priority:Merging Mind, Brain, and Human Values. Oxford, Blackwell.

Timmons, M. (2002). Moral Theory:An Introduction. Lanham, Rowman & Littlefield.

Tonkens, R. (2012). "Out of Character:On the Creation of Virtuous Machines." Ethics and Information Technology 14 (2):137–149.

Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind 59 (236):433–460.

von Wright, G. H. (1951). "Deontic Logic." Mind 60 (237):1–15.

Wallech, W. and C. Allen (2009). Moral Machines. Oxford; New York, Oxford University Press.

Weizenbaum, J. and J. McCarthy (1977). Computer Power and Human Reason:From Judgment to Calculation. San Francisco, WH Freeman.

Wiggins, D. (1982). Needs, Values and Truth:Essays in the Philosophy of Value. Oxford, Oxford University Press.

Wood, A. W. (2008). Kantian Ethics. Cambridge; New York, Cambridge University Press.

Method

Allen, C. , A. Varner and J. Zinser (2000). "Prolegomena to Any Future Artificial Moral Agent." Journal of Experimental and Theoretical Artificial Intelligence 12 (3):251–261.

Arnold, T. and M. Scheutz (2016). "Against the Moral Turing Test:Accountable Design and the Moral Reasoning of Autonomous Systems." Ethics and Information Technology 18 (2):103–115.

Beck, K. (2003). Test-Driven Development:By Example. Boston, MA, Addison-Wesley.

Requirements

Arkin, R. C. (2009). Governing Lethal Behaviour in Autonomous Robots. Boca Rouge, CRC Press.

DARPA (2016). "Explainable Artificial Intelligence." Retrieved 13th Sep., 2016, from www.darpa.mil/program/explainable-artificial-intelligence.

Goodfellow, I. , Y. Bengio and A. Courville (2016). Deep Learning. Cambridge, MA, MIT Press.

Goodman, B. , & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a "right to explanation". Retrieved 14th Sept, 2017, from <https://arxiv.org/abs/1606.08813>

Guarini, M. (2006). "Particularism and Classification and Reclassification of Moral Cases." IEEE Intelligent Systems [H.W.Wilson–AST] 21 (4):22.

Microsoft (2016). "Learning from Tay's Introduction." Retrieved 9th Sep., 2016, from <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>.

Mnih, V. , K. Kavukcuoglu , D. Silver , A. A. Rusu , J. Veness , M. G. Bellemare , A. Graves , M. Riedmiller , A. K. Fidjeland and G. Ostrovski (2015). "Human-Level Control through Deep Reinforcement Learning." Nature 518 (7540):529–533.

Parfit, D. (2011). *On What Matters*. Oxford; New York, Oxford University Press.
Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.

Solution design

Arkin, R. C. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Rouge, CRC Press.

Belnap, N. and M. Perloff (1988). "Seeing to It That: A Canonical Form for Agentives." *Theoria* 54:175–199.

Bringsjord, S. , C. Arkoudas and P. Bello (2006). "Toward a General Logician Methodology for Engineering Ethical Correct Robots." *IEEE Intelligent Systems* 21 (4):38–44.

Bringsjord, S. and G. Naveen Sundar (2013). "Deontic Cognitive Event Calculus." Retrieved 11th Nov., 2015, from www.cs.rpi.edu/~govinn/dcec.pdf.

Bringsjord, S. and J. Taylor (2012). *The Divine-Command Approach to Robot Ethics*. *Robot Ethics: The Ethical and Social Implications of Robotics*. P. Lin , K. Abney and G. Bekey . Cambridge, MIT Press:85–108.

Bringsjord, S. , C. Arkoudas and P. Bello (2006). "Toward a General Logician Methodology for Engineering Ethical Correct Robots." *IEEE Intelligent Systems* 21 (4):38–44

Brooks, R. A. (1990). "Elephants Don't Play Chess." *Robotics and Autonomous Systems* 6 (1):3–15.

Brooks, R. A. (1991). "Intelligence without Representation." *Artificial Intelligence* 47 (1): 139–159.

Castañeda, H.-N. (1981). *The Paradoxes of Deontic Logic: The Simplest Solution to All of Them in One Fell Swoop*. *New Studies in Deontic Logic*. R. Hilpinen . Dordrecht, D. Reidel Publishing Company:37–86.

Dancy, J. (2004). *Ethics Without Principles*. Oxford, OUP.

Franklin, S. (1995). *Artificial Minds*. Cambridge, MA, MIT Press.

Gert, J. (2012). *Normative Bedrock*. Oxford, Oxford University Press.

Guarini, M. (2010). "Particularism, Analogy and Moral Cognition." *Minds and Machines* 20:385–422.

Haikonen, P. (2012). *Consciousness and Robot Sentience*. Hackensack, NJ, World Scientific.

Hansen, J. (2006). "The Paradoxes of Deontic Logic: Alive and Kicking." *Theoria* 72: 221–232.

Horty, J. F. (2001). *Agency and Deontic Logic*. Oxford, Oxford University Press.

Hughes, G. E. and M. J. Cresswell (1996). *A New Introduction to Modal Logic*. New York, Routledge.

Maslow, A. (1954). *Motivation and Personality*. New York, Harper & Row.

McLaren B. (2005). "Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning, 2005." *AAAI Fall Symposium on Machine Ethics*, AAAI Technical Report FS-05-06, 2005

Parfit, D. (2011). *On What Matters*. Oxford; New York, Oxford University Press.

Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.

Poole, D. L. and A. K. Mackworth (2010). *Artificial Intelligence: Foundations of Computational Agents*. Cambridge, Cambridge University Press.

Reader, S. (2007). *Needs and Moral Necessity*. London; New York, Routledge.

Rethink Robotics (2015). "Baxter." Retrieved 17th Feb., 2015, from www.rethinkrobotics.com/baxter/.

Scanlon, T. (2011). *How I am Not a Kantian*. *On What Matters*. S. Scheffler . Oxford, Oxford University Press. 2:119–138.

Wallech, W. and C. Allen (2009). *Moral Machines*. Oxford; New York, Oxford University Press.

Xu, M. (2012). "Actions as Events." *Journal of Philosophical Logic* 41 (4):765–809.

Development – specific norm tests

- Arkin, R. C. (1990). "Integrating Behavioral, Perceptual, and World Knowledge in Reactive Navigation." *Robotics and Autonomous Systems* 6 (1–2):105–122.
- Arkin, R. C. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Rouge, CRC Press.
- Article 36 (2013). "Killer Robots:UK Government Policy on Fully Autonomous Weapons." Retrieved 26th May, 2015, from www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.
- Beauchamp, Z. and J. Savulescu (2013). *Robot Guardians:Teleoperated Combat Vehicles in Humanitarian Intervention. Killing by Remote Control:The Ethics of an Unmanned Military*. B. J. Strawser . Oxford, USA, Oxford University Press:106–125.
- Belnap, N. and M. Perloff (1988). "Seeing to It That:A Canonical Form for Agentives." *Theoria* 54:175–199.
- Bringsjord, S. , C. Arkoudas and P. Bello (2006). "Toward a General Logicist Methodology for Engineering Ethical Correct Robots." *IEEE Intelligent Systems* 21 (4):38–44.
- Bringsjord, S. and G. Naveen Sundar (2013). "Deontic Cognitive Event Calculus." Retrieved 8th Aug., 2015, from www.cs.rpi.edu/~govinn/dcec.pdf.
- Croitoru, M. , N. Oren , S. Miles and M. Luck (2012). "Graphical Norms via Conceptual Graphs." *Knowledge-Based Systems* 29:31–43.
- Everett, J. A. , D. A. Pizarro and M. J. Crockett (2016). "Inference of Trustworthiness from Intuitive Moral Judgments." *Journal of Experimental Psychology:General* 145 (6):772.
- Hansen, J. (2006). "The Paradoxes of Deontic Logic:Alive and Kicking." *Theoria* 72: 221–232.
- Hansson, S. O. (2013). *Alternative Semantics for Deontic Logic. Handbook of Deontic Logic and Normative Systems*. D. Gabbay , J. Horty , X. Parent , R. Van der Mayden and L. Van der Torre . Milton Keynes, College Publications:445–498.
- Harris, S. (2010). *The Moral Landscape:How Science Can Determine Human Values*. London, Bantam.
- Hilpinen, R. (1981). Preface. *New Studies in Deontic Logic*. R. Hilpinen . Dordrecht, D. Reidel Publishing Company:vii–vix.
- Horty, J. F. (2001). *Agency and Deontic Logic*. Oxford, Oxford University Press.
- Hursthouse, R. (1999). *On Virtue Ethics*. Oxford, Oxford University Press.
- ICRC (1996). "Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996 (Protocol II to the 1980 CCW Convention as amended on 3 May 1996)." Retrieved 6th Mar., 2015, from www.icrc.org/ihl/INTRO/575.
- ICRC (1997). "Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction, 18 September 1997." Retrieved 6th Mar., 2015, from www.icrc.org/ihl/INTRO/580.
- Kagan, S. (1989). *The Limits of Morality*. Oxford, Oxford University Press.
- Malle, B. , M. Scheutz , T. Arnold , J. Voiklis and C. Cusimano (2015). *Sacrifice One for the Good of Many:People Apply Different Moral Norms to Humans and Robots*. 10th ACM/IEEE International Conference on Human-Robot Interaction 2015, Portland, ACM.
- Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.
- Pigden, C. R. (1989). "Logic and the Autonomy of Ethics." *Australasian Journal of Philosophy* 67 (2):127–151.
- Rachels, J. (1975). "Active and Passive Euthanasia." *The New England Journal of Medicine* 292 (2):78.
- Ross, W. D. (1930). *The Right and the Good*. Oxford, The Clarendon Press.
- Rudaw (2017). "US Commander Hails Iraqi Forces Beating ISIS Drones and VBIEDs." Retrieved 3rd Feb., 2017, from www.rudaw.net/english/middleeast/iraq/110120171.
- Scharre, P. (2016). "Autonomous Weapons and Operational Risk." Retrieved 1st Mar., 2016, from www.cnas.org/autonomous-weapons-and-operational-risk#.VtSsW_197IU.
- Sperry, R. W. (1983). *Science & Moral Priority:Merging Mind, Brain, and Human Values*. Oxford, Blackwell.

van den Hoven, J. and G.-J. Lokhorst (2002). "Deontic Logic and Computer-Supported Computer Ethics." *Metaphilosophy* 33:376–386.

Development – knowledge representation

Aristotle (c. 350 BC) . "Nichomachean Ethics." Retrieved 29th Nov., 2015, from <http://classics.mit.edu/Aristotle/nicomachaen.html>.

Glimscher, P. and E. Fehr (2013). *Neuroeconomics*. Burlington, Elsevier Science.

Hansson, S. O. (2013). The Varieties of Permission. *Handbook of Deontic Logic and Normative Systems*. D. Gabbay , J. Horty , X. Parent , R. Van der Mayden and L. Van der Torre . Milton Keynes, College Publications. 1:195–240.

Jackson, F. (1992). "Critical Notice." *Australasian Journal of Philosophy* 70 (4): 475–488.

Krotzsch, M. , F. Simanzik and I. Horrocks (2012). "A Description Logic Primer." Retrieved 12th Dec., 2014, from <http://arxiv.org/pdf/1201.4089.pdf>.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA, Belknap Press.

Robinson, I. , J. Webber and E. Eifrem (2015). *Graph Databases*. Sebastapol, CA, O'Reilly.

Development – basic physical needs cases

Anderson, S. L. (2011). The Unacceptability of Asimov's Three Laws as a Basis for Machine Ethics. *Machine Ethics*. M. Anderson and S. L. Anderson . Cambridge, Cambridge University Press:285–296.

Aquinas, T. (1947). *Summa Theologica*. (trans. Fathers of the Dominican Province) New York, Benziger Bros.

Bourget, D. and D. Chalmers (2014). "What Do Philosophers Believe?" *Philosophical Studies* 170 (3):465–500.

Brock, G. (2005). "Needs and Global Justice." *The Philosophy of Need*. S. Reader . Cambridge, Cambridge University Press. *Royal Institute of Philosophy Supplement* 57:51–72.

Everett, J. A. , D. A. Pizarro and M. J. Crockett (2016). "Inference of Trustworthiness from Intuitive Moral Judgments." *Journal of Experimental Psychology:General* 145 (6):772.

Foot, P. (1967). "The Problem of Abortion and the Principle of Double Effect." *Oxford Review* 5:5–15.

Gert, J. (2012). *Normative Bedrock*. Oxford, Oxford University Press.

Greene, J. D. (2007). *The Secret Joke of Kant's Soul. Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*. W. Sinnott-Armstrong . Cambridge, MA, MIT Press. 3:35–80.

Hauser, M. D. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York, HarperCollins.

IMDB (2015). "The Martian." Retrieved 14th Oct., 2016, from www.imdb.com/title/tt3659388/.

Jackson, F. (1992). "Critical Notice." *Australasian Journal of Philosophy* 70 (4): 475–488.

McIntyre, A. (2001). "Doing Away with Double Effect." *Ethics* 111 (2):219–255.

Melaugh, M. (2016). "The Hunger Strike of 1981–List of Dead and Other Hunger Strikers." Retrieved 27th Oct., 2016, from <http://cain.ulst.ac.uk/events/hstrike/chronology.htm>.

Peel, M. (1997). "Hunger Strikes." *BMJ* 315:829–830.

Pereira, L. M. and A. Saptawijaya (2016). *Programming Machine Ethics*. London, Springer.

Prior, A. N. (1957). *Time and Modality*. Oxford, Oxford University Press.

Prior, A. N. (1960). "The Autonomy of Ethics." *Australasian Journal of Philosophy* 38 (3): 199–206.

Prior, A. N. (1967). *Past, Present and Future*. Oxford, Clarendon Press Oxford.

Rachels, J. (1975). "Active and Passive Euthanasia." *The New England Journal of Medicine* 292 (2):78.

Rawls, J. (1972). *A Theory of Justice*. Oxford, Clarendon Press.

Reader, S. (2007). *Needs and Moral Necessity*. London; New York, Routledge.

Ross, A. (1941). "Imperatives and Logic." *Theoria* 7 (1941):53–71.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge, MA, Harvard University Press.

Singer, P. (1997). "The Drowning Child and the Expanding Circle." *New Internationalist* 289:28–30.

Surf Life Saving Australia (2016). *Public Safety and Aquatic Rescue*. Chatswood NSW, Elsevier Australia.

Timmons, M. (2002). *Moral Theory:An Introduction*. Lanham, Rowman & Littlefield.

Weir, A. (2014). *The Martian*. London, Del Ray.

Wiggins, D. (1982). *Needs, Values and Truth:Essays in the Philosophy of Value*. Oxford, Oxford University Press.

Wood, A. (2011). *Trolley Problems. On What Matters*. D. Parfit . Oxford, Oxford University Press. 2:66–82.

Development – fairness and autonomy cases

Aquinas, T. (1947). *Summa Theologica*. (trans. Fathers of the Dominican Province) New York:Benziger Bros.

Arntz, M. , T. Gregory and U. Zierahn (2016). "The Risk of Automation for Jobs in OECD Countries:A Comparative Analysis," OECD Social, Employment and Migration Working Papers, No. 189, OECD Publishing, Paris. <http://dx.doi.org/10.1787/5jlz9h56dvq7-en>

Ashford, E. and T. Mulgan (2012). "Contractualism." *Stanford Encyclopedia of Philosophy*. Retrieved 31st Oct., 2016, from <http://plato.stanford.edu/archives/fall2012/entries/contractualism/>.

Brock, G. (2005). *Needs and Global Justice. The Philosophy of Need*. S. Reader . Cambridge, Cambridge University Press. *Royal Institute of Philosophy Supplement* 57:51–72.

Dunlop, T. (2016). *Why the Future Is Workless*. Sydney, NewSouth.

Ford, M. (2015). *Rise of the Robots:Technology and the Threat of a Jobless Future*. London, Oneworld.

Frey, C. B. , M. A. Osborne , C. Holmes , E. Rahbari , E. Curmi , R. Garlick , J. Chua , G. Friedlander , P. Chalif , G. McDonald and M. Wilkie (2016). "Technology at Work v. 2.0:The Future Is Not What It Used To Be." Retrieved 28th Sep., 2016, from www.oxford-martin.ox.ac.uk/downloads/reports/Citi_GPS_Technology_Work_2.pdf.

Frohlich, N. and J. Oppenheimer (1992). *Choosing Justice:An Experimental Approach to Ethical Theory*. Berkeley, CA, University of California Press.

Hauser, M. D. (2006). *Moral Minds:How Nature Designed Our Universal Sense of Right and Wrong*. New York, HarperCollins.

IMDB (2015). "The Martian." Retrieved 14th Oct., 2016, from www.imdb.com/title/tt3659388/.

Kant, I. (1785). "Groundwork of the Metaphysics of Morals." Retrieved 29th Nov, 2015, from <http://www.gutenberg.org/ebooks/5682>.

Nozick, R. (1974). *Anarchy, State and Utopia*. Oxford, Blackwell.

O'Neill, O. (2004). "Consequences for Non-Consequentialists." *Utilitas* 16 (01):1–11.

Parfit, D. (2011). *On What Matters*. Oxford; New York, Oxford University Press.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Cambridge, MA, Harvard University Press.

Weir, A. (2014). *The Martian*. London, Del Ray.

Moral variation

Korsgaard, C. M. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford; New York: Oxford University Press.

Rachels, S. and J., Rachels (2014). *The Elements of Moral Philosophy*. Dubuque: McGraw-Hill Education.

Wolf, S. (2011). *Hiking the Range. On What Matters*. S. Schleffer . Oxford, Oxford University Press. 2: 33–57.

Testing

Dancy, J. (2004). *Ethics without Principles*. Oxford, Oxford University Press.

Production

Arkin, R. C. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Rouge, CRC Press.

Arkin, R. C. (2010). "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9 (4):332–341.

Bekey, G. A. (2005). *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge, MA, MIT press.

British Standards Institution (2016). *Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*. London, BSI Standards Ltd.

COSMIC (2015). "The COSMIC Functional Size Measurement Method." Retrieved 30th Aug., 2015, from <http://cosmic-sizing.org/publications/introduction-to-the-cosmic-method-of-measuring-software/>.

Croitoru, M. , N. Oren , S. Miles and M. Luck (2012). "Graphical Norms via Conceptual Graphs." *Knowledge-Based Systems* 29:31–43.

Defense Science Board (2012). "The Role of Autonomy in DoD Systems." Retrieved from <http://fas.org/irp/agency/dod/dsb/autonomy.pdf>.

EPSRC (2010). "Principles of Robotics." Retrieved 19th Jan., 2017, from www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Galliot, J. (2015). *Responsibility and War Machines: Towards a Forward-Looking and Functional Account. Rethinking Machine Ethics in the Age of Ubiquitous Technology*. J. White and R. Searle . Hershey, PA, IGI Global:152–165.

Heyns, C. (2015). "Panel on Human Rights and Lethal Autonomous Weapons Systems (LAWS) Comments by Christof Heyns, United Nations Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions (as Finalised after the Meeting)." Retrieved 28th May, 2015, from [http://unog.ch/80256EDD006B8954/\(httpAssets\)/1869331AFF45728BC1257E2D0050EFE0/\\$file/2015_LAWS_MX_Heyns_Transcript.pdf](http://unog.ch/80256EDD006B8954/(httpAssets)/1869331AFF45728BC1257E2D0050EFE0/$file/2015_LAWS_MX_Heyns_Transcript.pdf).

Leveringhaus, A. (2016). *Ethics and Autonomous Weapons*. London, Palgrave Macmillan.

Lin, P. (2015). "The Right to Life and the Martens Clause." Retrieved 21st Apr., 2015, from [http://unog.ch/80256EDD006B8954/\(httpAssets\)/2B52D16262272AE2C1257E2900419C50/\\$file/24+Patrick+Lin_Patrick+SS.pdf](http://unog.ch/80256EDD006B8954/(httpAssets)/2B52D16262272AE2C1257E2900419C50/$file/24+Patrick+Lin_Patrick+SS.pdf).

Madl, T. and S. Franklin (2015). *Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture. A Construction Manual for Robots' Ethical Systems*. R. Trappl . London, Springer:137–153.

Scharre, P. (2016). "Autonomous Weapons and Operational Risk." Retrieved 1st Mar., 2016, from www.cnas.org/autonomous-weapons-and-operational-risk#.VtSsW_I97IU.

Sparrow, R. (2012). *Can Machines Be People? Reflections on the Turing Triage Test. Robot Ethics: The Ethical and Social Implications of Robotics*. P. Lin , K. Abney and G. Bekey . Cambridge MA, MIT Press:301–316.