

Jonathan Kwik

Lawfully Using Autonomous Weapon Technologies

 ASSER PRESS

 Springer

Jonathan Kwik
TMC Asser Instituut
The Hague, The Netherlands

ISBN 978-94-6265-630-7 ISBN 978-94-6265-631-4 (eBook)
<https://doi.org/10.1007/978-94-6265-631-4>

Published by T.M.C. ASSER PRESS, The Hague, The Netherlands www.asserpress.nl
Produced and distributed for T.M.C. ASSER PRESS by Springer-Verlag Berlin Heidelberg

© T.M.C. ASSER PRESS and the author 2024

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

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

This T.M.C. ASSER PRESS imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.

The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

If disposing of this product, please recycle the paper.

Contents

Part I Framing the Study

1	Introduction	3
1.1	Aim of This Book	4
1.2	The Technology Under Consideration	10
1.3	Dramatis Personae	12
1.3.1	Deployer	12
1.3.2	Operator	12
1.3.3	Providing Entity	13
1.4	Goodwill	14
1.5	A Roadmap	15
1.5.1	Controlling AWS	16
1.5.2	Technical Analysis	17
1.5.3	Legal-Operational Analysis	17
1.5.4	Criminal Liability	18
1.5.5	Synthesis and Recommendations	18
	References	19

Part II Controlling AWS

2	Controlling AWS: A Cyclical Process	27
2.1	Introduction	28
2.2	Background, Data Collection, Methodology	30
2.3	Results: Integrated MHC Framework	31
2.4	Facets	33
2.4.1	Awareness	33
2.4.2	Weapon Selection	35
2.4.3	Context Controls	36
2.4.4	Predictability	37
2.4.5	Accountability	37
2.4.6	Assessment Awareness	38

- 2.5 Processes 38
 - 2.5.1 Awareness Informs Weapon Selection and Context Control 38
 - 2.5.2 Awareness and Context Control Permit Prediction 39
 - 2.5.3 All Previous Facets Contribute to Accountability 40
- 2.6 MHC Outside of the Operational Level 41
- 2.7 Discussion 43
- References 44

Part III Understanding AWS

- 3 AI Fundamentals and the Military Environment 51**
 - 3.1 Part III: Stage-Setting 52
 - 3.2 Modern Artificial Intelligence: A Primer 54
 - 3.2.1 Symbolic AI 56
 - 3.2.2 Machine Learning 56
 - 3.3 The AI Decision-Making Process 62
 - 3.3.1 Sensing, Deciding and Acting 63
 - 3.3.2 Acquisition 64
 - 3.3.3 Analysis 66
 - 3.3.4 Decision 67
 - 3.3.5 Action 68
 - 3.4 The Military Environment 68
 - 3.4.1 Importance of Continuous Intelligence 69
 - 3.4.2 Attributes of a Military Environment 70
 - 3.5 The ‘Intended Operational Environment’ (IOE) 71
 - References 73
- 4 Measures of Performance 79**
 - 4.1 Introduction 80
 - 4.2 Accuracy 81
 - 4.2.1 Confusion Matrix and Accuracy Metrics 81
 - 4.2.2 Be Wary of Accuracy 83
 - 4.3 Robustness 86
 - 4.4 Reliability 90
 - 4.5 Understandability 92
 - 4.5.1 Opaque AWS and Solutions to Opacity 93
 - 4.5.2 Utility of Understandability 95
 - 4.6 Recommendations 98
 - References 99
- 5 Causes of Failure 105**
 - 5.1 Introduction 106
 - 5.2 Classical Component Failures 107
 - 5.2.1 Human Error 108
 - 5.2.2 Mechanical Faults Leading to AI Failure 109

- 5.3 AI Failures: Introduction 109
- 5.4 Training Data Issues 111
- 5.5 Input Data Issues 113
- 5.6 Out-of-Distribution and Drift 115
- 5.7 Proxy Mistakes 118
- 5.8 Bias 120
- 5.9 System Failures 121
- 5.10 Recommendations 124
- References 125
- 6 Adversarials: Anti-AI Countermeasures 129**
 - 6.1 Introduction 130
 - 6.2 Digital But Non-AI Specific Countermeasures 132
 - 6.3 Adversarial Input 133
 - 6.3.1 Concept 134
 - 6.3.2 Counters 135
 - 6.3.3 Prerequisites: What Do Opponents Need? 137
 - 6.4 Poisoning 139
 - 6.4.1 Concept 139
 - 6.4.2 Threat Analysis 141
 - 6.5 Security Analysis, Threat Models and Counters 144
 - 6.5.1 Opponent’s Goals and Incentives 145
 - 6.5.2 Knowledge 146
 - 6.5.3 Power 148
 - 6.5.4 Other Assumptions Related to Adversaries 150
 - 6.6 Recommendations and Future Work 151
 - References 152
- Part IV AWS in the Legal-Operational Context**
- 7 Legal-Operational Analysis: Introduction 159**
 - 7.1 Progression Within the MHC Framework 160
 - 7.2 Notes on Legal Methodology 162
 - 7.3 Structure 164
 - References 165
- 8 Legally-Salient Variables for AWS 167**
 - 8.1 Introduction 168
 - 8.2 Variables Related to System Performance 172
 - 8.2.1 Measures of Performance Generally 172
 - 8.2.2 Negative Class Frequency 174
 - 8.3 Temporal and Spatial Variables 178
 - 8.3.1 Temporal Variables 178
 - 8.3.2 Environmental Dynamicity 183
 - 8.3.3 Spatial Dimension 183

- 8.4 Variables Related to Incidental Harm 184
 - 8.4.1 Proximate Civilian Entities and Effectors 185
 - 8.4.2 Negative Class Detection 186
- 8.5 Variables Related to the Military Objectives 188
 - 8.5.1 Type of Objective 189
 - 8.5.2 Target Specificity 192
- 8.6 Opacity 195
- 8.7 Context Controls 196
 - 8.7.1 In the MHC Framework 197
 - 8.7.2 Reducing False Positives 198
 - 8.7.3 Reducing Risk of Incidental Harm 201
 - 8.7.4 Selecting Systems with Higher Levels
of Understandability 204
 - 8.7.5 Multifactorial Controls 204
- 8.8 Linked Control: Do It Right 209
 - 8.8.1 Proper Interface 209
 - 8.8.2 Sufficient Delay 210
 - 8.8.3 Operator Requirements 211
 - 8.8.4 Coordination Between Deployer and Operator 211
- 8.9 Synthesis: What Conditions Make AWS Use More
Difficult? 212
- References 217
- 9 AWS and Targeting 227**
 - 9.1 Introduction 228
 - 9.2 The Scope of an AWS Attack 229
 - 9.2.1 Contours 231
 - 9.2.2 Factors Influencing Scope 233
 - 9.3 Indiscriminate Means 236
 - 9.3.1 When is an AWS Inherently Indiscriminate? 236
 - 9.3.2 IOE-Incompatibility Presumes the AWS is
Indiscriminate 239
 - 9.4 ‘Unaimed’ Use of AWS 240
 - 9.4.1 An Absence of Targeting 240
 - 9.4.2 How Generally May Deployer Define
the Objectives? 242
 - 9.4.3 Too High Opacity as an Unaimed Attack 243
 - 9.5 Calculating Proportionality 244
 - 9.5.1 Uncertainty 246
 - 9.5.2 Attack Scope 248
 - 9.6 Minimisation 250
 - 9.6.1 Context Controls as Obligatory Precautionary
Measures 251
 - 9.6.2 Feasibility and Military Considerations 253

- 9.7 Warnings 257
 - 9.7.1 Deployer- versus AWS-Issued Warnings 258
 - 9.7.2 Feasibility 259
 - 9.7.3 The ‘Effective’ Requirement 260
- 9.8 The Duty to Cancel 261
 - 9.8.1 Attack- or Instance-Level Obligation? 261
 - 9.8.2 Attack-Level Cancelling 262
 - 9.8.3 Instance-Level Canceling 265
- References 268
- 10 Desirability of Using AWS 275**
 - 10.1 Introduction 276
 - 10.2 Purported AWS Benefits 277
 - 10.2.1 Advantages Not Unique to AI 278
 - 10.2.2 Advantages Related to Psychology and Physiology 279
 - 10.2.3 Advantages Unique to AI 281
 - 10.3 How Useful is an AWS? 282
 - 10.3.1 Indispensability 282
 - 10.3.2 Comparative Advantage 284
 - 10.4 Synthesis 285
 - 10.4.1 Assessing Facility and Benefit 286
 - 10.4.2 Desirability: Comparing Facility and Benefit 287
 - 10.4.3 Discussion 288
 - 10.5 Closing Thoughts 291
 - References 292
- Part V Criminal Liability**
- 11 Criminal Liability: Problem Setting 299**
 - 11.1 Introduction 300
 - 11.2 The Responsibility Gap: From Broad to Narrow 301
 - 11.3 Candidates for Liability 303
 - 11.4 The Responsibility Gap Problem 305
 - 11.5 A Roadmap for the Current Part 308
 - 11.6 Housekeeping: Terms and Jurisdictions 309
 - References 311
- 12 Obstacles to Criminal Liability: A Systemic Analysis 315**
 - 12.1 Preliminary Notes 316
 - 12.1.1 Explicit and Implicit Elements of Criminal Liability 316
 - 12.1.2 Deployer Liability Within an Idealised MHC Framework 319

12.2	Questions of Causality and Control	321
12.2.1	Whether and How as a Form of Conduct	321
12.2.2	Pinpointing Causes and the Problem of Many Hands	323
12.2.3	Issues of Control	325
12.2.4	Causality and Control: Findings	326
12.3	Actus Reus	326
12.4	Mens Rea	328
12.4.1	The Epistemic Problem	329
12.4.2	Reduced Perceivability Leading to Lower Cognition	330
12.4.3	Mens Rea Requires Foreseeability	333
12.5	Summation of Findings	334
	References	335
13	Across the Spectrum of Intent	339
13.1	Introduction	340
13.2	Dolus Directus: Desiring a Result	341
13.3	Dolus Indirectus: Certain Risk	343
13.4	Known Risk-Taking	345
13.4.1	Quantitative Risk and AI: Knowing Too Much?	346
13.4.2	Gravitation Toward Risk-Taking, Mitigations, and Small Gaps	350
13.5	Unknown and Generic Risk-Taking	355
13.5.1	Unknown Risk	356
13.5.2	Generic Risk	357
13.5.3	Iteration and Manufactured Ignorance	360
13.6	Judicious Gaps	364
13.6.1	Criminal Law Acting as Intended	364
13.6.2	What Makes a Gap Judicious?	366
13.7	Conclusion: Across the Spectrum	368
	References	372
 Part VI Conclusions		
14	Integration and Closing	379
14.1	Introduction	380
14.2	Summary of Findings	381
14.2.1	Meaningful Human Control: A Novel Operationalisation	381
14.2.2	Awareness: Knowing Your System	382
14.2.3	Legal and Operational Analysis: The ‘May I’ and ‘Should I’	384
14.2.4	Criminal Liability: What If Something Goes Wrong?	388

- 14.3 A Practical Guide for Deployers 389
 - 14.3.1 High-Level Flowchart 390
 - 14.3.2 Legal Flowchart 392
 - 14.3.3 Desirability Flowchart 395
- 14.4 Reflections and Future Work 395
- References 399

- Appendix** 401
- Index** 409

Abbreviations

3D	Three-dimensional
AADA	Acquisition, Analysis, Decision, Action
AC	Accountability (<i>MHC Facet</i>)
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AIV	Adviesraad Internationale Vraagstukken (Netherlands)
APC	Armoured Personnel Carrier
API	Additional Protocol I to the Geneva Conventions
AS	Attack Scope
AV	Autonomous Vehicle
AW	Awareness (<i>MHC Facet</i>)
AWA	Assessment Awareness (<i>MHC Facet</i>)
AW _C	Contextual Awareness (<i>MHC Facet</i>)
AW _I	Interaction Awareness (<i>MHC Facet</i>)
AWS	Autonomous Weapon System
AW _T	Technical Awareness (<i>MHC Facet</i>)
BCC	Broader Chain Candidate for liability
CC	Context Control
CC _L	Linked Control (<i>MHC Facet</i>)
CC _O	Operational Control (<i>MHC Facet</i>)
CC _S	System Control (<i>MHC Facet</i>)
CCW	Convention on Certain Conventional Weapons
CCW-I	First Protocol to the Convention on Certain Conventional Weapons
CCW-II	Second Protocol to the Convention on Certain Conventional Weapons
CCW-IV	Fourth Protocol to the Convention on Certain Conventional Weapons
CDE	Collateral Damage Estimation
CDEM	Collateral Damage Estimation Methodology
CEP	Circular Error Probable
CIV	Civilian
CIWS	Close-In Weapon System
CPU	Central Processing Unit

C-RAM	Counter Rocket, Artillery, Mortar
D	Accused
DARPA	Defense Advanced Research Projects Agency (US)
DDD	Dirty, Dull, Dangerous
DMZ	Demilitarised Zone
DoD	Department of Defense (US)
DPH	Direct Participation in Hostilities
EU	European Union
GAN	Generative Adversarial Network
GARD	Guaranteeing AI Robustness Against Deception (DARPA agency)
GGE	Group of Governmental Experts (CCW Conference Geneva)
GPS	Global Positioning System
GPU	Graphics Processing Unit
HAF	Haftar Armed Forces (Libya)
HARM	High-speed Anti-Radiation Missile
HFACS	Human Factors Analysis and Classification System (US DoD)
HIV	Human Immunodeficiency Virus
HLEG	High-Level Expert Group (EU)
HMI	Human-Machine Interface
HPCR	Humanitarian Policy and Conflict Research
HQ	Headquarters
IAC	International Armed Conflict
ICC	International Criminal Court
ICJ	International Court of Justice
ICL	International Criminal Law
ICRC	International Committee of the Red Cross
ICTY	International Criminal Tribunal for the Former Yugoslavia
IH	Incidental Harm (collateral damage)
IHL	International Humanitarian Law
IOE	Intended Operational Environment
ISR	Intelligence, Surveillance, Reconnaissance
JCE	Joint Criminal Enterprise
LAR	Lethal Autonomous Robot
LAWS	Lethal Autonomous Weapon System
LC	Linked Control
LOAC	Law of Armed Conflict
MA	Military Advantage
MHC	Meaningful Human Control
ML	Machine Learning
MOD	Ministry of Defence (UK)
MoP	Measure of Performance
NASA	National Aeronautics and Space Administration (US)
NATO	North Atlantic Treaty Organisation
NGO	Non-Governmental Organisation
NIAC	Non-International Armed Conflict

NLP	Natural Language Processing
OOD	Out-of-distribution
OODA	Observe, Orient, Decide, Act
\mathbb{P}	Probability
PGM	Precision-Guided Munition
POW	Prisoner of War
<i>PR</i>	Predictability (<i>MHC Facet</i>)
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ROE	Rules of Engagement
RSK	Republic of Serbian Krajina
TEV&V	Test, Evaluation, Verification and Validation
TTP	Tactics, Techniques, and Procedures
UAV	Unmanned Aerial Vehicle
UK	United Kingdom
UN	United Nations
US	United States
USAF	United States Air Force
V	Victim
W_p	Weapon
<i>WS</i>	Weapon Selection (<i>MHC Facet</i>)
XAI	eXplainable AI